

Edwige Godlewski  
Pierre-Arnaud Raviart

# Numerical Approximation of Hyperbolic Systems of Conservation Laws

*Second Edition*

# **Applied Mathematical Sciences**

## **Volume 118**

### **Series Editors**

Anthony Bloch, Department of Mathematics, University of Michigan, Ann Arbor, MI, USA

[abloch@umich.edu](mailto:abloch@umich.edu)

C. L. Epstein, Department of Mathematics, University of Pennsylvania, Philadelphia, PA, USA

[cle@math.upenn.edu](mailto:cle@math.upenn.edu)

Alain Goriely, Department of Mathematics, University of Oxford, Oxford, UK

[goriely@maths.ox.ac.uk](mailto:goriely@maths.ox.ac.uk)

Leslie Greengard, New York University, New York, NY, USA

[Greengard@cims.nyu.edu](mailto:Greengard@cims.nyu.edu)

### **Advisory Editors**

J. Bell, Center for Computational Sciences and Engineering, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

P. Constantin, Department of Mathematics, Princeton University, Princeton, NJ, USA

R. Durrett, Department of Mathematics, Duke University, Durham, CA, USA

R. Kohn, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

R. Pego, Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

L. Ryzhik, Department of Mathematics, Stanford University, Stanford, CA, USA

A. Singer, Department of Mathematics, Princeton University, Princeton, NJ, USA

A. Stevens, Department of Applied Mathematics, University of Münster, Münster, Germany

S. Wright, Computer Sciences Department, University of Wisconsin, Madison, WI, USA

### **Founding Editors**

F. John, New York University, New York, NY, USA

J. P. LaSalle, Brown University, Providence, RI, USA

L. Sirovich, Brown University, Providence, RI, USA

The mathematization of all sciences, the fading of traditional scientific boundaries, the impact of computer technology, the growing importance of computer modeling and the necessity of scientific planning all create the need both in education and research for books that are introductory to and abreast of these developments. The purpose of this series is to provide such books, suitable for the user of mathematics, the mathematician interested in applications, and the student scientist. In particular, this series will provide an outlet for topics of immediate interest because of the novelty of its treatment of an application or of mathematics being applied or lying close to applications. These books should be accessible to readers versed in mathematics or science and engineering, and will feature a lively tutorial style, a focus on topics of current interest, and present clear exposition of broad appeal. A compliment to the Applied Mathematical Sciences series is the Texts in Applied Mathematics series, which publishes textbooks suitable for advanced undergraduate and beginning graduate courses.

More information about this series at <http://www.springer.com/series/34>

Edwige Godlewski • Pierre-Arnaud Raviart

# Numerical Approximation of Hyperbolic Systems of Conservation Laws

Second Edition



Springer

Edwige Godlewski  
Laboratoire Jacques-Louis Lions  
Sorbonne University  
Paris, France

Pierre-Arnaud Raviart  
Laboratoire Jacques-Louis Lions  
Sorbonne University  
Paris, France

ISSN 0066-5452  
Applied Mathematical Sciences  
ISBN 978-1-0716-1342-9  
<https://doi.org/10.1007/978-1-0716-1344-3>

ISSN 2196-968X (electronic)  
ISBN 978-1-0716-1344-3 (eBook)

Mathematics Subject Classification: 35L65, 35L67, 65M06, 65M08, 65M12, 76Nxx, 35L50, 35L60, 35Q35, 65Mxx, 35Q20, 35Q86, 76P05, 76W05, 80A32

© Springer Science+Business Media, LLC, part of Springer Nature 1996, 2021  
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature

The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

# Preface to the Second Edition

There was an obvious need to complete the first edition of this textbook with the treatment of source terms. Thus, a new chapter (Chap. VII) has been added, which also provides a few important principles concerning non-conservative systems that are naturally introduced with the derivation of well-balanced or asymptotic preserving schemes. Note that most theoretical results are only referred to since it is out of scope to give detailed proofs; these may be tricky and are often quite technical.

We took the opportunity of this second edition to include more examples in the introduction chapter (now Chap. I), such as MHD, shallow water, and flow in a nozzle, and to give some insights on multiphase flow models; this last subject deserves a much longer treatment. Then we thought it is important to emphasize the change of frame from Eulerian to Lagrangian coordinates and the specificity of fluid systems. Additionally, the low Mach limit has been addressed in the chapter devoted to multidimensional systems (now Chap. V) with the final section introducing all Mach schemes.

For 25 years, there has been a tremendous lot of work dedicated to the numerical approximation of hyperbolic systems, among which we choose to introduce the relaxation approach, now at the end of Chap. IV and the case of discontinuous fluxes, and interface coupling, a topic covered in Chap. VII. Both subjects are treated in some specific outlines.

Then, some complements may be found here and there, such as recalling some results of our earlier publication at the beginning of Chap. IV, or more examples of systems of two equations in Chap. II.

We must finally confess that it took us some time to complete the work of this second edition, for different reasons. In fact, most of this work was achieved several years ago, which may explain why only few very recent results are presented, some of them are just mentioned in the notes at the end of each chapter, to give a hint and provide references where the subject is more thoroughly treated.

# Preface to the First Edition

This work is devoted to the theory and approximation of nonlinear hyperbolic systems of conservation laws in one or two space variables. It follows directly a previous publication on hyperbolic systems of conservation laws by the same authors, and we shall make frequent references to Godlewski and Raviart (1991) (hereafter noted G.R.), though the present volume can be read independently. This earlier publication, apart from a first chapter, especially covered the scalar case. Thus, we shall detail here neither the mathematical theory of multidimensional *scalar* conservation laws nor their approximation in the one-dimensional case by finite-difference conservative schemes, both of which were treated in G.R., but we shall mostly consider systems. The theory for systems is in fact much more difficult and not at all completed. This explains why we shall mainly concentrate on some theoretical aspects that are needed in the applications, such as the solution of the Riemann problem, with occasional insights into more sophisticated problems.

The present book is divided into six chapters, including an introductory chapter<sup>1</sup>. For the reader's convenience, we shall resume in this Introduction the notions that are necessary for a self-sufficient understanding of this book –the main definitions of hyperbolicity, weak solutions, and entropy– present the practical examples that will be thoroughly developed in the following chapters, and recall the main results concerning the scalar case.

Chapter I is devoted to the resolution of the Riemann problem for a general hyperbolic system in one space dimension, introducing the classical notions of Riemann invariants and simple waves, the rarefaction and shock curves, and characteristics and entropy conditions. The theory is then applied to the p-system.

In Chap. II, we make a closer study of the one-dimensional system of gas dynamics. We solve the Riemann problem in detail and then present the

---

<sup>1</sup> The numbering of the chapters has changed in the second edition, the Introduction is now Chap. I. Hence in what follows, Chap. I refers to what is now Chap. II and so on.

simplest models of reacting flow, first the Chapman-Jouguet theory and then the Z.N.D. model for detonation.

After this theoretical approach, we go into the numerical approximation of hyperbolic systems by conservative finite-difference methods. The most usual schemes for one-dimensional systems are developed in Chap. III, with special emphasis on the application to gas dynamics. The last section begins with a short account on the kinetic theory so as to introduce kinetic schemes.

Chapter IV is devoted to the study of finite volume methods for bidimensional systems, preceded by some theoretical considerations on multidimensional systems.

For the sake of completeness, we could not avoid the problem of boundary conditions. Chapter V is but an introduction to the complex theory and presents some numerical boundary treatment.

The authors wish to thank R. Abgrall, F. Coquel, F. Dubois, and particularly T. Gallouet, B. Perthame, and D. Serre, from whom they learned a great deal and who answered willingly and most amiably their many questions.

They owe thanks to the SMAI reading committee and to the reviewers, who made very valuable suggestions.

The first author is grateful to all her colleagues who encouraged her in completing this huge work, especially to H. Le Dret and F. Murat for so often giving her their time, and to L. Ruprecht for her kind and competent assistance in the retyping of the final manuscript; such friendly help was invaluable.

Paris, France  
September 1995

E. Godlewski and P.-A. Raviart

# Contents

|           |   |     |
|-----------|---|-----|
| <b>I</b>  | <b>Introduction</b>   | 1   |
| 1         | Definitions and Examples  | 1   |
| 2         | Fluid Systems in Eulerian and Lagrangian Frames                         | 6   |
| 3         | Some Averaged Models: Shallow Water, Flow in a Duct, and Two-Phase Flow | 20  |
| 4         | Weak Solutions of Systems of Conservation Laws                          | 27  |
| 4.1       | Characteristics in the Scalar One-Dimensional Case                      | 27  |
| 4.2       | Weak Solutions: The Rankine-Hugoniot Condition                          | 30  |
| 4.3       | Example of Nonuniqueness of Weak Solutions                              | 35  |
| 5         | Entropy Solution  | 37  |
| 5.1       | A Mathematical Notion of Entropy  | 37  |
| 5.2       | The Vanishing Viscosity Method  | 44  |
| 5.3       | Existence and Uniqueness of the Entropy Solution in the Scalar Case     | 50  |
| Notes     |   | 52  |
| <b>II</b> | <b>Nonlinear Hyperbolic Systems in One Space Dimension</b>              | 55  |
| 1         | Linear Hyperbolic Systems with Constant Coefficients                    | 55  |
| 2         | The Nonlinear Case, Definitions and Examples                            | 58  |
| 2.1       | Change of Variables, Change of Frame                                    | 60  |
| 2.2       | The Gas Dynamics Equations  | 66  |
| 2.3       | Ideal MHD   | 75  |
| 3         | Simple Waves and Riemann Invariants                                     | 80  |
| 3.1       | Rarefaction Waves   | 80  |
| 3.2       | Riemann Invariants  | 84  |
| 4         | Shock Waves and Contact Discontinuities                                 | 92  |
| 5         | Characteristic Curves and Entropy Conditions                            | 103 |
| 5.1       | Characteristic Curves   | 103 |
| 5.2       | The Lax Entropy Conditions  | 107 |
| 5.3       | Other Entropy Conditions  | 110 |

|            |  |            |
|------------|--|------------|
| 6          | Solution of the Riemann Problem . . . . .  | 116        |
| 7          | Examples of Systems of Two Equations . . . . .                                       | 120        |
| 7.1        | The Case of a Linear or a Linearly Degenerate System . . . . .                       | 120        |
| 7.2        | The Riemann Problem for the $p$ -System . . . . .                                    | 122        |
| 7.3        | The Riemann Problem for the Barotropic Euler System . . . . .                        | 133        |
|            | Notes . . . . .  | 137        |
| <b>III</b> | <b>Gas Dynamics and Reacting Flows . . . . .</b>                                     | <b>141</b> |
| 1          | Preliminaries . . . . .  | 141        |
| 1.1        | Properties of the Physical Entropy . . . . .   | 141        |
| 1.2        | Ideal Gases . . . . .  | 149        |
| 2          | Entropy Satisfying Shock Conditions . . . . .  | 153        |
| 3          | Solution of the Riemann Problem . . . . .  | 171        |
| 4          | Reacting Flows: The Chapman-Jouguet Theory . . . . .                                 | 188        |
| 5          | Reacting Flows: The Z.N.D. Model for Detonations . . . . .                           | 207        |
|            | Notes . . . . .  | 212        |
| <b>IV</b>  | <b>Finite Volume Schemes for One-Dimensional Systems . . . . .</b>                   | <b>215</b> |
| 1          | Generalities on Finite Volume Methods for Systems . . . . .                          | 215        |
| 1.1        | Extension of Scalar Schemes to Systems: Some Examples . . . . .                      | 221        |
| 1.2        | $L^2$ Stability . . . . .  | 230        |
| 1.3        | Dissipation and Dispersion . . . . .   | 232        |
| 2          | Godunov's Method . . . . .   | 236        |
| 2.1        | Godunov's Method for Systems . . . . .   | 236        |
| 2.2        | The Gas Dynamics Equations in a Moving Frame . . . . .                               | 240        |
| 2.3        | Godunov's Method in Lagrangian Coordinates . . . . .                                 | 242        |
| 2.4        | Godunov's Method in Eulerian Coordinates<br>(Direct Method) . . . . .                | 245        |
| 2.5        | Godunov's Method in Eulerian Coordinates<br>(Lagrangian Step + Projection) . . . . . | 246        |
| 2.6        | Godunov's Method in a Moving Grid . . . . .  | 249        |
| 3          | Godunov-Type Methods . . . . .   | 250        |
| 3.1        | Approximate Riemann Solvers and Godunov-Type Methods . . . . .                       | 250        |
| 3.2        | Roe's Method and Variants . . . . .  | 259        |
| 3.3        | The H.L.L. Method . . . . .  | 269        |
| 3.4        | Osher's Scheme . . . . .   | 274        |
| 4          | Roe-Type Methods for the Gas Dynamics System . . . . .                               | 283        |
| 4.1        | Roe's Method for the Gas Dynamics Equations: (I)<br>The Ideal Gas Case . . . . .     | 283        |
| 4.2        | Roe's Method for the Gas Dynamics Equations: (II)<br>The "Real Gas" Case . . . . .   | 294        |

|     |  |     |
|-----|--|-----|
| 4.3 | A Roe-Type Linearization Based on Shock Curve Decomposition .....                  | 299 |
| 4.4 | Another Roe-Type Linearization Associated with a Path .....                        | 303 |
| 4.5 | The Case of the Gas Dynamics System in Lagrangian Coordinates .....                | 309 |
| 5   | Flux Vector Splitting Methods .....  | 320 |
| 5.1 | General Formulation .....  | 320 |
| 5.2 | Application to the Gas Dynamics Equations: (I) Steger and Warming's Approach ..... | 322 |
| 5.3 | Application to the Gas Dynamics Equations: (II) Van Leer's Approach .....          | 326 |
| 6   | Van Leer's Second-Order Method .....   | 329 |
| 6.1 | Van Leer's Method for Systems .....  | 329 |
| 6.2 | Solution of the Generalized Riemann Problem .....                                  | 333 |
| 6.3 | The G.R.P. for the Gas Dynamics Equations in Lagrangian Coordinates .....          | 336 |
| 6.4 | Use of the G.R.P. in van Leer's Method .....                                       | 345 |
| 7   | Kinetic Schemes for the Euler Equations .....                                      | 354 |
| 7.1 | The Boltzmann Equation .....   | 354 |
| 7.2 | The B.G.K. Model .....   | 363 |
| 7.3 | The Kinetic Scheme .....   | 368 |
| 7.4 | Some Extensions of the Kinetic Approach .....                                      | 388 |
| 8   | Relaxation Schemes .....   | 394 |
| 8.1 | Introduction to Relaxation .....   | 394 |
| 8.2 | Model Examples .....   | 399 |
| 8.3 | A Relaxation Scheme for the Euler System .....                                     | 407 |
|     | Notes .....  | 420 |
| V   | <b>The Case of Multidimensional Systems</b> .....                                  | 425 |
| 1   | Generalities on Multidimensional Hyperbolic Systems .....                          | 425 |
| 1.1 | Definitions .....  | 425 |
| 1.2 | Characteristics .....  | 428 |
| 1.3 | Simple Plane Waves .....   | 433 |
| 1.4 | Shock Waves .....  | 437 |
| 2   | The Gas Dynamics Equations in Two Space Dimensions .....                           | 439 |
| 2.1 | Entropy and Entropy Variables .....  | 440 |
| 2.2 | Invariance of the Euler Equations .....  | 443 |
| 2.3 | Eigenvalues .....  | 450 |
| 2.4 | Characteristics .....  | 455 |
| 2.5 | Plane Wave Solutions: Self-Similar Solutions .....                                 | 460 |
| 3   | Multidimensional Finite Difference Schemes .....                                   | 468 |
| 3.1 | Direct Approach .....  | 468 |
| 3.2 | Dimensional Splitting .....  | 480 |
| 4   | Finite-Volume Methods .....  | 487 |
| 4.1 | Definition of the Finite-Volume Method .....                                       | 488 |

|            |  |            |
|------------|--|------------|
| 4.2        | General Results .....  | 499        |
| 4.3        | Usual Schemes .....  | 517        |
| 5          | Second-Order Finite-Volume Schemes .....                                 | 533        |
| 5.1        | MUSCL-Type Schemes .....   | 533        |
| 5.2        | Other Approaches .....   | 546        |
| 6          | An Introduction to All-Mach Schemes for the System of Gas Dynamics ..... | 547        |
| 6.1        | The Low Mach Limit of the System of Gas Dynamics .....                   | 548        |
| 6.2        | Asymptotic Analysis of the Semi-Discrete Roe Scheme .....                | 552        |
| 6.3        | An All-Mach Semi-Discrete Roe Scheme .....                               | 561        |
| 6.4        | Asymptotic Analysis of the Semi-Discrete HLL Scheme .....                | 568        |
| 6.5        | An All-Mach Semi-Discrete HLL Scheme .....                               | 574        |
|            | Notes .....  | 578        |
| <b>VI</b>  | <b>An Introduction to Boundary Conditions .....</b>                      | <b>581</b> |
| 1          | The Initial Boundary Value Problem in the Linear Case .....              | 581        |
| 1.1        | Scalar Advection Equations .....   | 582        |
| 1.2        | One-Dimensional Linear Systems. Linearization .....                      | 587        |
| 1.3        | Multidimensional Linear Systems .....                                    | 590        |
| 2          | The Nonlinear Approach .....   | 599        |
| 2.1        | Nonlinear Equations .....  | 599        |
| 2.2        | Nonlinear Systems .....  | 602        |
| 3          | Gas Dynamics .....   | 606        |
| 3.1        | Fluid Boundary (Linearized Approach) .....                               | 607        |
| 3.2        | Solid or Rigid Wall Boundary .....                                       | 610        |
| 4          | Absorbing Boundary Conditions .....                                      | 610        |
| 5          | Numerical Treatment .....  | 618        |
| 5.1        | Finite Difference Schemes .....  | 618        |
| 5.2        | Finite Volume Approach .....   | 621        |
|            | Notes .....  | 625        |
| <b>VII</b> | <b>Source Terms .....</b>  | <b>627</b> |
| 1          | Introduction to Source Terms .....                                       | 627        |
| 1.1        | Some General Considerations for Systems with Source Terms .....          | 628        |
| 1.2        | Simple Examples of Source Terms in the Scalar Case .....                 | 629        |
| 1.3        | Numerical Treatment of Source Terms .....                                | 632        |
| 1.4        | Examples of Systems with Source Terms .....                              | 639        |
| 2          | Systems with Geometric Source Terms .....                                | 643        |
| 2.1        | Nonconservative Systems .....  | 644        |
| 2.2        | Stationary Waves and Resonance .....                                     | 650        |

|     |  |     |
|-----|--|-----|
| 2.3 | Case of a Nozzle with Discontinuous Section . . . . .                              | 656 |
| 2.4 | The Example of the Shallow Water System . . . . .                                  | 662 |
| 3   | Specific Numerical Treatment of Source Terms . . . . .                             | 665 |
| 3.1 | Some Numerical Considerations for Flow in a Nozzle . . . . .                       | 665 |
| 3.2 | Preserving Equilibria, Well-Balanced Schemes . . . . .                             | 667 |
| 3.3 | Schemes for the Shallow Water System . . . . .                                     | 675 |
| 4   | Simple Approximate Riemann Solvers . . . . .                                       | 679 |
| 4.1 | Definition of Simple Approximate Riemann Solvers . . . . .                         | 679 |
| 4.2 | Well-Balanced Simple Schemes . . . . .   | 682 |
| 4.3 | Simple Approximate Riemann Solvers in Lagrangian or Eulerian Coordinates . . . . . | 685 |
| 4.4 | The Example of the Gas Dynamics Equations with Gravity and Friction . . . . .      | 688 |
| 4.5 | Link with Relaxation Schemes . . . . .   | 697 |
| 5   | Stiff Source Terms, Asymptotic Preserving Numerical Schemes . . . . .              | 705 |
| 5.1 | Introduction . . . . .   | 705 |
| 5.2 | Some Simple Examples . . . . .   | 707 |
| 5.3 | Derivation of an AP Scheme for the Linear Model . . . . .                          | 711 |
| 5.4 | Euler System with Gravity and Friction . . . . .                                   | 721 |
| 6   | Interface Coupling . . . . .   | 731 |
| 6.1 | Introduction to Interface Coupling . . . . .                                       | 731 |
| 6.2 | The Interface Coupling Condition . . . . .   | 734 |
| 6.3 | Numerical Coupling . . . . .   | 744 |
|     | Notes . . . . .  | 746 |
|     | <b>References</b> . . . . .  | 749 |
|     | <b>Index</b> . . . . .   | 831 |



# I

## Introduction

### 1 Definitions and Examples

In this section, we present the general form of systems of conservation laws in several space variables, and we give some important examples of such systems that arise in continuum physics.

Let  $\Omega$  be an open subset of  $\mathbb{R}^p$ , and let  $\mathbf{f}_j$ ,  $1 \leq j \leq d$ , be  $d$  smooth functions from  $\Omega$  into  $\mathbb{R}^p$ ; the general form of a system of conservation laws in several space variables is

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) = \mathbf{0}, \quad \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d, \quad t > 0, \quad (1.1)$$

where

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}$$

is a vector-valued function from  $\mathbb{R}^d \times [0, +\infty[$  into  $\Omega$ . The set  $\Omega$  is called the set of states and the functions

$$\mathbf{f}_j = \begin{pmatrix} f_{1j} \\ \vdots \\ f_{pj} \end{pmatrix}$$

are called the flux functions. One says that system (1.1) is written in *conservative form*.

In the sequel, we will also write the system (1.1) in the form

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathbf{f}(\mathbf{u}) = \mathbf{0}$$

where  $\mathbf{f}$  stands for the matrix-valued function

$$\mathbf{f} = (f_{ij})_{1 \leq i \leq p, 1 \leq j \leq d}$$

and  $\nabla \cdot$  is the divergence operator (we will equivalently use the notation  $\operatorname{div}$ )

$$\nabla \cdot \mathbf{f} = \sum_{j=1}^d \frac{\partial}{\partial x_j} f_j.$$

Formally, the system (1.1) expresses the conservation of the  $p$  quantities  $u_1, \dots, u_p$ . In fact, let  $D$  be an arbitrary domain of  $\mathbb{R}^d$ , and let  $\mathbf{n} = (n_1, \dots, n_d)^T$  be the outward unit normal to the boundary  $\partial D$  of  $D$ . Then, it follows from (1.1) that

$$\frac{d}{dt} \int_D \mathbf{u} \, d\mathbf{x} + \sum_{j=1}^d \int_{\partial D} \mathbf{f}_j(\mathbf{u}) \, n_j \, dS = \mathbf{0}.$$

This balance equation has now a very natural meaning: the time variation of  $\int_D \mathbf{u} \, d\mathbf{x}$  is equal to the losses through the boundary  $\partial D$ .

In all the following, we shall be concerned with the study of *hyperbolic* systems of conservation laws, which we define in the following way. For all  $j = 1, \dots, d$ , let

$$\mathbf{A}_j(\mathbf{u}) = \left( \frac{\partial f_{ij}}{\partial u_k}(\mathbf{u}) \right)_{1 \leq i, k \leq p}$$

be the Jacobian matrix of  $\mathbf{f}_j(\mathbf{u})$ ; the system (1.1) is called *hyperbolic* if, for any  $\mathbf{u} \in \Omega$  and any  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_d) \in \mathbb{R}^d$ ,  $\boldsymbol{\omega} \neq \mathbf{0}$ , the matrix

$$\mathbf{A}(\mathbf{u}, \boldsymbol{\omega}) = \sum_{j=1}^d \omega_j \mathbf{A}_j(\mathbf{u})$$

has  $p$  real eigenvalues  $\lambda_1(\mathbf{u}, \boldsymbol{\omega}) \leq \lambda_2(\mathbf{u}, \boldsymbol{\omega}) \leq \dots \leq \lambda_p(\mathbf{u}, \boldsymbol{\omega})$  and  $p$  linearly independent corresponding eigenvectors  $\mathbf{r}_1(\mathbf{u}, \boldsymbol{\omega}), \dots, \mathbf{r}_p(\mathbf{u}, \boldsymbol{\omega})$ , i.e.,

$$\mathbf{A}(\mathbf{u}, \boldsymbol{\omega}) \mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}) = \lambda_k(\mathbf{u}, \boldsymbol{\omega}) \mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}), \quad 1 \leq k \leq p.$$

If, in addition, the eigenvalues  $\lambda_k(\mathbf{u}, \boldsymbol{\omega})$  are all distinct, the system (1.1) is called *strictly hyperbolic*.

In fact, little is known about systems in more than one space variable unless they are *symmetrizable*, i.e., there exists for all  $\mathbf{u} \in \Omega$  a symmetric positive-definite matrix  $\mathbf{A}_0(\mathbf{u})$  smoothly varying with  $\mathbf{u}$  such that the matrices

$$\mathbf{A}_0(\mathbf{u}) \mathbf{A}_j(\mathbf{u}), \quad 1 \leq j \leq d$$

are symmetric. Symmetrizable systems of conservation laws are clearly hyperbolic. Note that most of the systems of conservation laws that arise in practice are symmetrizable; this is a consequence of the existence of an entropy function (Godunov-Mock theorem [543], see Theorem 5.1 below).

For such systems, we shall study the *Cauchy problem*, or initial value problem (IVP): find a function  $\mathbf{u} : (\mathbf{x}, t) \in \mathbb{R}^d \times [0, \infty[ \rightarrow \mathbf{u}(\mathbf{x}, t) \in \Omega$  that is a solution of (1.1) satisfying the initial condition

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \quad (1.2)$$

where  $\mathbf{u}_0 : \mathbb{R}^d \rightarrow \Omega$  is a given function. The initial boundary value problem (I.B.V.P.) will be considered in Chap. VI. One aim of this introduction is to make precise in which sense (1.1), (1.2) is to be taken.

In the one-dimensional case, when  $\mathbf{u}_0$  has the following particular form,

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_\ell, & x < 0 \\ \mathbf{u}_r, & x > 0, \end{cases} \quad (1.3)$$

with constant states  $\mathbf{u}_\ell, \mathbf{u}_r$ , this Cauchy problem is called the (one-dimensional) *Riemann problem* (see Chap. V, Remark 2.8, for the definition of a 2-D Riemann problem).

In the scalar case (i.e.,  $p = 1$ ), the simplest example of a nonlinear conservation law is given by Burgers' equation.

*Example 1.1. The Burgers' equation.* The scalar parabolic equation

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} = 0 \quad (1.4a)$$

was introduced in particular by Burgers as the simplest differential model for a fluid flow and is therefore often called the (viscous) Burgers' equation. Though very simple, this equation can be regarded as a model for decaying free turbulence (see Cole [325]). A number of authors have developed the asymptotic theory of Navier-Stokes equations in terms of Burgers' equation, and it is thus often used in numerical tests [1059]. Burgers studied the limit equation when  $\nu$  tends to zero, which we write in conservation form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left( \frac{u^2}{2} \right) = 0. \quad (1.4b)$$

Equation (1.4b) is the inviscid Burgers' equation (or Burgers' equation without viscosity), which, for brevity, we shall simply call from now on Burgers' equation. It occurs in particular in wave theory to depict the distortion of waveform in simple waves (see Lighthill [800, Sec. 2.9], Whitham [1188, Sec. 2.8]).

We shall see that Burgers' equation possesses all the features of a scalar convex equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \quad (1.5)$$

where  $f : \mathbb{R} \rightarrow \mathbb{R}$  is a *convex* smooth function. In particular, the Cauchy problem for Burgers' equation may have discontinuous weak solutions even for a smooth initial function  $u_0$ , and the solution of the Riemann problem

is either a shock propagating or a rarefaction wave (see Fig. 4.2), both being kinds of waves that are involved in the solution of the Riemann problem for a system.

*Remark 1.1.* It will be very easy to derive the properties of Eq. (1.5) when the flux  $f$  is concave, from the convex case. An example of a concave flux is illustrated by the classical LWR traffic model (for Lighthill-Whitham-Richards) for which

$$f(\rho) = u_{max}\rho\left(1 - \frac{\rho}{\rho_{max}}\right),$$

where  $\rho \in [0, \rho_{max}]$  measures the density of cars and  $u_{max}$  is a maximum velocity.  $\square$

Finally, it is worth mentioning that the Cauchy problem for (1.4a) has an explicit solution, obtained using the Cole-Hopf transform

$$u = -2\nu \frac{\varphi_x}{\varphi}.$$

It eliminates the nonlinear term and transforms (1.4a) into the heat equation

$$\frac{\partial \varphi}{\partial t} = \nu \frac{\partial^2 \varphi}{\partial x^2},$$

for which explicit expressions of the solution are known. For details, we refer to the original papers of Hopf [630], Cole [325], and Whitham (1974, Chapter 4) [1188] where a thorough study of Eq. (1.4a) (including limit as  $\nu \rightarrow 0$  and shock structure) can be found.  $\square$

*Example 1.2. The Buckley-Leverett equation.* In petroleum engineering, the Buckley-Leverett equation is a one-dimensional model for a two-phase flow, an oil and water mixture, in a porous medium. We assume that the porosity is constant (taken to be 1) and we ignore capillarity and gravity effects. Then the reduced water saturation  $s$  satisfies and Eq. (1.5) with flux

$$f(s) = \frac{k_{r,w}(s)/\mu_w}{k_{r,w}(s)/\mu_w + k_{r,o}(s)/\mu_o}, \quad (1.6)$$

where  $\mu$  denotes the viscosity (supposed to be constant) and  $k_r(s)$  the relative permeability of a fluid (the indices  $w$  and  $o$  refer to water and oil, respectively); the set of states is the interval  $[0, 1]$ . A typical example is obtained by taking  $k_{r,w} = s^2$ ,  $k_{r,o}(s) = (1-s)^2$ , so that  $f$  writes

$$f(s) = \frac{s^2}{s^2 + (1-s)^2 \mu_w / \mu_o},$$

and it is an increasing, S-shaped (nonconvex) function (see [1147] for a possible extension of this model and references therein).  $\square$

Let us next introduce some fundamental examples of the systems of conservation laws that we shall study in the following chapters.

*Example 1.3. The p-system.* A model for one-dimensional isentropic gas dynamics in Lagrangian coordinates is given by the following system of two equations:

$$\begin{cases} \frac{\partial v}{\partial t} - \frac{\partial u}{\partial x} = 0 \\ \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} p(v) = 0, \end{cases} \quad (1.7)$$

where  $v$  is the specific volume,  $u$  is the velocity, and the pressure  $p = p(v)$  is a given function of  $v$ . For a polytropic isentropic ideal gas, we have

$$p(v) = Av^{-\gamma}$$

for some constants  $A = A(s) > 0$  (depending on the constant entropy) and  $\gamma \geq 1$ .

System (1.7) is of the form

$$\frac{\partial \mathbf{w}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{w}) = \mathbf{0},$$

where

$$\mathbf{w} = \begin{pmatrix} v \\ u \end{pmatrix}, \quad \mathbf{f}(\mathbf{w}) = \begin{pmatrix} -u \\ p(v) \end{pmatrix}, \quad \Omega = \{(v, u) \in \mathbb{R}^2; v > 0\}.$$

The Jacobian matrix of  $\mathbf{f}$  is easily computed

$$\mathbf{A}(\mathbf{w}) = \begin{pmatrix} 0 & -1 \\ p'(v) & 0 \end{pmatrix},$$

and the system (1.7) is strictly hyperbolic provided that we assume  $p'(v) < 0$ . In that case, the matrix  $\mathbf{A}(\mathbf{w})$  (which depends only on  $v$ ) has indeed two real distinct eigenvalues

$$\lambda_1 = -\sqrt{(-p'(v))} < \lambda_2 = \sqrt{(-p'(v))}.$$

Besides the scalar case, the  $p$ -system (1.7) is the simplest nontrivial example of a nonlinear system of conservation laws. Note that any nonlinear wave equation

$$\frac{\partial^2 g}{\partial t^2} - \frac{\partial}{\partial x} \left( \sigma \left( \frac{\partial g}{\partial x} \right) \right) = 0$$

can be put in the form (1.7) by setting

$$u = \frac{\partial g}{\partial t}, \quad v = \frac{\partial g}{\partial x}, \quad p(v) = -\sigma(v),$$

as one can easily check.  $\square$

## 2 Fluid Systems in Eulerian and Lagrangian Frames

System (1.7) is the simplest model for a fluid system in Lagrangian coordinates. We now consider the full system of gas dynamics, an example that appears to be fundamental in the applications, without assuming the flow is isentropic, and in the general multidimensional case.

*Example 2.1. The equations of gas dynamics in Eulerian coordinates.* In Eulerian coordinates, the Euler equations for a compressible inviscid fluid (where we neglect heat conduction) can be written in the following conservative form:

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \sum_{j=1}^3 \frac{\partial}{\partial x_j} (\rho u_j) = 0, \\ \frac{\partial}{\partial t} (\rho u_i) + \sum_{j=1}^3 \frac{\partial}{\partial x_j} (\rho u_i u_j + p \delta_{ij}) = 0, \quad 1 \leq i \leq 3, \\ \frac{\partial}{\partial t} (\rho e) + \sum_{j=1}^3 \frac{\partial}{\partial x_j} ((\rho e + p) u_j) = 0. \end{array} \right. \quad (2.1)$$

In (2.1),  $\rho$  is the density of the fluid,  $\mathbf{u} = (u_1, u_2, u_3)^T$  the velocity,  $p$  the pressure,  $e$  the specific (i.e., per unit mass) internal energy, and  $e = \varepsilon + \frac{|\mathbf{u}|^2}{2}$  the specific total energy. Equations (2.1) express respectively the laws of conservation of mass, momentum, and total energy for the fluid.

Note that (2.1) may be equivalently written

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \\ \frac{\partial}{\partial t} (\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + p \mathbf{I}) = \mathbf{0}, \\ \frac{\partial}{\partial t} (\rho e) + \nabla \cdot ((\rho e + p) \mathbf{u}) = 0 \end{array} \right.$$

where  $\mathbf{I}$  is the  $d \times d$  identity matrix and  $\mathbf{u} \otimes \mathbf{u}$  denotes the matrix-valued function  $(\mathbf{u} \otimes \mathbf{u})_{ij} = u_i u_j$ . More generally, for  $\mathbf{v} \in \mathbb{R}^p$ ,  $\mathbf{w} \in \mathbb{R}^q$ , we define the  $p \times q$  matrix  $\mathbf{v} \otimes \mathbf{w}$  by

$$(\mathbf{v} \otimes \mathbf{w})_{ij} = v_i w_j, \quad 1 \leq i \leq p, 1 \leq j \leq q.$$

Note also that

$$\nabla \cdot (p \mathbf{I}) = \nabla p$$

where  $\nabla$  is the gradient operator.

We need to add to (2.1) an equation for  $p$ . We shall see in Chap. V that Galilean invariance requires that  $p$  does not depend on  $\mathbf{u}$ . The “equation of state” can be taken in the form

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

e.g., for a polytropic ideal gas, the equation of state is given by

$$p = (\gamma - 1)\rho \varepsilon, \quad \gamma > 1.$$

Now, setting

$$q_i = \rho u_i, \quad 1 \leq i \leq 3, \quad E = \rho e,$$

the system (2.1) can be put into the general framework of system (1.1) if we take

$$\mathbf{U} = \begin{pmatrix} \rho \\ q_1 \\ q_2 \\ q_3 \\ E \end{pmatrix}, \quad \mathbf{f}_1(\mathbf{U}) = \begin{pmatrix} q_1 \\ p + q_1^2/\rho \\ q_1 q_2/\rho \\ q_1 q_3/\rho \\ (E + p)q_1/\rho \end{pmatrix}, \quad (2.2a)$$

$$\mathbf{f}_2(\mathbf{U}) = \begin{pmatrix} q_2 \\ q_1 q_2/\rho \\ p + q_2^2/\rho \\ q_2 q_3/\rho \\ (E + p)q_2/\rho \end{pmatrix}, \quad \mathbf{f}_3(\mathbf{U}) = \begin{pmatrix} q_3 \\ q_1 q_3/\rho \\ q_2 q_3/\rho \\ p + q_3^2/\rho \\ (E + p)q_3/\rho \end{pmatrix} \quad (2.2b)$$

with

$$p = p\left(\rho, \frac{E}{\rho} - \frac{|\mathbf{q}|^2}{2\rho^2}\right)$$

and if the set of states is

$$\Omega = \left\{ (\rho, \mathbf{q} = (q_1, q_2, q_3), E); \rho > 0, \mathbf{q} \in \mathbb{R}^3, E - \frac{|\mathbf{q}|^2}{2\rho} > 0 \right\}.$$

We shall see that in the one-dimensional case (resp. bidimensional), the system (2.2a), (2.2b) of the gas dynamics equations with usual equation of state is indeed a strictly hyperbolic (resp. hyperbolic) symmetrizable nonlinear system of conservation laws.

In many applications, we do not have to solve the full system but only a somewhat reduced system of equations. On the one hand, one can often suppose, for instance, that the flow is barotropic so that the equation of state reduces to

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

and it suffices to solve the system of equations for conservation of mass and momentum. On the other hand, if we assume that the flow has some symmetry, one can reduce the number of space variables. For instance, assuming slab symmetry, the gas dynamics equations become (using obvious notations)

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = 0. \end{array} \right. \quad (2.3)$$

Again, choosing  $\rho, q = \rho u$ , and  $E = \rho e$  as dependent variables, (2.3) fits into our framework with  $d = 1, p = 3$ .

More generally, if we only assume that the flow is invariant with respect to  $(x_2, x_3)$ , the gas dynamics equations become with obvious notations

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) = 0, \\ \frac{\partial}{\partial t}(\rho w) + \frac{\partial}{\partial x}(\rho uw) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = 0. \end{array} \right. \quad (2.4)$$

Again (2.4) fits into our framework with  $d = 1, p = 5$ .  $\square$

*Example 2.2. The equations of ideal magnetohydrodynamics (MHD) in Eulerian coordinates.* The MHD equations model the flow of a conducting fluid in the presence of a magnetic field. The ideal MHD equations are obtained when we neglect displacement current, electrostatic forces, viscosity, and heat conduction. They read using convenient units

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \\ \frac{\partial}{\partial t}(\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + p^* \mathbf{I} - \frac{1}{\mu} \mathbf{B} \otimes \mathbf{B}) = \mathbf{0}, \\ \frac{\partial \mathbf{B}}{\partial t} + \nabla \cdot (\mathbf{B} \otimes \mathbf{u} - \mathbf{u} \otimes \mathbf{B}) = \mathbf{0}, \\ \frac{\partial}{\partial t}(\rho e^*) + \nabla \cdot ((\rho e^* + p^*) \mathbf{u} - \frac{1}{\mu} (\mathbf{u} \cdot \mathbf{B}) \mathbf{B}) = 0 \end{array} \right. \quad (2.5)$$

with the additional requirement

$$\nabla \cdot \mathbf{B} = 0. \quad (2.6)$$

In (2.5),  $\mathbf{B} = (B_1, B_2, B_3)^T$  is the magnetic field;  $\mu$  is the resistivity of the fluid (supposed to be constant),  $p^* = p + \frac{1}{2\mu}|\mathbf{B}|^2$  is the total pressure, sum of the static pressure  $p$  and the magnetic pressure  $\frac{1}{2\mu}|\mathbf{B}|^2$ ; and  $e^* = \varepsilon + \frac{1}{2}|\mathbf{u}|^2 + \frac{1}{2\mu\rho}|\mathbf{B}|^2$  is the total specific energy.

It is a simple matter to check that

$$\nabla \cdot (\nabla \cdot (\mathbf{B} \otimes \mathbf{u} - \mathbf{u} \otimes \mathbf{B})) = \mathbf{0}$$

so that, by the third equation (2.5), the constraint (2.6) is automatically satisfied as soon it is satisfied at time  $t = 0$ . In fact, the third equation (2.5) is usually written in the form

$$\frac{\partial \mathbf{B}}{\partial t} + \nabla \times (\mathbf{B} \times \mathbf{u}) = \mathbf{0}$$

which is equivalent under the constraint (2.6) (the expression  $\nabla \times (\mathbf{B} \times \mathbf{u})$  is also classically noted  $\text{curl}(\mathbf{B} \wedge \mathbf{u})$ ).

Again the ideal MHD system can be put into the framework of system (1.1) if we set

$$\mathbf{U} = \begin{pmatrix} \rho \\ q_1 \\ q_2 \\ q_3 \\ B_1 \\ B_2 \\ B_3 \\ E^* \end{pmatrix}, \quad \mathbf{f}_1(\mathbf{U}) = \begin{pmatrix} q_1 \\ p + q_1^2/\rho + p^* - \frac{1}{\mu}B_1^2 \\ q_1q_2/\rho - \frac{1}{\mu}B_1B_2 \\ q_1q_3/\rho - \frac{1}{\mu}B_1B_3 \\ 0 \\ B_2u_1 - B_1u_2 \\ B_3u_1 - B_1u_3 \\ (E^* + p^*)q_1/\rho - \frac{1}{\mu\rho}(\mathbf{q} \cdot \mathbf{B})B_1 \end{pmatrix}, \dots$$

with

$$p^* = p(\rho, \frac{E^*}{\rho} - \frac{|\mathbf{q}|^2}{2\rho^2} - \frac{|\mathbf{B}|^2}{2\mu\rho}) + \frac{|\mathbf{B}|^2}{2\mu}.$$

The set of states  $\Omega$  is then defined as

$$\Omega = \left\{ (\rho, \mathbf{q}, \mathbf{B}, E^*); \rho > 0, \mathbf{q} \in \mathbb{R}^3, \mathbf{B} \in \mathbb{R}^3, E^* - \frac{|\mathbf{q}|^2}{2\rho} - \frac{|\mathbf{B}|^2}{2\mu} > 0 \right\}.$$

If we assume that the flow is invariant with respect to  $(x_2, x_3)$ , the ideal MHD equations read (using obvious notations)

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p^*) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho u v - \frac{B_x B_y}{\mu}) = 0, \\ \frac{\partial}{\partial t}(\rho w) + \frac{\partial}{\partial x}(\rho u w - \frac{B_x B_z}{\mu}) = 0, \\ \frac{\partial B_y}{\partial t} + \frac{\partial}{\partial x}(B_y u - B_x v) = 0, \\ \frac{\partial B_z}{\partial t} + \frac{\partial}{\partial x}(B_z u - B_x w) = 0, \\ \frac{\partial}{\partial t}(\rho e^*) + \frac{\partial}{\partial x}((\rho e^* + p^*)u - \frac{B_x}{\mu}(B_x u + B_y v + B_z w)) = 0, \\ B_x = \text{constant}. \end{array} \right. \quad (2.7)$$

We shall show in the next chapter that the ideal MHD system is indeed hyperbolic (but not strictly hyperbolic) and symmetrizable in the whole set of states  $\Omega$ .  $\square$

More generally, a number of nonlinear hyperbolic systems of conservation laws arising in fluid mechanics can be written in the form

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \\ \frac{\partial}{\partial t}(\rho \Phi) + \nabla \cdot (\rho \Phi \otimes \mathbf{u} + \mathbf{g}(\rho, \Phi)) = \rho \mathbf{s}(\rho, \Phi) \end{array} \right. \quad (2.8)$$

where

$$\Phi = \begin{pmatrix} \Phi_1 \\ \vdots \\ \Phi_{p-1} \end{pmatrix}, \quad \mathbf{g} = (g_{ij})_{1 \leq i \leq p-1, 1 \leq j \leq 3}.$$

In (2.8),  $\rho$  is again the mass density of the fluid,  $\mathbf{u}$  is its velocity while  $\rho \Phi$  stands for the other conservative variables, and  $\mathbf{s}$  is a source term. Note that, in (2.8), we have distinguished the mass conservation equation from the other conservation laws. Indeed, the equations of gas dynamics and the ideal MHD equations fit into the framework of such a general system (2.8) with

$$\Phi = \begin{pmatrix} \mathbf{u} \\ e \end{pmatrix}, \quad \mathbf{g}(\rho, \Phi) = \begin{pmatrix} p \mathbf{I} \\ p \mathbf{u} \end{pmatrix} \quad (2.9)$$

and

$$\Phi = \begin{pmatrix} \mathbf{u} \\ \mathbf{B} \\ \rho \\ e^* \end{pmatrix}, \quad \mathbf{g}(\rho, \Phi) = \begin{pmatrix} p^* \mathbf{I} - \frac{1}{\mu} \mathbf{B} \otimes \mathbf{B} \\ -\mathbf{u} \otimes \mathbf{B} \\ p^* \mathbf{u} - \frac{1}{\mu} (\mathbf{u} \cdot \mathbf{B}) \mathbf{B} \end{pmatrix} \quad (2.10)$$

respectively, and  $\mathbf{s} = \mathbf{0}$ .

Let us now write the system (2.8) using Lagrangian coordinates. Consider the differential system

$$\frac{d\mathbf{x}}{dt} = \mathbf{u}(\mathbf{x}, t) \quad (2.11)$$

and, for all  $\xi \in \mathbb{R}^3$ , we denote by  $t \rightarrow \mathbf{x}(\xi, t)$  the solution of (2.11) that satisfies the initial condition

$$\mathbf{x}(\xi, 0) = \xi. \quad (2.12)$$

Then  $(\xi = (\xi_1, \xi_2, \xi_3), t)$  are the Lagrangian coordinates associated with the velocity field  $\mathbf{u}$ . If we set

$$J(\xi, t) = \det \left( \frac{\partial x_i}{\partial \xi_j}(\xi, t) \right), \quad (2.13)$$

it is a standard exercise to show that

$$\frac{\partial J}{\partial t}(\xi, t) = J(\xi, t) (\nabla \cdot \mathbf{u})(\mathbf{x}(\xi, t), t). \quad (2.14)$$

Now, given a function  $\varphi = \varphi(\mathbf{x}, t)$  expressed in Eulerian coordinates, we denote by  $\bar{\varphi} = \bar{\varphi}(\xi, t)$  this function expressed in Lagrangian coordinates, i.e.,

$$\bar{\varphi}(\xi, t) = \varphi(\mathbf{x}(\xi, t), t).$$

Then, using (2.14), an easy computation shows that

$$\frac{\partial}{\partial t} (\bar{\varphi} J) = J \left( \frac{\partial \bar{\varphi}}{\partial t} + \overline{\nabla \cdot (\varphi \mathbf{u})} \right) = J \left( \frac{\partial \bar{\varphi}}{\partial t} + \sum_{j=1}^3 \overline{\frac{\partial}{\partial x_j} (\varphi u_j)} \right).$$

Hence, the system (2.8) becomes

$$\begin{cases} \frac{\partial}{\partial t} (\bar{\rho} J) = 0, \\ \frac{\partial}{\partial t} (\bar{\rho} \Phi J) + J \overline{\nabla \cdot \mathbf{g}} = \mathbf{0}. \end{cases} \quad (2.15)$$

Let us again assume that the flow is invariant with respect to  $(x_2, x_3)$  so that (2.8) reads

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u), \\ \frac{\partial}{\partial t}(\rho \Phi) + \frac{\partial}{\partial x}(\rho \Phi u + \mathbf{g}) = J \rho s \end{cases} \quad (2.16)$$

and (2.15) becomes

$$\begin{cases} \frac{\partial}{\partial t}(\bar{\rho} J) = 0, \\ \frac{\partial}{\partial t}(\bar{\rho} \Phi J) + J \frac{\partial \mathbf{g}}{\partial x} = J \bar{\rho} s. \end{cases} \quad (2.17)$$

Since

$$\frac{\partial}{\partial \xi} = \left( \frac{\partial x}{\partial \xi} \right) \frac{\partial}{\partial x} = J \frac{\partial}{\partial x},$$

we obtain, suppressing the bars for simplicity

$$\begin{cases} \frac{\partial}{\partial t}(\rho J) = 0, \\ \frac{\partial}{\partial t}(\rho \Phi J) + J \frac{\partial \mathbf{g}}{\partial \xi} = J \rho s, \quad \mathbf{g} = \mathbf{g}(\rho, \Phi). \end{cases} \quad (2.18)$$

Let us put this system in a more classical form. The first equation (of mass conservation) gives

$$\rho J = \rho_0,$$

where  $\rho_0(\xi) = \rho(\xi, 0)$ . If we introduce the specific volume

$$\tau = \frac{1}{\rho},$$

we get

$$J = \rho_0 \tau$$

so that the one-dimensional analogue of (2.14)

$$\frac{\partial J}{\partial t} = J \frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi}$$

becomes

$$\rho_0 \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial \xi} = 0,$$

and the second equation (2.15) gives

$$\rho_0 \frac{\partial \Phi}{\partial t} + \frac{\partial \mathbf{g}}{\partial \xi} = \rho_0 \mathbf{s}.$$

Finally, we introduce a mass variable  $m$  such that

$$dm = \rho_0 d\xi.$$

Then the system (2.16) written in Lagrangian coordinates takes the simple form

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial \Phi}{\partial t} + \frac{\partial g}{\partial m} = s, \end{cases} \quad (2.19)$$

with  $s = s(1/\tau, \Phi)$ .

*Remark 2.1.* Let us have a look at the two-dimensional case. We now write (2.8) in the form (without source for simplicity)

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho \Phi) + \frac{\partial}{\partial x}(\rho u \Phi + g(\rho, \Phi)) + \frac{\partial}{\partial y}(\rho v \Phi + h(\rho, \Phi)) = \mathbf{0} \end{cases}$$

with obvious notations, and it becomes (see (2.15))

$$\begin{cases} \frac{\partial}{\partial t}(\bar{\rho} J) = 0, \\ \frac{\partial}{\partial t}(\bar{\rho} \bar{\Phi} J) + J \left( \frac{\partial \bar{g}}{\partial x} + \frac{\partial \bar{h}}{\partial y} \right) = \mathbf{0}. \end{cases}$$

In order to obtain a more useful form, denoting by  $(\xi, \eta, t)$  the Lagrangian coordinates, we express the derivatives  $\frac{\partial \bar{g}}{\partial x}$  and  $\frac{\partial \bar{h}}{\partial y}$  in terms of derivatives wrt.  $(\xi, \eta)$ . We use the fact that the  $2 \times 2$  Jacobian matrix of the map  $(x, y) \mapsto (\xi, \eta)$  is the inverse of the Jacobian matrix of the map  $(\xi, \eta) \mapsto (x, y)$  which is easily computed, and together with the definition (2.13) of  $J$ , we get

$$\begin{aligned} J \left( \frac{\partial \bar{g}}{\partial x} + \frac{\partial \bar{h}}{\partial y} \right) &= J \left( \frac{\partial \bar{g}}{\partial \xi} \frac{\partial \xi}{\partial x} + \frac{\partial \bar{g}}{\partial \eta} \frac{\partial \eta}{\partial x} + \frac{\partial \bar{h}}{\partial \xi} \frac{\partial \xi}{\partial y} + \frac{\partial \bar{h}}{\partial \eta} \frac{\partial \eta}{\partial y} \right) \\ &= \frac{\partial \bar{g}}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial \bar{g}}{\partial \eta} \frac{\partial y}{\partial \xi} - \frac{\partial \bar{h}}{\partial \xi} \frac{\partial x}{\partial \eta} + \frac{\partial \bar{h}}{\partial \eta} \frac{\partial x}{\partial \xi} \\ &= \frac{\partial}{\partial \xi} \left( \bar{g} \frac{\partial y}{\partial \eta} - \bar{h} \frac{\partial x}{\partial \eta} \right) + \frac{\partial}{\partial \eta} \left( -\bar{g} \frac{\partial y}{\partial \xi} + \bar{h} \frac{\partial x}{\partial \xi} \right). \end{aligned}$$

Setting  $a = \frac{\partial x}{\partial \xi}, b = \frac{\partial y}{\partial \xi}, c = \frac{\partial x}{\partial \eta}, d = \frac{\partial y}{\partial \eta}$ , we have  $J = ad - bc$ , and we notice that the system in Lagrangian coordinates can be equivalently written (dropping the bars)

$$\left\{ \begin{array}{l} \rho_0 \frac{\partial \tau}{\partial t} - \frac{\partial}{\partial \xi}(ud - vc) - \frac{\partial}{\partial \eta}(-ub + va) = 0, \\ \rho_0 \frac{\partial \Phi}{\partial t} + \frac{\partial}{\partial \xi}(\mathbf{g}d - \mathbf{h}c) + \frac{\partial}{\partial \eta}(-\mathbf{g}b + \mathbf{h}a) = \mathbf{0}, \\ \frac{\partial}{\partial t} \begin{pmatrix} a \\ b \end{pmatrix} - \frac{\partial}{\partial \xi} \begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{0}, \\ \frac{\partial}{\partial t} \begin{pmatrix} c \\ d \end{pmatrix} - \frac{\partial}{\partial \eta} \begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{0}. \end{array} \right.$$

The three-dimensional case can be handled in a similar way. For a deeper analysis, we refer to [639] and [416, 419].  $\square$

*Example 2.3. The equations of gas dynamics in Lagrangian coordinates.* If we choose in (2.19)

$$\Phi = \begin{pmatrix} u \\ v \\ w \\ e \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} p \\ 0 \\ 0 \\ pu \end{pmatrix}$$

we obtain the system of gas dynamics (2.4) which reads in Lagrangian coordinates

$$\left\{ \begin{array}{l} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial v}{\partial t} = 0, \\ \frac{\partial w}{\partial t} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial(pu)}{\partial m} = 0 \end{array} \right. \quad (2.20)$$

with  $p = p(\tau, \varepsilon) = p(\tau, e - \frac{u^2}{2})$ . By setting

$$\mathbf{V} = \begin{pmatrix} \tau \\ u \\ v \\ w \\ e \end{pmatrix}, \quad \mathbf{f}(\mathbf{V}) = \begin{pmatrix} -u \\ p \\ 0 \\ 0 \\ pu \end{pmatrix}, \quad \Omega = \left\{ \mathbf{V}; \tau > 0, \mathbf{u} = (u, v, w) \in \mathbb{R}^3, e - \frac{|\mathbf{u}|^2}{2} > 0 \right\},$$

we obtain a nonlinear system of conservation laws

$$\frac{\partial \mathbf{V}}{\partial t} + \frac{\partial}{\partial m} \mathbf{f}(\mathbf{V}) = \mathbf{0} \quad (2.21)$$

that again will be seen to be strictly hyperbolic and symmetrizable under natural physical assumptions. On the other hand, the system (2.3) written

in Lagrangian coordinates is of the form (2.21) with

$$\mathbf{V} = \begin{pmatrix} \tau \\ u \\ e \end{pmatrix}, \quad \mathbf{f}(\mathbf{V}) = \begin{pmatrix} -u \\ p \\ pu \end{pmatrix}.$$

If we assume in addition that the flow is barotropic, the equation of state reduces to  $p = p(\tau)$ . Again, it suffices to solve the system of equations of conservation of mass and momentum

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0. \end{cases}$$

We obtain the  $p$ -system (1.7) but with different notations.  $\square$

*Example 2.4. The ideal MHD equations in Lagrangian coordinates.* If we choose in (2.19) the pair  $(\Phi, \mathbf{g}(\rho, \Phi))$  as

$$\Phi = \begin{pmatrix} u \\ v \\ w \\ \tau B_y \\ \tau B_z \\ e \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} p^* \\ -\frac{B_x}{\mu} B_y \\ -\frac{B_x}{\mu} B_z \\ -B_x v \\ -B_x w \\ p^* u - \frac{B_x}{\mu} (B_x u + B_y v + B_z w) \end{pmatrix},$$

we obtain the ideal MHD equations which become in Lagrangian coordinates

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p^*}{\partial m} = 0, \\ \frac{\partial v}{\partial t} - \frac{\partial}{\partial m} \left( \frac{B_x}{\mu} B_y \right) = 0, \\ \frac{\partial w}{\partial t} - \frac{\partial}{\partial m} \left( \frac{B_x}{\mu} B_z \right) = 0, \\ \frac{\partial}{\partial t} (\tau B_y) - \frac{\partial}{\partial m} (B_x v) = 0, \\ \frac{\partial}{\partial t} (\tau B_z) - \frac{\partial}{\partial m} (B_x w) = 0, \\ \frac{\partial e_*}{\partial t} + \frac{\partial}{\partial m} \left( p^* u - \frac{B_x}{\mu} (B_x u + B_y v + B_z w) \right) = 0 \end{cases} \quad (2.22)$$

where

$$p^* = p(\tau, \varepsilon) + \frac{|\mathbf{B}|^2}{2\mu}, \quad e^* = \varepsilon + \frac{1}{2}(u^2 + v^2 + w^2) + \frac{\tau|\mathbf{B}|^2}{2\mu}.$$

Again this system is of the form (2.21)  $\square$

In fact, in many applications, we need to consider systems of conservation laws with source terms

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) = \mathbf{s}(\mathbf{u}), \quad \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d, \quad t > 0, \quad (2.23)$$

where  $\mathbf{s}$  is a smooth function from  $\Omega$  into  $\mathbb{R}^p$ ,  $\mathbf{s}$  may also depend on  $\mathbf{x}$ . This source term may be present in different situations; it may be linked to external forces, such as gravity, and to the presence of chemical reaction; or it may have a geometric nature. Let us give below several examples of such situations.

*Example 2.5. The system of gas dynamics with gravity and friction.* If we take into account both gravity and friction terms, the Euler system of gas dynamics becomes

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \\ \frac{\partial}{\partial t}(\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + p \mathbf{I}) = \rho(\mathbf{g} - \alpha \boldsymbol{\varphi}(\mathbf{u})), \\ \frac{\partial}{\partial t}(\rho e) + \nabla \cdot ((\rho e + p)\mathbf{u}) = \rho(\mathbf{g} \cdot \mathbf{u} - \alpha \psi(\mathbf{u})) \end{array} \right. \quad (2.24)$$

where  $\mathbf{g}$  stands for the gravity force and  $\alpha > 0$  is a friction constant coefficient. The functions  $\boldsymbol{\varphi}$  and  $\psi$  are typically of the forms

$$\boldsymbol{\varphi}(\mathbf{u}) = |\mathbf{u}|^\chi \mathbf{u}, \quad \psi(\mathbf{u}) = a|\mathbf{u}|^{\chi+2}$$

where  $\chi = 0$  for a linear friction,  $\chi = 1$  for a quadratic one, and  $a > 0$  is a constant.

Using the notations of Example 2.1, we have here

$$\mathbf{s}(\mathbf{U}) = \begin{pmatrix} 0 \\ \rho(\mathbf{g} - \alpha \boldsymbol{\varphi}(\frac{\mathbf{q}}{\rho})) \\ \mathbf{g} \cdot \mathbf{q} - \alpha \rho \psi(\frac{\mathbf{q}}{\rho}) \end{pmatrix}.$$

Assuming slab symmetry, we obtain the system

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = \rho(g - \alpha\varphi(u)), \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = \rho(gu - \alpha\psi(u)). \end{cases} \quad (2.25)$$

Passing in Lagrangian coordinates, (2.25) becomes

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial u}{\partial m} = g - \alpha\varphi(u), \\ \frac{\partial e}{\partial t} + \frac{\partial}{\partial m}(pu) = gu - \alpha\psi(u). \end{cases} \quad (2.26)$$

□

*Example 2.6. The equations of reacting gas flows.* Consider the more complex example of a system of reacting gas flows. We assume that the system is multicomponent, consisting of  $n$  reacting species. If we neglect again viscosity and heat conduction, the equations of the flow are obtained by adjoining to the Euler system of gas dynamics the equations governing the chemical reactions between species

$$\frac{\partial}{\partial t}(\rho Y_i) + \nabla \cdot (\rho Y_i \mathbf{u}) = m_i \omega_i, \quad 1 \leq i \leq n-1. \quad (2.27)$$

In (2.27),  $Y_i$  is the mass fraction;  $m_i$  is the molar mass, also termed the molecular weight, of the  $i$ th species; and  $\omega_i$  is the molar production rate of the  $i$ th species with

$$\sum_{i=1}^n Y_i = 1, \quad \sum_{i=1}^n m_i \omega_i = 0, \quad (2.28)$$

so that we have replaced the equation of conservation of the  $n$ th species by the total mass conservation which is part of the Euler system.

In order to close the system of equations, we need an equation of state which can be taken in the form

$$\varepsilon = \varepsilon(\rho, p, \mathbf{Y}), \quad \mathbf{Y} = (Y_1, \dots, Y_{n-1})^T \quad (2.29)$$

and expressions for the production rates

$$m_i \omega_i = m_i \omega_i(\rho, p, \mathbf{Y}), \quad 1 \leq i \leq n-1. \quad (2.30)$$

For example, if we suppose that each species behaves as a polytropic ideal gas with adiabatic exponent  $\gamma_i$ , we have

$$\varepsilon = \sum_{i=1}^n Y_i h_i^0 + \frac{1}{\gamma(\mathbf{Y}) - 1} \frac{p}{\rho}$$

where  $h_i^0$  is the enthalpy of formation of the species  $i$  at 0°K and

$$\gamma(\mathbf{Y}) = \frac{\sum_{i=1}^n Y_i C_{p,i}}{\sum_{i=1}^n Y_i C_{v,i}}$$

$C_{v,i}$  (resp.  $C_{p,i}$ ) denotes its specific heat at constant volume (resp. at constant pressure). For more details on ideal gases, see Chap. III, Sect. 1.2. If in addition  $\gamma_i = \frac{C_{p,i}}{C_{v,i}}$  is independent of  $i$ , i.e., all the species have the same adiabatic exponent  $\gamma$ , we obtain  $\gamma(\mathbf{Y}) = \gamma$ .

Now, going back to the general case, we set

$$\rho \mathbf{r} = (m_1 \omega_1, \dots, m_{n-1} \omega_{n-1})^T$$

and

$$\mathbf{W} = \begin{pmatrix} \mathbf{U} \\ \rho \mathbf{Y} \end{pmatrix}, \quad \mathbf{g}_j = \begin{pmatrix} \mathbf{f}_j \\ q_j \mathbf{Y} \end{pmatrix}, 1 \leq j \leq 3, \quad \mathbf{s} = \begin{pmatrix} \mathbf{0} \\ \rho \mathbf{r} \end{pmatrix}$$

where we have used the notations (2.2a), (2.2b) of Example 2.1. We obtain the system

$$\frac{\partial \mathbf{W}}{\partial t} + \nabla \cdot \mathbf{g} = \mathbf{s}$$

which is indeed of the form (2.23) (note that we may write the flux with  $\mathbf{Y} = \frac{\rho \mathbf{Y}}{\varrho}$ , i.e., in terms of the conservative variables).

Assuming slab symmetry, the equations of reacting gas flows become

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x} (\rho u) = 0, \\ \frac{\partial}{\partial t} (\rho u) + \frac{\partial}{\partial x} (\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t} (\rho e) + \frac{\partial}{\partial x} ((\rho e + p) u) = 0, \\ \frac{\partial}{\partial t} (\rho \mathbf{Y}) + \frac{\partial}{\partial x} (\rho \mathbf{Y} u) = \rho \mathbf{r}. \end{array} \right. \quad (2.31)$$

In Lagrangian coordinates, the last equation (2.31) gives

$$\frac{\partial}{\partial t} (\overline{\rho \mathbf{Y}} J) = J \overline{\rho \mathbf{r}}$$

or equivalently by the mass conservation equation

$$\frac{\partial \bar{\mathbf{Y}}}{\partial t} = \bar{\mathbf{r}}.$$

Hence, (2.31) gives in Lagrangian coordinates (dropping the bars)

$$\left\{ \begin{array}{l} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial (pu)}{\partial m} = 0, \\ \frac{\partial \mathbf{Y}}{\partial t} = \mathbf{r}. \end{array} \right. \quad (2.32)$$

We next restrict ourselves to the simple case where we consider only two species, the unburnt gas and the burnt gas. We denote by  $z$  the mass fraction of the burnt gas so that  $1 - z$  is the mass fraction of the unburnt gas. The simplest model is obtained by assuming that on the one hand the unburnt gas is converted to the burnt gas through an irreversible chemical reaction with a production rate

$$r = r(\rho, p, z) \quad (2.33)$$

and on the other hand that both unburnt and burnt gases behave as polytropic ideal gases with the same adiabatic exponent  $\gamma$  so that the equation of state reads

$$\varepsilon(\rho, p, z) = zh_b^0 + (1 - z)h_u^0 + \frac{1}{\gamma - 1} \frac{p}{\rho}.$$

Observe that  $h_b^0 < h_u^0$  for an exothermic reaction. In this simple case, (2.27) resumes to

$$\frac{\partial}{\partial t}(\rho z) + \nabla \cdot (\rho z \mathbf{u}) = \rho r.$$

If we assume that the kinetics of the chemical reaction obeys an Arrhenius mechanism, we have

$$r = K(1 - z) \exp\left(-\frac{E^*}{RT}\right),$$

where  $E^*$  is the activation energy and  $T$  is the temperature. In practice, a simpler expression for  $r$  is often used, for instance,

$$r = K(1 - z)^\alpha,$$

with  $\alpha = \frac{1}{2}$  or 1. □

One can even consider systems of conservation laws more general than (2.23), namely, systems of the form

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot \mathbf{f}(\mathbf{x}, t, \mathbf{u}) = \mathbf{s}(\mathbf{x}, t, \mathbf{u}), \quad \mathbf{x} \in \mathbb{R}^d, \quad t > 0, \quad (2.34)$$

where the flux function  $\mathbf{f} = (\mathbf{f}_j)_{1 \leq j \leq d}$  and the source term  $\mathbf{s}$  are smooth functions from  $\mathbb{R}^d \times [0, +\infty[ \times \Omega$  into  $(\mathbb{R}^p)^d$  and  $\mathbb{R}^p$ , respectively.

### 3 Some Averaged Models: Shallow Water, Flow in a Duct, and Two-Phase Flow

The examples we consider now are obtained by averaging a given complete fluid model, the *incompressible* Navier-Stokes system on a column of water for the Saint-Venant system, respectively the Euler system on a section for the flow in a duct (or a nozzle). The corresponding models necessarily result from some simplifying assumption, in order that the averaged quantities give a good description of the flow. We will see that the averaging procedure leads to systems with a geometric source term (depending on  $x$ ), so that they do not benefit from a useful Lagrangian description. We begin by stating a classical lemma (Leibniz integral rule), which is involved in the derivation of the following models and explains why the averaging procedure leads in general to nonconservative terms.

*Lemma 3.1*

Let  $\alpha, \beta$  (resp.  $\varphi$ ) be sufficiently smooth real functions on  $\mathbb{R}^2$  (resp.  $\mathbb{R}^3$ ). We have the identity, for  $i = 1, 2$ ,

$$\int_{\alpha(x_1, x_2)}^{\beta(x_1, x_2)} \frac{\partial}{\partial x_i} \varphi(x_1, x_2, z) dz = \frac{\partial}{\partial x_i} \int_{\alpha(x_1, x_2)}^{\beta(x_1, x_2)} \varphi(x_1, x_2, z) dz - \left( \frac{\partial \beta}{\partial x_i} \varphi(x_1, x_2, \beta(x_1, x_2)) - \frac{\partial \alpha}{\partial x_i} \varphi(x_1, x_2, \alpha(x_1, x_2)) \right).$$

*Example 3.1.* The Saint-Venant system. It models shallow water with topography and reads

$$\begin{cases} \frac{\partial h}{\partial t} + \nabla \cdot (h \mathbf{u}) = 0, \\ \frac{\partial}{\partial t} (h \mathbf{u}) + \nabla \cdot (h \mathbf{u} \otimes \mathbf{u}) + \frac{g}{2} \nabla h^2 = -gh \nabla Z, \end{cases} \quad (3.1)$$

where  $h = h(x, y, t)$  denotes the water height,  $\mathbf{u}(x, y, t) = (u, v)^T(x, y, t)$  is the flow velocity, and  $Z = Z(x, y)$  characterizes the topography. This system, also called the Saint-Venant system [87], writes as the barotropic system of gas dynamics with a source term; the first equation expresses the conservation of mass of a column of water, and in the second equation, the pressure law is taken as  $p(h) = \frac{g}{2} h^2$ .

System (3.1) is obtained by averaging the incompressible Euler system with free boundary on a vertical column of water, from a bottom  $z = Z(x, y)$  to the free surface, say  $z = \eta(x, y, t)$ , with kinematic boundary conditions. Thus,  $h = \eta - Z$  and the velocity is the average  $\bar{\mathbf{u}} = \frac{1}{h} \int_Z^\eta \mathbf{u}(x, y, z, t) dz$ , and we have dropped the bars for simplicity. The hydrostatic approximation amounts to introduce a *small parameter*  $\epsilon = \frac{H}{L}$  where  $H$  and  $L$  are two characteristic dimensions along the vertical and horizontal direction, respectively, to rescale the equations and neglect the vertical acceleration of the fluid (of order  $\mathcal{O}(\epsilon^2)$ ) in the rescaled Euler system. This assumption yields that the pressure is simply the hydrostatic pressure  $\rho_0 gh$ , where  $g$  is the gravitational constant and  $\rho_0$  the (constant) fluid density usually taken equal to 1 for water. The advantage of the averaging procedure is to transform a problem with a free surface into a problem where the height of water becomes a dependent variable, and this explains why it leads to a compressible flow model. Note that the source term is a byproduct of the integration step, resulting from Lemma 3.1 (we do not go into the details of the computation; see [512] and the references therein).

Let us introduce the energy,

$$E(h, \mathbf{u}, Z) = \frac{1}{2}(h|\mathbf{u}|^2 + gh^2) + gZh.$$

Then smooth solutions of (3.1) can be shown to satisfy the energy conservation equation

$$\partial_t E + \operatorname{div}(\mathbf{u}(E + \frac{g}{2}h^2)) = 0.$$

We will see that this means that system (3.1) admits a mathematical entropy, which is the physical energy; in case of a discontinuous solution, it becomes an inequality. Let us also mention that this energy conservation equation can be obtained directly from the averaging of the energy equation in the incompressible Navier-Stokes system together with the shallow water approximation.

If we assume that the flow is one-dimensional, in the direction  $x$ , we get

$$\begin{cases} \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0, \\ \frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}(hu^2 + \frac{g}{2}h^2) = -ghZ'(x). \end{cases} \quad (3.2)$$

A more complete model, including viscosity and friction, can be derived starting from the Navier-Stokes system together with the hydrostatic assumption; we refer to [512]. We will give more references concerning extended shallow water models in Chap. VII concerning source terms.  $\square$

*Example 3.2. Flow in a duct with variable cross section.* We now give another example of a system for which the source term comes from an averaging

procedure (which means here by integrating on a section) together with some symmetry assumption. We consider the flow in a duct with variable section (a nozzle) aligned with the  $x$  axis, with cross section  $S(x)$  smoothly varying with space, and we note the area  $|S(x)| = A(x)$ . By symmetry, we can consider a 2d duct (in the plane  $z = 0$ ), limited by a curve with equation  $y = A(x)/2$  (resp.  $y = -A(x)/2$ ) for the upper (resp. lower part).

In order to derive a simplified PDE model, we start by integrating the system (2.1) on a small streamtube, say  $\omega$ , around the  $x$  axis, between  $x$  and  $x + \delta x$ , with section  $S(\xi), \xi \in [x, x + \delta x]$ . First we define for any quantity  $\varphi$  its average over a section  $S(x)$  noted  $\bar{\varphi}$  by

$$A(x)\bar{\varphi}(x, t) = \int_{S(x)} \varphi(x, y, t) dy.$$

We consider the integral of the first equation of (1.7) and get for the first term

$$\int_{\omega} \partial_t \rho(\mathbf{x}, t) d\mathbf{x} = \partial_t \left( \int_{\omega} \rho(\mathbf{x}, t) d\mathbf{x} \right) = \partial_t \int_x^{x+\delta x} A(\xi) \bar{\rho}(\xi, t) d\xi$$

where  $\bar{\rho}(x, t)$  is the average density over the section  $S(x)$ . Then we write

$$\int_{\omega} \nabla \cdot (\rho \mathbf{u}) d\mathbf{x} = \int_{\partial\omega} \rho \mathbf{u} \cdot \mathbf{n} d\sigma.$$

Now, the surface  $\partial\omega$  is made of the two parallel sections, say  $S_x$  (resp.  $S_{x+\delta x}$ ) with outward normal  $-\mathbf{e}_1$  (resp.  $\mathbf{e}_1$ ), where  $\mathbf{e}_1$  denotes the unit vector along the  $x$  axis, and with length  $A(x)$  (resp.  $A(x+\delta x)$ ) and a lateral surface which we denote  $\Sigma$ , then

$$\int_{\partial\omega} = \int_{S_x} + \int_{S_{x+\delta x}} + \int_{\Sigma}.$$

We assume that the flow propagates in a streamtube; hence,  $\mathbf{u} \cdot \mathbf{n} = 0$  on  $\Sigma$  and the corresponding integral vanishes. The two remaining integrals are easily computed, because  $\mathbf{u} \cdot \mathbf{e}_1 = u$ . Indeed, for any quantity  $\varphi$ , we have

$$\int_{S_x} \varphi \mathbf{u} \cdot \mathbf{n} d\sigma + \int_{S_{x+\delta x}} \varphi \mathbf{u} \cdot \mathbf{n} d\sigma = -A(x)\bar{\varphi}u(x, t) + A(x+\delta x)\bar{\varphi}u(x+\delta x, t).$$

Thus, dividing by  $\delta x$  and letting  $\delta x \rightarrow 0$ , we get an equation for the averaged density

$$\partial_t(A\bar{\rho}) + \partial_x(A\bar{\rho}u) = 0.$$

Similarly, we integrate the first momentum conservation equation, which we can write

$$\partial_t(\rho u) + \nabla \cdot (\rho u \mathbf{u} + p \mathbf{e}_1) = 0.$$

Following the same lines, we obtain the terms  $\partial_t \bar{\rho} \bar{u} + \partial_x (\bar{\rho} u^2 + \bar{p})$ . However, there is now an added contribution coming from the pressure term  $p \mathbf{e}_1$  integrated on the lateral surface  $\Sigma$ . The tangent vector to the upper part of  $\Sigma$  is  $(1, A'(x)/2)^T$  and the normal to the upper part (resp. lower) is  $(-A'(x)/2, 1)^T$  (resp.  $(-1, -A'(x)/2)^T$ ), and we add the equal contributions of the upper and lower parts. Hence, if  $\mathbf{n}$  denotes the unit normal vector (obtained by normalizing the normal vector  $(-A'(x)/2, 1)^T$ ) and  $d\sigma$  is the arc length measure along  $\Sigma$ , we get

$$\int_{\Sigma} p(\mathbf{x}, t) \mathbf{e}_1 \cdot \mathbf{n} d\sigma = \frac{1}{2} \int_x^{x+\delta x} \left( p(\xi, A(\xi)/2, t) + p(\xi, -A(\xi)/2, t) \right) A'(\xi) d\xi.$$

Thus, dividing by  $\delta x$  and letting  $\delta x \rightarrow 0$ , we get a term  $A'(x)\bar{p}$ , where  $\bar{p}$  denotes the mean pressure  $\bar{p}(x, t) = \frac{1}{2}(p(x, A(x)/2, t) + p(x, -A(x)/2, t))$  on  $\Sigma$ . Thus, averaging the momentum conservation equation yields

$$\partial_t (A \bar{\rho} \bar{u}) + \partial_x (A(\bar{\rho} u^2 + \bar{p})) - \bar{p} A'(x) = 0.$$

In the same way, integrating the second momentum conservation equation,

$$\partial_t (\rho v) + \nabla \cdot (\rho v \mathbf{u} + p \mathbf{e}_2) = 0,$$

yields

$$\partial_t (A \bar{\rho} \bar{v}) + \partial_x (A(\bar{\rho} \bar{u} \bar{v})) - \bar{\delta} p = 0,$$

with  $\bar{\delta} p(x, t) = \frac{1}{2}(p(x, A(x)/2, t) - p(x, -A(x)/2, t))$ . Now, if we add the assumption that the flow is symmetric, the average velocity  $\bar{v}$  vanishes, similarly  $\bar{\rho} \bar{v} = 0$ , and also the term corresponding to the pressure  $\int_{\Sigma} p(\mathbf{x}, t) \mathbf{e}_2 \cdot \mathbf{n} d\sigma = 0$ , we can then ignore the second momentum equation.

The energy equation comes as the other ones since we integrate the term  $\nabla \cdot ((\varrho e + p)\mathbf{u})$ , and again, only a term with  $\mathbf{u} \cdot \mathbf{n}$  is involved. This term vanishes on  $\Sigma$  and is easily computed on the cross sections  $S(x), S(x + \delta x)$ .

Let us assume last that both pressure terms  $\bar{\delta} p$  and  $\bar{p}$  can be replaced by  $p(\bar{\rho}, \bar{\varepsilon})$ , and  $\bar{\rho} u^2$  by  $(\bar{\rho} u)^2 / \bar{\rho}$ , then dropping the bars for simplicity, we get the system

$$\begin{cases} \frac{\partial}{\partial t} (A \rho) + \frac{\partial}{\partial x} (A \varrho u) = 0, \\ \frac{\partial}{\partial t} (A \rho u) + \frac{\partial}{\partial x} (A(\rho u^2 + p)) = p A'(x), \\ \frac{\partial}{\partial t} (A \varrho e) + \frac{\partial}{\partial x} (A(\varrho e + p) u) = 0, \end{cases} \quad (3.3)$$

which also looks like the Euler system, provided we define a surfacic density  $r = A \rho$ , with a source term which is generally called a *geometric* source term (when  $A$  is not constant). However, we have to look more precisely at the pressure law given in terms of conservative variables. Assuming for simplicity

that the fluid is barotropic, system (3.3) gives

$$\begin{cases} \frac{\partial}{\partial t}(A\rho) + \frac{\partial}{\partial x}(A\rho u) = 0, \\ \frac{\partial}{\partial t}(A\rho u) + \frac{\partial}{\partial x}(A(\rho u^2 + p(\rho))) = p(\rho)A'(x), \end{cases}$$

which now looks like the barotropic Euler system with a source. Then the pressure term can be given in terms of  $r = A\rho$ ,  $P(r) = Ap(r/A)$ . However, this definition of  $P$  introduces a flux depending on  $x$  which does not simplify the study. In fact, the system can be written equivalently replacing  $A'(x)$  by  $\frac{\partial}{\partial x}A$  and completed by an equation which says that  $A$  does not vary with time. This approach is now classical for a geometric source term (which thus does not depend on  $t$ ) or some systems of conservation laws with flux depending on  $x$ . It yields here

$$\begin{cases} \frac{\partial}{\partial t}(A\rho) + \frac{\partial}{\partial x}(A\rho u) = 0, \\ \frac{\partial}{\partial t}(A\rho u) + \frac{\partial}{\partial x}(A(\rho u^2 + p(\varrho))) - p(\rho)\frac{\partial}{\partial x}A = 0, \\ \frac{\partial}{\partial t}A = 0. \end{cases} \quad (3.4)$$

System (3.4) is a first example of a system with a *nonconservative* product since, when  $A$  is not constant, the term  $p(\rho)\frac{\partial}{\partial x}A$  cannot be put in divergence form.  $\square$

Similarly, system (3.2) can be transformed and completed by an equation  $\frac{\partial}{\partial t}Z = 0$ , and it provides another example of a system with a nonconservative term. We shall come again on this approach later on in Chap. VII and study the corresponding nonconservative systems in more details.

*Example 3.3. Simple models of two-phase flow.* In the context of two-phase flows, a similar averaging approach leads to a model which may be justified for a stratified flow. Assuming the flow evolves between two plates and consists of two inviscid fields separated by a smooth single-valued surface, the averaging procedure allows to reduce a two-dimensional (stratified) flow to a one-dimensional model. Denoting by  $z = \alpha(x, t)$  the interface between the two fluids, one integrates the 2D-Euler system on  $[0, \alpha]$  (resp.  $[\alpha, 1]$ ) for fluid 1 (resp. fluid 2), with equation of state  $p_1(\rho, s)$  (resp.  $p_2(\rho, s)$ ). In this approach,  $\alpha$  plays the role of a volume fraction. In the following, for ease of notation, we will set  $\alpha_1 = \alpha$ ,  $\alpha_2 = 1 - \alpha$ . The resulting system has six equations, but it involves interface quantities, and more modeling assumptions are needed to close the system, in particular assumptions concerning the values of the

velocity and pressure at the interface, among which some simplifications are usually done to lead to a tractable model. We do not go into details for which we refer to [965, 1073], and give only some simple examples.

Assume for simplicity that the flow is isentropic. We first consider the case where the two fields have an *equal pressure*  $p$ . It leads to a four-equation model, which writes, with  $k = 1, 2$  and  $\alpha_1 + \alpha_2 = 1$ ,

$$\begin{cases} \frac{\partial}{\partial t}(\alpha_k \rho_k) + \frac{\partial}{\partial x}(\alpha_k \rho_k u_k) = 0, \\ \frac{\partial}{\partial t}(\alpha_k \rho_k u_k) + \frac{\partial}{\partial x}(\alpha_k \rho_k u_k^2) + \alpha_k \frac{\partial}{\partial x} p + \Delta p \frac{\partial}{\partial x} \alpha_k = M_k, \end{cases} \quad (3.5)$$

with equations of state  $\rho_i = \rho_i(p)$  and where we have noted  $\Delta p = p - p_I$ , a pressure default. Here  $p_I$  denotes an *interfacial* pressure and  $\Delta p$  is some given function of the state  $\mathbf{u} = (\alpha_1, (1 - \alpha)_2, \alpha_1 u_1, (1 - \alpha)_2 u_2)^T$ . In the right-hand side, the  $M_k$ 's are some source terms, satisfying  $M_1 + M_2 = 0$ , which we assume do not contain derivatives of the variables: they model interfacial forces (such as interface drag which is proportional to the relative velocity  $u_r = u_1 - u_2$ ) or external forces (such as gravity). When  $\Delta p = 0$ , we get the so-called one-pressure model. Note that the second equation in (3.5) can also be written

$$\frac{\partial}{\partial t}(\alpha_k \rho_k u_k) + \frac{\partial}{\partial x}(\alpha_k(\rho_k u_k^2 + p)) - p_I \frac{\partial}{\partial x} \alpha_k = M_k,$$

which clearly resembles the second equation obtained in the preceding example modeling the flow in a nozzle, with now  $\alpha_k(x, t)$  unknown and depending also on  $t$ . System (3.5) is nonconservative, and it can be put in the form

$$\frac{\partial}{\partial t}\mathbf{u} + \mathbf{A}(\mathbf{u}) \frac{\partial}{\partial x}\mathbf{u} = \mathbf{0}.$$

However, the resulting model may fail to be hyperbolic, in the sense that  $\mathbf{A}(\mathbf{u})$  does not necessarily have real eigenvalues for all physically acceptable quantities. In particular, the one pressure model ( $\Delta p = 0$ ) is not hyperbolic in general (not in the range of relative velocities in common physical applications), so that, usefully, one builds some “pressure correction” term  $\Delta p$  proportional to  $u_r^2$ , where again  $u_r = u_1 - u_2$  is the relative velocity [896], or a correction term involving derivatives (see [1130]). This system, though simple, gives an example where the computation of the eigenvalues, which are solution of a fourth-degree polynomial, cannot be achieved explicitly, and the result is obtained by perturbation, using some Taylor expansions.

A related model keeps two different pressures, leading to the so-called *five-equation two-pressure* model which writes, ignoring the external forces and transfer terms between the phases,

$$\left\{ \begin{array}{l} \frac{\partial}{\partial t}(\alpha_k \rho_k) + \frac{\partial}{\partial x}(\alpha_k \rho_k u_k) = 0, \quad k = 1, 2, \\ \frac{\partial}{\partial t}(\alpha_k \rho_k u_k) + \frac{\partial}{\partial x}(\alpha_k(\rho_k u_k^2 + p_k)) - p_I \frac{\partial}{\partial x} \alpha_k = 0, \quad k = 1, 2, \\ \frac{\partial}{\partial t} \alpha_1 + u_I \frac{\partial}{\partial x} \alpha_1 = \delta, \end{array} \right. \quad (3.6)$$

with equations of state  $p_i = p_i(\rho_i)$  and where  $u_I, p_I$  are the *interfacial* velocity and pressure which are supposed to be known functions of the state. A frequent choice is  $u_I = u_2, p_I = p_1$ ; the model is often called the (isentropic) model of Baer-Nunziato. The eigenvalues are explicitly computed [965], and the system is hyperbolic (outside resonance). In the last equation, when the source term  $\delta$  is taken proportional to the relative pressure  $p_1 - p_2$ , with a coefficient which can be stiff, one speaks of a *pressure relaxation* term. At least formally, the limit for zero relaxation time forces the equality of pressures  $p_1 = p_2$  and gives the four-equation two-fluid model (3.5); for the full seven-equation model with energy, a hyperbolic five-equation reduced model is derived in [890] via an asymptotic analysis for zero relaxation time (see also [16]).

An *equilibrium* mixture model corresponds to the assumption that the two fluids have equal velocity, and we have only one total momentum equation, obtained by adding the momentum equations of each phase; it writes in the barotropic case as a three-equation model

$$\left\{ \begin{array}{l} \frac{\partial}{\partial t}(\alpha_k \rho_k) + \frac{\partial}{\partial x}(\alpha_k \rho_k u) = 0, \quad k = 1, 2, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \end{array} \right. \quad (3.7)$$

where  $\rho = \sum_{k=1,2} \alpha_k \rho_k$  is the total density, and the system is closed with equations of state for each fluid  $p_k(\rho_k)$  and equating the pressure  $p = p_1(\rho_1) = p_2(\rho_2)$ . We may equivalently replace one equation of conservation of the mass of one phase by the sum of the two equations which writes

$$\frac{\partial}{\partial t} \rho + \frac{\partial}{\partial x} \rho u = 0.$$

The resulting system (3.7) is hyperbolic, and the eigenvalues are explicit [279].

For modeling a phase transition, one can consider a relaxation source term of the form  $\pm \frac{1}{\varepsilon}(g_1 - g_2)$ ,  $g_i$  denoting the chemical potential of phase  $i$ , that can be added in the right-hand side of each equation of conservation of mass (the sum of the term is zero so as to keep the conservation of total mass). Then the limit system as  $\varepsilon \rightarrow 0$  is the (barotropic) Euler system (conservation of total mass and momentum) with a pressure law which takes into account the transition between the two phases; we refer to [238] (see also [602] for the system with energy).

It is just an insight into the extreme variety of models for two-phase flow, which, according to the modeling assumptions, differ by many aspects, in particular the number of equations, the presence of nonconservative products [1010], not mentioning the source terms, and also for which the hyperbolic feature is not an obvious issue (see [346, 364, 693], and also [511] for multiphase flow modeling via Hamilton's principle). Though we did not go into details on the modeling assumptions, because we focus on giving different examples of systems for which a general numerical approach is possible, it should be emphasized that a given model is valid only in some flow configuration and may fail to reproduce other configurations (see [965, 1073], for a recent survey on multi-fluid models [195]).  $\square$

For the sake of simplicity, we now essentially concentrate in the following on systems of the form (1.1) and occasionally on systems of the form (2.23) since they cover a large range of physical situations. The treatment of source terms will be given some consideration later on, in a special chapter.

## 4 Weak Solutions of Systems of Conservation Laws

Let us go back to the Cauchy problem (1.1), (1.2) for a general system of conservation laws. We shall say that a function  $\mathbf{u} : \mathbb{R}^d \times [0, \infty[ \rightarrow \Omega$  is a *classical solution* of (1.1), (1.2) if  $\mathbf{u}$  is a  $C^1$  function that satisfies Eqs. (1.1), (1.2) pointwise.

An essential feature of this problem is that there do not exist in general classical solutions of (1.1), (1.2) beyond some finite time interval, even when the initial condition  $\mathbf{u}_0$  is a very smooth function. Let us illustrate this fact by studying the simple case of a one-dimensional scalar equation.

### 4.1 Characteristics in the Scalar One-Dimensional Case

Thus, assume  $p = d = 1$  and let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a  $C^1$  function. We consider the problem

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \quad x \in \mathbb{R}, t > 0, \quad (4.1)$$

$$u(x, 0) = u_0(x), \quad x \in \mathbb{R}, \quad (4.2)$$

and set

$$a(u) = f'(u). \quad (4.3)$$

Let  $u$  be a classical solution of (4.1) so that we can write (4.1) in *nonconservative* form

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

and let us introduce the *characteristic* curves associated with (4.1). They are defined as the integral curves of the differential equation

$$\frac{dx}{dt} = a(u(x(t), t)). \quad (4.4)$$

*Proposition 4.1.*

Assume that  $u$  is a smooth solution of (4.1). The characteristic curves (4.4) are straight lines along which  $u$  is constant.

*Proof.* Consider a characteristic curve passing through the point  $(x_0, 0)$ , i.e., a solution of the ordinary differential system

$$\begin{cases} \frac{dx}{dt} = a(u(x(t), t)) \\ x(0) = x_0. \end{cases}$$

It exists at least on a small time interval  $[0, t_0[$ . Along such a curve,  $u$  is constant since

$$\begin{aligned} \frac{d}{dt}u(x(t), t) &= \frac{\partial u}{\partial t}(x(t), t) + \frac{\partial u}{\partial x}(x(t), t)\frac{dx}{dt}(t) \\ &= \left(\frac{\partial u}{\partial t} + a(u)\frac{\partial u}{\partial x}\right)(x(t), t) = 0. \end{aligned}$$

Therefore, it follows from (4.4) that the characteristic curves are straight lines whose constant slopes depend on the initial data, and the characteristic straight line passing through the point  $(x_0, 0)$  is defined by the equation

$$x = x_0 + t a(u_0(x_0)). \quad (4.5)$$

This important property gives a way to construct smooth solutions.

One sets

$$u(x, t) = u_0(x_0), \quad (4.6)$$

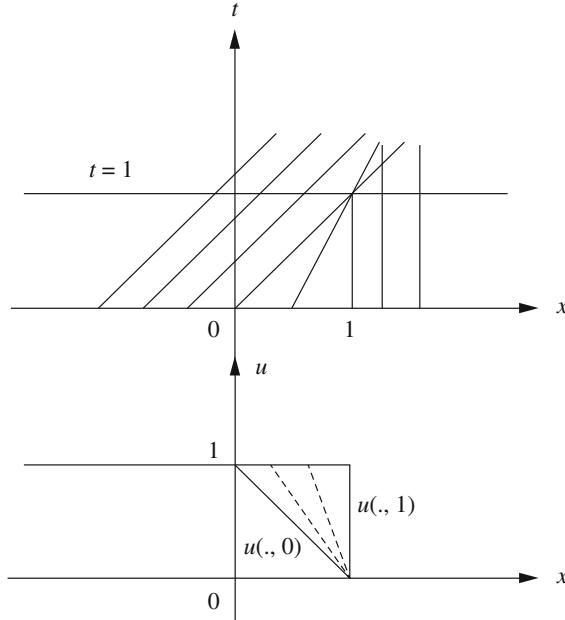
where  $x_0$  is solution of (4.5). This is the so-called *method of characteristics* (see example 4.1).  $\square$

Let us next assume that there exist two points  $x_1 < x_2$  such that

$$m_1 = \frac{1}{a(u_0(x_1))} < m_2 = \frac{1}{a(u_0(x_2))}.$$

Then, the characteristics  $C_1$  and  $C_2$  drawn from the points  $(x_1, 0)$  and  $(x_2, 0)$ , respectively, have slopes  $m_1$  and  $m_2$  and intersect necessarily at some point  $P$ .

At this point  $P$ , the solution  $u$  should take both values  $u_0(x_1)$  and  $u_0(x_2)$ , which is clearly impossible. Hence, the solution  $u$  cannot be continuous at the point  $P$ . Note that this phenomenon is independent of the smoothness of the



**Fig. 4.1** Method of characteristics for Burgers' equation

functions  $u_0$  and  $f$ . Indeed, using (4.5), we see that the two characteristics intersect at time  $t$  if

$$t(a(u_0(x_1)) - a(u_0(x_2))) = x_2 - x_1.$$

Thus, unless the function  $x \rightarrow a(u_0(x))$  is monotone increasing, in which case this equation has no positive solution  $t$ , we cannot define a classical solution  $u$  for all time  $t > 0$  (see Fig. 4.1). Moreover, one can determine the critical time  $T^*$  up to which a classical solution exists and can be constructed by the method of characteristics;  $T^*$  is given by

$$T^* = -\frac{1}{\min(\alpha, 0)}, \quad \alpha = \min_{y \in \mathbb{R}} \frac{d}{dy} a(u_0(y)).$$

The multidimensional case will be considered in Chap. V, Sect. 1.2.

*Example 4.1.* We want to solve the following Cauchy problem for Burgers' equation (1.4b) with the initial condition

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

By using the method of characteristics, we can solve up to the time when the characteristics intersect. We already know by (4.5) that the characteristic passing through the point  $(x_0, 0)$  is given by

$$x = x(x_0, t) = x_0 + tu_0(x_0)$$

so that

$$x(x_0, t) = \begin{cases} x_0 + t, & x_0 \leq 0, \\ x_0 + t(1 - x_0), & 0 \leq x_0 \leq 1, \\ x_0, & x_0 \geq 1. \end{cases}$$

For  $t < 1$ , the characteristics do not intersect (see Fig. 4.1). Hence, given a point  $(x, t)$  with  $t < 1$ , we draw the (backward) characteristic passing through this point, and we determine the corresponding point  $x_0$

$$x_0 = \begin{cases} x - t, & \text{if } x \leq t < 1, \\ \frac{x - t}{1 - t}, & \text{if } t \leq x \leq 1, \\ x, & \text{if } x \geq 1. \end{cases}$$

It consists of a front moving to the right and steepening until it becomes a “shock” (see Fig. 4.1). This discontinuity corresponds to the fact that at time  $t = 1$  the characteristics intersect.  $\square$

In short, using the method of characteristics (involving the implicit function theorem), one can prove that for  $u$  smooth enough a classical solution of (4.1), (4.2) exists in a small time interval. On the other hand, we have seen that in the nonlinear case  $a'(u) \neq 0$ , discontinuities may develop after a finite time. The above considerations lead us to introduce *weak solutions* (which are indeed weaker than the classical solutions!).

## 4.2 Weak Solutions: The Rankine-Hugoniot Condition

Consider the Cauchy problem (1.1), (1.2), and assume  $\mathbf{u}_0 \in \mathbf{L}_{\text{loc}}^\infty(\mathbb{R}^d)^p$ , where  $\mathbf{L}_{\text{loc}}^\infty$  is the space of locally bounded measurable functions; we want to state precisely in which sense (1.1), (1.2) is to be taken. Let  $\mathbf{C}_0^1(\mathbb{R}^d \times [0, +\infty[)$  denote the space of  $C^1$  functions  $\varphi$  with compact support in  $\mathbb{R}^d \times [0, +\infty[$  (which means that  $\varphi$  is the restriction to  $\mathbb{R}^d \times [0, +\infty[$  of a  $C^1$  function with compact support in an open set containing  $\mathbb{R}^d \times [0, +\infty[$ ). We begin by noticing that if  $\mathbf{u}$  is  $C^1$  and  $\varphi \in \mathbf{C}_0^1(\mathbb{R}^d \times [0, +\infty[)^p$ , we obtain by Green’s theorem (or integration by parts)

$$\begin{aligned}
& - \int_0^\infty \int_{\mathbb{R}^d} \left\{ \frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) \right\} \cdot \boldsymbol{\varphi} \, d\mathbf{x} dt \\
& = \int_0^\infty \int_{\mathbb{R}^d} \left\{ \mathbf{u} \cdot \frac{\partial \boldsymbol{\varphi}}{\partial t} + \sum_{j=1}^d \mathbf{f}_j(\mathbf{u}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_j} \right\} \, d\mathbf{x} \, dt + \int_{\mathbb{R}^d} \mathbf{u}(\mathbf{x}, 0) \cdot \boldsymbol{\varphi}(\mathbf{x}, 0) \, d\mathbf{x},
\end{aligned}$$

where the dot  $\cdot$  denotes the Euclidean inner product on  $\mathbb{R}^p$ . Thus, any *classical* solution  $\mathbf{u}$  of (1.1), (1.2) satisfies  $\forall \boldsymbol{\varphi} \in \mathbf{C}_0^1(\mathbb{R}^d \times [0, +\infty[)^p$

$$\int_0^\infty \int_{\mathbb{R}^d} \left\{ \mathbf{u} \cdot \frac{\partial \boldsymbol{\varphi}}{\partial t} + \sum_{j=1}^d \mathbf{f}_j(\mathbf{u}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_j} \right\} \, d\mathbf{x} \, dt + \int_{\mathbb{R}^d} \mathbf{u}_0(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x}, 0) \, d\mathbf{x} = 0. \quad (4.7)$$

Next, we remark that (4.7) makes sense if  $\mathbf{u} \in \mathbf{L}_{\text{loc}}^\infty(\mathbb{R}^d \times [0, +\infty[)^p$ . Hence, we introduce the following definition.

*Definition 4.1*

Assume that  $\mathbf{u}_0 \in \mathbf{L}_{\text{loc}}^\infty(\mathbb{R}^d)^p$ . A function  $\mathbf{u} \in \mathbf{L}_{\text{loc}}^\infty(\mathbb{R}^d \times [0, +\infty[)^p$  is called a weak solution of the Cauchy problem (1.1), (1.2) if  $\mathbf{u}(\mathbf{x}, t) \in \Omega$  a.e. and satisfies (4.7) for any function  $\boldsymbol{\varphi} \in \mathbf{C}_0^1(\mathbb{R}^d \times [0, +\infty[)^p$ .

By construction, a classical solution of problem (1.1), (1.2) is also a weak solution. Conversely, by choosing  $\boldsymbol{\varphi}$  in  $\mathbf{C}_0^\infty(\mathbb{R}^d \times ]0, \infty[^p)$ , where  $\mathbf{C}_0^\infty(\mathbb{R}^d \times ]0, \infty[)$  is the space of  $C^\infty$  functions with compact support in  $\mathbb{R}^d \times ]0, \infty[$ , we obtain that any weak solution  $\mathbf{u}$  satisfies (1.1) in the sense of distributions on  $\mathbb{R}^d \times ]0, \infty[$ . Moreover, if  $\mathbf{u}$  happens to be a  $C^1$  function, then it is a classical solution. Indeed, let  $\boldsymbol{\varphi} \in \mathbf{C}_0^1(\mathbb{R}^d \times ]0, +\infty[^p)$ ; integrating (4.7) by parts gives

$$\int_0^\infty \int_{\mathbb{R}^d} \left\{ \frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) \right\} \cdot \boldsymbol{\varphi} \, d\mathbf{x} \, dt = 0,$$

so that (1.1) holds pointwise.

Next, if we multiply (1.1) by a test function  $\boldsymbol{\varphi} \in \mathbf{C}_0^1(\mathbb{R}^d \times [0, +\infty[)^p$ , integrate by parts, and compare with (4.7), we obtain

$$\int_{\mathbb{R}^d} (\mathbf{u}(\mathbf{x}, 0) - \mathbf{u}_0(\mathbf{x})) \cdot \boldsymbol{\varphi}(\mathbf{x}, 0) \, d\mathbf{x} = 0,$$

which yields (1.2) pointwise.

We shall now consider solutions of (1.1) in the sense of distributions that are only piecewise smooth and therefore admit discontinuities.

First, we notice that the above argument shows that any distributional solution  $\mathbf{u}$  is a classical solution of (1.1) in any domain where  $\mathbf{u}$  is  $C^1$ . We restrict the study to a particular type of discontinuous function; in fact, this is not restrictive for the examples that we shall encounter (if one regards the structure of BV (bounded variation) functions, this simplification is not too unreasonable (see, for instance, DiPerna [423])). For the sake of brevity, we

say that a function  $\mathbf{u}$  is “piecewise  $C^1$ ” if there exist a finite number of smooth orientable surfaces  $\Sigma$  in the  $(t, \mathbf{x})$ -space outside of which  $\mathbf{u}$  is a  $C^1$  function and across which  $\mathbf{u}$  has a jump discontinuity. Given a surface of discontinuity  $\Sigma$ , we denote by  $\mathbf{n} = (n_t, n_{x_1}, n_{x_2}, \dots, n_{x_d})^T (\neq \mathbf{0})$  a normal vector to  $\Sigma$  and by  $\mathbf{u}_+$  and  $\mathbf{u}_-$  the limits of  $\mathbf{u}$  on each side of  $\Sigma$ ,

$$\mathbf{u}_\pm(\mathbf{x}, t) = \lim_{\varepsilon \rightarrow 0, \varepsilon > 0} \mathbf{u}((\mathbf{x}, t) \pm \varepsilon \mathbf{n}).$$

In the one-dimensional case, we assume in addition that any line of discontinuity  $\Sigma$  has a parametrization of the form  $(t, \xi(t))$ , where  $\xi : t \mapsto \xi(t)$  is a  $C^1$  function from some time interval  $(t_1, t_2)$  into  $\mathbb{R}$ , and we set in the same way

$$\mathbf{u}_\pm(\xi(t), t) = \lim_{\varepsilon \rightarrow 0, \varepsilon > 0} \mathbf{u}(\xi(t) \pm \varepsilon, t).$$

We now show that even in the frame of “piecewise  $C^1$ ” functions, not every discontinuity is admissible. The values of  $\mathbf{u}$  and  $\mathbf{f}(\mathbf{u})$  on each side of  $\Sigma$  are linked by a relation that is hidden in the equations, as results from the following theorem.

*Theorem 4.1*

*Let  $\mathbf{u} : \mathbb{R}^d \times [0, +\infty[ \rightarrow \Omega$  be a piecewise  $C^1$  function (in the above sense). Then,  $\mathbf{u}$  is a solution of (1.1) in the sense of distributions on  $\mathbb{R}^d \times ]0, +\infty[$  if and only if the two following conditions are satisfied:*

- (i)  $\mathbf{u}$  is a classical solution of (1.1) in the domains where  $\mathbf{u}$  is  $C^1$ ;
- (ii)  $\mathbf{u}$  satisfies the jump condition

$$(\mathbf{u}_+ - \mathbf{u}_-) n_t + \sum_{j=1}^d (\mathbf{f}_j(\mathbf{u}_+) - \mathbf{f}_j(\mathbf{u}_-)) n_{x_j} = \mathbf{0} \quad (4.8)$$

along the surfaces of discontinuity.

The jump relation (4.8) is known as the *Rankine-Hugoniot condition*.

*Proof.* Suppose that  $\mathbf{u}$  is a piecewise  $C^1$  function, solution of (1.1) in the sense of distributions on  $\mathbb{R}^d \times ]0, +\infty[$ . We have already observed that  $\mathbf{u}$  satisfies property (i).

Now, let  $\Sigma$  be a surface of discontinuity of  $\mathbf{u}$ ,  $M$  a point of  $\Sigma$ , and  $D$  a small ball centered at  $M$  (for simplicity, we assume that  $\Sigma \cap D$  is the only surface of discontinuity of  $\mathbf{u}$  in  $D$ ). We denote by  $D_\pm$  the two open components of  $D$  on each side of  $\Sigma$ . Let  $\varphi \in \mathbf{C}_0^\infty(D)^p$ . We write

$$0 = \int_D \left\{ \mathbf{u} \cdot \frac{\partial \varphi}{\partial t} + \sum_{j=1}^d \mathbf{f}_j(\mathbf{u}) \cdot \frac{\partial \varphi}{\partial x_j} \right\} d\mathbf{x} dt = \int_{D+} + \int_{D-} .$$

Suppose, for instance, that the normal vector  $\mathbf{n}$  to the surface  $\Sigma$  points in the direction of  $D_+$ . Then, applying Green's formula in  $D_+$  and  $D_-$  gives

$$\begin{aligned} 0 = & - \int_{D_+} \left\{ \frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) \right\} \cdot \boldsymbol{\varphi} \, d\mathbf{x} \, dt \\ & - \int_{\Sigma \cap D} \{ n_t \mathbf{u}_+ + \sum_{j=1}^d n_{x_j} \mathbf{f}_j(\mathbf{u}_+) \} \cdot \boldsymbol{\varphi} \, dS \\ & - \int_{D_-} \left\{ \frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) \right\} \cdot \boldsymbol{\varphi} \, d\mathbf{x} \, dt \\ & + \int_{\Sigma \cap D} \{ n_t \mathbf{u}_- + \sum_{j=1}^d n_{x_j} \mathbf{f}_j(\mathbf{u}_-) \} \cdot \boldsymbol{\varphi} \, dS. \end{aligned}$$

Since  $\mathbf{u}$  is a classical solution of (1.1) in  $D_+$  and  $D_-$ , the first and third integrals vanish; we obtain

$$\int_{\Sigma \cap D} \{ -n_t(\mathbf{u}_+ - \mathbf{u}_-) - \sum_{j=1}^d n_{x_j} (\mathbf{f}_j(\mathbf{u}_+) - \mathbf{f}_j(\mathbf{u}_-)) \} \cdot \boldsymbol{\varphi} \, dS = 0.$$

This holds for arbitrary  $\boldsymbol{\varphi}$ , and hence we obtain the jump relation (4.8) at the point  $M$ .

On the other hand, if  $\mathbf{u}$  is a piecewise  $C^1$  function that satisfies properties (i) and (ii), it is a simple matter to check that  $\mathbf{u}$  is indeed a distributional solution of (1.1).  $\square$

Denote by

$$[\mathbf{u}] = \mathbf{u}_+ - \mathbf{u}_- \tag{4.9}$$

the *jump* of  $\mathbf{u}$  across  $\Sigma$  and similarly by

$$[\mathbf{f}_j(\mathbf{u})] = \mathbf{f}_j(\mathbf{u}_+) - \mathbf{f}_j(\mathbf{u}_-)$$

the jump of  $\mathbf{f}_j(\mathbf{u})$ ,  $1 \leq j \leq d$ ; then (4.8) can be written

$$n_t [\mathbf{u}] + \sum_{j=1}^d n_{x_j} [\mathbf{f}_j(\mathbf{u})] = 0.$$

If  $(n_{x_1}, \dots, n_{x_d}) \neq (0, \dots, 0)$ , we can take the normal vector in the form

$$\mathbf{n} = \begin{pmatrix} -s \\ \boldsymbol{\nu} \end{pmatrix},$$

where  $s \in \mathbb{R}$  and  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d)^T$  is a unit vector in  $\mathbb{R}^d$ . Then (4.8) can be equivalently written

$$s[\mathbf{u}] = \sum_{j=1}^d \nu_j [\mathbf{f}_j(\mathbf{u})].$$

We recall here that  $\Sigma$  has a standard orientation determined by the choice of the canonical volume form on its tangent space and thus a canonical normal vector field associated to this orientation. Now, if  $\Sigma$  is oriented and  $\frac{\mathbf{n}}{|\mathbf{n}|}$  is the outward unit normal vector to  $\Sigma$ ,  $\boldsymbol{\nu}$  and  $s$  may be interpreted respectively as the direction and the speed of propagation of the discontinuity. For instance, suppose that in the two-dimensional case ( $d = 2$ )  $\Sigma$  is a smooth surface in  $\mathbb{R}^3$  that has a parametrization of the form  $(t, x_1, x_2 = \xi(t, x_1))$ . We have

$$\mathbf{n} = (n_t, n_{x_1}, n_{x_2})^T = \left(1 + \left(\frac{\partial \xi}{\partial x_1}\right)^2\right)^{-1/2} \left(-\frac{\partial \xi}{\partial t}, -\frac{\partial \xi}{\partial x_1}, 1\right)^T,$$

and  $\boldsymbol{\nu}$  points in the direction of the positive  $x_2$ -axis.

In the one-dimensional case ( $d = 1$ ), we have assumed that  $\Sigma$  is a smooth curve with parametrization  $(t, \xi(t))$ , and we have

$$\mathbf{n} = (-s, 1)^T, \quad s = \frac{d\xi}{dt}, \quad (4.10)$$

so that the Rankine-Hugoniot jump condition (4.8) becomes

$$[\mathbf{f}(\mathbf{u})] = s[\mathbf{u}]. \quad (4.11)$$

For a scalar equation, we obtain

$$s = \frac{[f(u)]}{[u]},$$

whereas for a system, (4.11) represents in fact  $p$  equations

$$[f_i(\mathbf{u})] = s[u_i], \quad 1 \leq i \leq p$$

in which the same  $s$  appears.

*Remark 4.1.* Note that, as for the characteristic curves,  $\Sigma$  is traditionally represented in the  $(x, t)$ -space, and this is the main reason for which we have chosen to write  $u(x, t)$  instead of  $u(t, x)$ . Thus, in the figures, Fig. 5.1 for instance,  $s$  is the reciprocal of the slope. The wave propagation is such that  $s$  is always finite. This remark is valid for systems too.  $\square$

*Remark 4.2.* If the function  $\mathbf{u}$  is continuous, the Rankine-Hugoniot jump condition is automatically satisfied. In that case, it suffices to check that  $\mathbf{u}$  satisfies (1.1) in the domains where  $\mathbf{u}$  is  $C^1$  to prove that  $\mathbf{u}$  is indeed a distributional solution of (1.1).  $\square$

Next, we want to point out, again with a simple example, that a weak solution of the Cauchy problem (1.1), (1.2) is not necessarily unique.

### 4.3 Example of Nonuniqueness of Weak Solutions

We consider the Riemann problem (1.3) for Burgers' equation (1.4b), i.e.,

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \frac{u^2}{2} = 0 \\ u_0(x) = \begin{cases} u_\ell, & x < 0, \\ u_r, & x > 0. \end{cases} \end{cases} \quad (4.12)$$

If  $u_\ell \neq u_r$ , the Rankine-Hugoniot condition (4.11) shows that we obtain a weak solution of the problem by propagating the discontinuity at speed  $s = \frac{1}{2}(u_\ell + u_r)$ , which gives

$$u(x, t) = \begin{cases} u_\ell, & x < st \\ u_r, & x > st. \end{cases} \quad (4.13)$$

Now, let us check that there are many other weak solutions. In fact, let  $a$  be a constant such that  $a \geq \max(u_\ell, -u_r)$ . The function  $u$  defined by

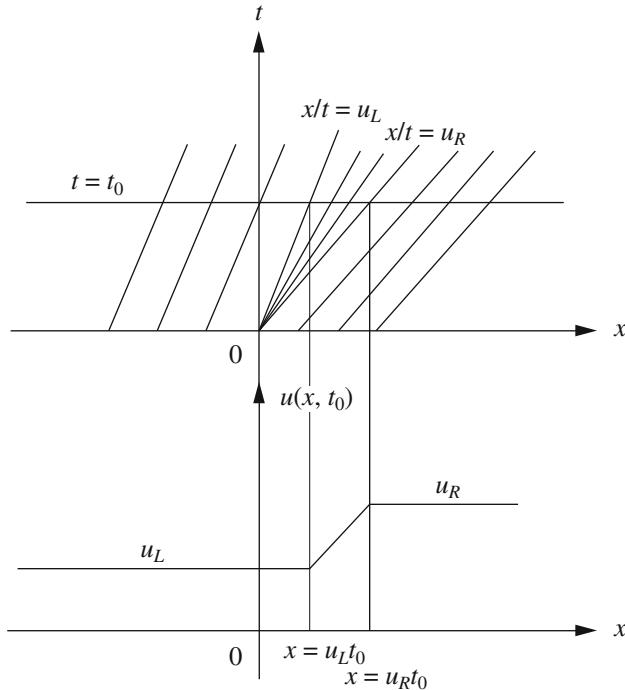
$$u(x, t) = \begin{cases} u_\ell, & x < s_1 t \\ -a, & s_1 t < x < 0 \\ a, & 0 < x < s_2 t \\ u_r, & x > s_2 t \end{cases}$$

is also a weak solution if  $s_1 = \frac{u_\ell - a}{2}$ ,  $s_2 = \frac{u_r + a}{2}$ , so that the Rankine-Hugoniot jump condition (4.11) is satisfied along each line of discontinuity of  $u$ . We thus obtain a one-parameter family of discontinuous weak solutions.

On the other hand, if we suppose  $u_\ell \leq u_r$ , we can also exhibit a continuous solution. In that case, the characteristics do not intersect (see Fig. 4.2). Clearly, the method of characteristics enables us to determine the solution everywhere except in the region  $u_\ell \leq \frac{x}{t} \leq u_r$ . However, we notice that the function  $v(x, t) = \frac{x}{t}$  is a classical solution of Burgers' equation since

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial t} = -\frac{x}{t^2} + \left(\frac{x}{t}\right)\left(\frac{1}{t}\right) = 0.$$

Hence, the continuous function



**Fig. 4.2** Example of rarefaction for Burgers' equation

$$u(x, t) = \begin{cases} u_\ell, & x \leq u_\ell t \\ x/t, & u_\ell t \leq x \leq u_r t \\ u_r, & x \geq u_r t \end{cases} \quad (4.14)$$

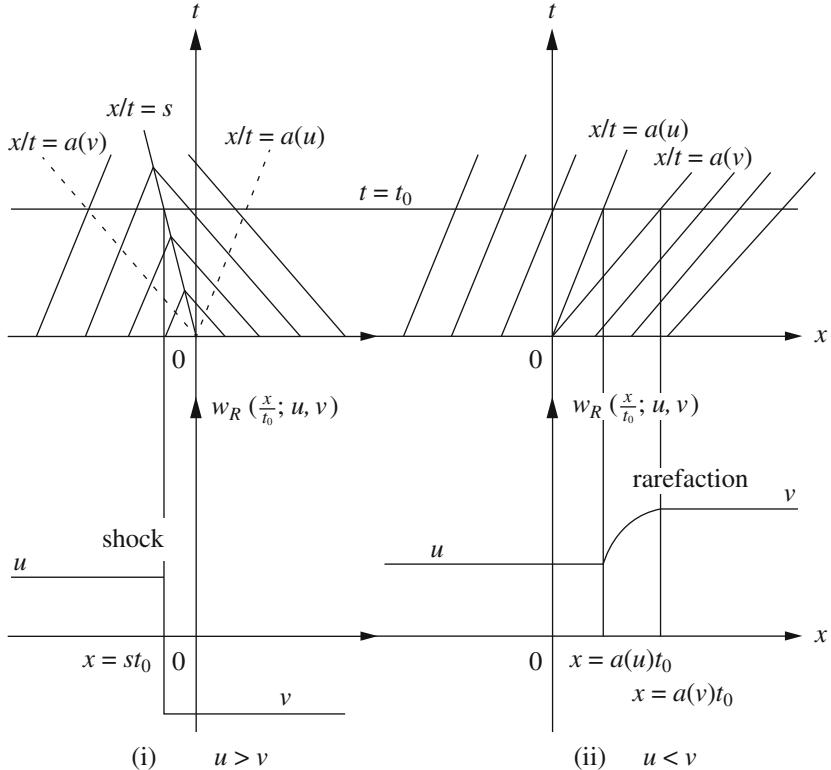
is also a weak solution of problem (4.12) (see Fig. 4.2).

Note that the function  $\frac{x}{t}$  was not found by chance. For more general strictly convex fluxes  $f$ , assuming  $u_\ell \leq u_r$ , we can still find a self-similar continuous solution of the Riemann problem (see an illustration in Fig. 5.1(ii))

$$u(x, t) = v\left(\frac{x}{t}\right), \quad f'(u_\ell) \leq \frac{x}{t} \leq f'(u_r), \text{ where } f'(v(\xi)) = \xi.$$

Let us also note the obvious property that in the “rarefaction fan”  $f'(u_\ell) \leq \frac{x}{t} \leq f'(u_r)$ , where we cannot use the method of characteristics from the data at time 0, this solution (4.14) is smooth and constant on the lines  $\frac{x}{t} = Cte$ , which are characteristics for  $t > 0$ .

We have just noticed that a weak solution of the Cauchy problem (1.1), (1.2) is not necessarily unique. Hence, we need to find some criterion that enables us to choose the “physically relevant” solution among all the weak solutions of (1.1), (1.2). This criterion is based on the concept of *entropy* that we introduce now. In the above example, when  $u_\ell \leq u_r$ , it happens that the “relevant”



**Fig. 5.1** Solution of the scalar Riemann problem (convex case)

solution is the continuous one. One heuristic reason is that for the other discontinuous solutions, we have added some arbitrary extra information  $a$  that is contained neither in the data  $u_0$  nor in the equation itself.

## 5 Entropy Solution

### 5.1 A Mathematical Notion of Entropy

Let us consider the following problem: given any smooth solution  $\mathbf{u}$  of (1.1), we wonder whether  $\mathbf{u}$  would satisfy an additional conservation law of the form

$$\frac{\partial}{\partial t} U(\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}) = 0, \quad (5.1)$$

where  $U$  and  $F_j, 1 \leq j \leq d$ , are sufficiently smooth functions from  $\Omega$  into  $\mathbb{R}$ . We want to check that this is indeed the case if

$$U'(\mathbf{u})\mathbf{f}'_j(\mathbf{u}) = F'_j(\mathbf{u}), \quad 1 \leq j \leq d, \quad (5.2)$$

where, for ease of notation, we identify the linear forms  $U'(\mathbf{u}), F'_j(\mathbf{u}) : \mathbb{R}^p \rightarrow \mathbb{R}$  with the corresponding row vectors

$$U' = \nabla U^T = \left( \frac{\partial U}{\partial u_1}, \dots, \frac{\partial U}{\partial u_p} \right), \quad F'_j = \nabla F_j^T = \left( \frac{\partial F_j}{\partial u_1}, \dots, \frac{\partial F_j}{\partial u_p} \right)$$

and the linear mapping  $\mathbf{f}'_j : \mathbb{R}^p \rightarrow \mathbb{R}^p$  with the matrix

$$\mathbf{f}'_j = \mathbf{A}_j = \left( \frac{\partial f_{ij}}{\partial u_k} \right) \quad 1 \leq i, k \leq p.$$

In fact, assuming that  $\mathbf{u}$  is a classical solution of (1.1) and carrying out the differentiation, we obtain

$$U'(\mathbf{u}) \left( \frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \mathbf{f}'_j(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x_j} \right) = 0$$

and by (5.2)

$$U'(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d F'_j(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x_j} = 0, \quad (5.3)$$

so that (5.1) follows.

Let us now define the notion of *entropy* in the sense of Lax. We shall restrict ourselves to convex entropy functions though one may consider a more general framework.

### *Definition 5.1*

Assume that  $\Omega$  is convex. Then, a convex function  $U : \Omega \rightarrow \mathbb{R}$  is called an entropy for the system of conservation laws (1.1) if there exist  $d$  functions  $F_j : \Omega \rightarrow \mathbb{R}, 1 \leq j \leq d$ , called entropy fluxes, such that the relations (5.2) hold.

Hence, if (5.2) holds, any classical solution of (1.1) satisfies the additional conservation law (5.1). On the other hand, this is not true in general of a weak solution and in particular of a piecewise  $C^1$  weak solution. Such a weak solution  $\mathbf{u}$  must satisfy the Rankine-Hugoniot condition (4.8) along the surfaces of discontinuity, whereas a solution of (5.1) should satisfy the corresponding jump condition

$$n_t[U(\mathbf{u})] + \sum_{j=1}^d n_{x_j}[F_j(\mathbf{u})] = 0, \quad (5.4)$$

which is in general incompatible with (4.8). We shall see below that, for an entropy solution, this last jump condition should be replaced by a jump inequality.

The problem is to find entropy functions and if possible all entropy functions associated with the nonlinear system of conservation laws (1.1). This is easy for a scalar conservation law ( $p = 1$ ) since in that case any convex function  $U$  is an entropy. In fact, it suffices to take for  $F_j$  a primitive of the function  $U' f'_j$ .

In the general case  $p > 1$ , finding the entropy functions is a much more difficult problem. Note that Eqs. (5.2) can be written in the form of a system of  $p \times d$  linear partial differential equations of the first order in the  $(d + 1)$  unknown functions  $U, F_j, 1 < j < d$ , namely,

$$\sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_k} \frac{\partial U}{\partial u_i} - \frac{\partial F_j}{\partial u_k} = 0, \quad 1 \leq j \leq d, 1 \leq k \leq p. \quad (5.5)$$

Except in the case  $p = 2, d = 1$  (which is treated, for instance, in Serre [1032]), the existence of entropy functions (also called the *extensions*) is a special property of the system. However, in all the practical examples derived from mechanics or physics, we are able to find an entropy function that has a physical meaning. A classification of hyperbolic systems ( $d = 1$ ) with respect to their entropies is given in Serre [1036]; see also [654].

*Example 5.1. The  $p$ -system (Example 1.3 revisited).* We consider again the  $p$ -system (1.7), where we assume  $p'(v) < 0$ . Let  $P$  be a primitive of  $p$ . By multiplying the first equation (1.7) by  $-p(v)$ , the second equation (1.7) by  $u$ , and adding, we obtain

$$\frac{\partial}{\partial t} \left( \frac{u^2}{2} - P(v) \right) + \frac{\partial}{\partial x} (p(v)u) = 0.$$

Therefore, the pair  $(U, F)$  defined by

$$U(v, u) = \frac{u^2}{2} - P(v), \quad F(v, u) = p(v)u$$

is a pair of an entropy function and an entropy flux for the  $p$ -system. Note that the Hessian matrix of  $U$ , given by

$$U''(v, u) = \begin{pmatrix} -p'(v) & 0 \\ 0 & 1 \end{pmatrix},$$

is positive-definite so that  $U$  is strictly convex.  $\square$

*Example 5.2. Symmetric systems.* Let us go back to the general situation (1.1) but assume that the  $p \times p$  matrices  $\mathbf{A}_j = \mathbf{f}'_j(\mathbf{u})$  are symmetric, i.e.,

$$\frac{\partial f_{ij}}{\partial u_k} = \frac{\partial f_{kj}}{\partial u_i}, \quad 1 \leq i, k \leq p.$$

The above relations are exactly the compatibility conditions for the existence of a function  $g_j = g_j(\mathbf{u})$  such that

$$\frac{\partial g_j}{\partial u_i} = f_{ij}.$$

Next, we observe that the following strictly convex function

$$U(\mathbf{u}) = \frac{1}{2} \sum_{i=1}^p u_i^2$$

is an entropy for the symmetric system (1.1) associated with the entropy fluxes

$$F_j(\mathbf{u}) = \sum_{i=1}^p u_i f_{ij}(\mathbf{u}) - g_j(\mathbf{u}).$$

Indeed, since

$$\frac{\partial}{\partial x_j} g_j(\mathbf{u}) = \sum_{i=1}^p \frac{\partial g_j}{\partial u_i}(\mathbf{u}) \frac{\partial u_i}{\partial x_j} = \sum_{i=1}^p f_{ij}(\mathbf{u}) \frac{\partial u_i}{\partial x_j},$$

we have

$$\begin{aligned} \frac{\partial}{\partial x_j} F_j(\mathbf{u}) &= \sum_{i=1}^p u_i \frac{\partial}{\partial x_j} (f_{ij}(\mathbf{u})) + \sum_{i=1}^p f_{ij}(\mathbf{u}) \frac{\partial u_i}{\partial x_j} - \sum_{i=1}^p f_{ij}(\mathbf{u}) \frac{\partial u_i}{\partial x_j} \\ &= \sum_{i=1}^p u_i \frac{\partial}{\partial x_j} (f_{ij}(\mathbf{u})) \end{aligned}$$

so that

$$\frac{\partial}{\partial t} U(\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}) = \sum_{i=1}^p u_i \left\{ \frac{\partial u_i}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} (f_{ij}(\mathbf{u})) \right\} = 0.$$

We shall now prove that the situation of Example 5.2 is in fact almost general.  $\square$

Let us see that a nonlinear system of conservation laws that admits a strictly convex entropy is *symmetrizable*. More precisely, we can state the following theorem (Godunov-Mock theorem).

### Theorem 5.1

Let  $U : \Omega \rightarrow \mathbb{R}$  be a strictly convex function. A necessary and sufficient condition for  $U$  to be an entropy for the system (1.1) is that the  $(p \times p)$  matrices  $U''(\mathbf{u})\mathbf{f}'_j(\mathbf{u}), 1 \leq j \leq d$ , are symmetric.

*Proof.* Since the function  $U$  is strictly convex, its Hessian  $U''(\mathbf{u})$  is a symmetric positive-definite matrix. Now, we assume that  $U$  is an entropy so

that the conditions (5.5) hold. Then, differentiating (5.5) with respect to  $u_\ell$  gives

$$\frac{\partial^2 F_j}{\partial u_k \partial u_\ell} - \sum_{i=1}^p \frac{\partial^2 f_{ij}}{\partial u_k \partial u_\ell} \frac{\partial U}{\partial u_i} = \sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_k} \frac{\partial^2 U}{\partial u_i \partial u_\ell}, \quad 1 \leq k, \ell \leq p.$$

Since the left-hand side member of the above equation is symmetric in  $k$  and  $\ell$ , the same is true of the right-hand side, i.e.,

$$\sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_k} \frac{\partial^2 U}{\partial u_i \partial u_\ell} = \sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_\ell} \frac{\partial^2 U}{\partial u_i \partial u_k}, \quad 1 \leq k, \ell \leq p. \quad (5.6)$$

This means exactly that the matrix  $U''(\mathbf{u})\mathbf{f}'_j(\mathbf{u})$  is symmetric,  $1 \leq j \leq d$ .

Conversely, assume that conditions (5.6) hold. Using

$$\frac{\partial}{\partial u_\ell} \left( \sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_k} \frac{\partial U}{\partial u_i} \right) = \sum_{i=1}^p \left( \frac{\partial f_{ij}}{\partial u_k} \frac{\partial^2 U}{\partial u_i \partial u_\ell} + \frac{\partial^2 f_{ij}}{\partial u_k \partial u_\ell} \frac{\partial U}{\partial u_i} \right),$$

we obtain

$$\frac{\partial}{\partial u_\ell} \left( \sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_k} \frac{\partial U}{\partial u_i} \right) = \frac{\partial}{\partial u_k} \left( \sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_\ell} \frac{\partial U}{\partial u_i} \right), \quad 1 \leq k, \ell \leq p.$$

These relations are the compatibility conditions that ensure the existence of a function  $F_j$  such that

$$\frac{\partial F_j}{\partial u_k} = \sum_{i=1}^p \frac{\partial f_{ij}}{\partial u_k} \frac{\partial U}{\partial u_i}.$$

Hence, it follows from (5.5) that  $U$  is an entropy function associated with the entropy fluxes  $F_j$ ,  $1 \leq j \leq d$ .  $\square$

As a corollary of Theorem 5.1, the existence of a strictly convex entropy  $U$  implies that the system (1.1) is symmetrizable. In fact, premultiplication by  $U''(\mathbf{u})$  gives the system

$$U''(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d U''(\mathbf{u}) \mathbf{f}'_j(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x_j} = 0,$$

where the matrix  $U''(\mathbf{u})$  is symmetric and positive-definite and the matrices  $U''(\mathbf{u})\mathbf{f}'_j(\mathbf{u})$  are symmetric.

The symmetrization of (1.1) can also be accomplished by introducing new dependent variables  $\mathbf{v}$ , i.e., by setting  $\mathbf{u} = \mathbf{u}(\mathbf{v})$ . Equation (1.1) then becomes

$$\mathbf{u}'(\mathbf{v}) \frac{\partial \mathbf{v}}{\partial t} + \sum_{j=1}^d \mathbf{f}'_j(\mathbf{u}) \mathbf{u}'(\mathbf{v}) \frac{\partial \mathbf{v}}{\partial x_j} = \mathbf{0}$$

so that (1.1) is symmetrized if again  $\mathbf{u}'(\mathbf{v})$  is a symmetric positive-definite matrix and the matrices  $\mathbf{f}'_j(\mathbf{u})$   $\mathbf{u}'(\mathbf{v})$  are symmetric.

*Theorem 5.2*

*A necessary and sufficient condition for the system (1.1) to possess a strictly convex entropy  $U$  is that there exists a change of dependent variables  $\mathbf{u} = \mathbf{u}(\mathbf{v})$  that symmetrizes (1.1).*

*Proof.* Suppose that (1.1) is symmetrized by the change of variables  $\mathbf{u} = \mathbf{u}(\mathbf{v})$ . The symmetry of the matrix

$$\mathbf{u}'(\mathbf{v}) = \left( \frac{\partial u_i}{\partial v_k} \right)_{1 \leq i, k \leq p}$$

implies the existence of a function  $q(\mathbf{v})$  such that

$$q'(\mathbf{v}) = \mathbf{u}(\mathbf{v})^T.$$

Similarly, the symmetry of the matrices  $\mathbf{f}'_j(\mathbf{u}(\mathbf{v}))$   $\mathbf{u}'(\mathbf{v})$ ,  $1 \leq j \leq d$ , implies the existence of functions  $p_j(\mathbf{v})$  such that

$$p'_j(\mathbf{v}) = \mathbf{f}'_j(\mathbf{u}(\mathbf{v}))^T, \quad i \leq j \leq d.$$

The positive-definiteness of  $\mathbf{u}'(\mathbf{v})$  is equivalent to the strict convexity of the function  $q(\mathbf{v})$ , which implies in turn that the mapping  $\mathbf{v} \mapsto q'(\mathbf{v})$  is one-to-one. Hence,  $\mathbf{v}$  can be regarded as a function of  $\mathbf{u}$ . We set

$$U(\mathbf{u}) = \mathbf{v}(\mathbf{u})^T \mathbf{u} - q(\mathbf{v}(\mathbf{u})), \quad F_j(\mathbf{u}) = \mathbf{v}(\mathbf{u})^T \mathbf{f}'_j(\mathbf{u}) - p'_j(\mathbf{v}(\mathbf{u})), \quad 1 \leq j \leq p.$$

By differentiating  $U$  and  $F_j$ , we obtain

$$U'(\mathbf{u}) = \mathbf{v}(\mathbf{u})^T + \mathbf{u}^T \mathbf{v}'(\mathbf{u}) - q'(\mathbf{v}(\mathbf{u})) \mathbf{v}'(\mathbf{u}) = \mathbf{v}(\mathbf{u})^T$$

and

$$F'_j(\mathbf{u}) = \mathbf{v}(\mathbf{u})^T \mathbf{f}'_j(\mathbf{u}) + \mathbf{f}'_j(\mathbf{u})^T \mathbf{v}'(\mathbf{u}) - p'_j(\mathbf{v}(\mathbf{u})) \mathbf{v}'(\mathbf{u}) = \mathbf{v}(\mathbf{u})^T \mathbf{f}'_j(\mathbf{u})$$

so that the relations (5.2) hold. Moreover, we have

$$U''(\mathbf{u}(\mathbf{v})) \cdot \mathbf{u}'(\mathbf{v}) = \mathbf{I},$$

and therefore the matrix

$$U''(\mathbf{u}(\mathbf{v})) = \mathbf{u}'(\mathbf{v})^{-1}$$

is positive-definite. This proves that  $U$  is a strictly convex entropy.

Conversely, if  $U$  is a strictly convex entropy, the mapping  $\mathbf{u} \rightarrow U'(\mathbf{u})$  is one-to-one. Hence, we can define the change of variables

$$\mathbf{v}^T = U'(\mathbf{u})$$

(the components of  $\mathbf{v}$  are called the *entropy variables*). Next, we define the conjugate, or polar, functions  $U^*$  and  $F_j^*$  of  $U$  and  $F_j$ ,  $1 \leq j \leq d$  by

$$U^*(\mathbf{v}) = \mathbf{v}^T \mathbf{u}(\mathbf{v}) - U(\mathbf{u}(\mathbf{v})), F_j^*(\mathbf{v}) = \mathbf{v}^T \mathbf{f}_j(\mathbf{u}(\mathbf{v})) - F_j(\mathbf{u}(\mathbf{v})).$$

Differentiating with respect to  $\mathbf{v}$  gives

$$U^{*' }(\mathbf{v}) = \mathbf{u}(\mathbf{v})^T + \mathbf{v}^T \mathbf{u}'(\mathbf{v}) - U'(\mathbf{u}(\mathbf{v}))\mathbf{u}'(\mathbf{v}) = \mathbf{u}(\mathbf{v})^T$$

and

$$F_j^{*' }(\mathbf{v}) = \mathbf{f}_j(\mathbf{u}(\mathbf{v}))^T + \mathbf{v}^T \mathbf{f}'_j(\mathbf{u}(\mathbf{v}))\mathbf{u}'(\mathbf{v}) - F'_j(\mathbf{u}(\mathbf{v}))\mathbf{u}'(\mathbf{v}) = \mathbf{f}_j(\mathbf{u}(\mathbf{v}))^T.$$

Hence, the matrices  $\mathbf{u}'(\mathbf{v}) = U^{**'}(\mathbf{v})$  and  $\mathbf{f}'_j(\mathbf{u}(\mathbf{v}))\mathbf{u}'(\mathbf{v}) = F_j^{**'}(\mathbf{v})$ ,  $1 \leq j \leq d$ , are symmetric. Moreover, the matrix

$$\mathbf{u}'(\mathbf{v}) = U''(\mathbf{u}(\mathbf{v}))^{-1}$$

is positive-definite, which proves that the change of variables  $\mathbf{v}^T = U'(\mathbf{u})$  symmetrizes (1.1).  $\square$

*Example 5.3. Gas dynamics equations.* Let us consider again the gas dynamics equations in Eulerian coordinates (Example 2.1). We define the (thermodynamic) specific entropy  $s$  by

$$T \, ds = d\varepsilon + p \, d\tau, \quad \tau = \frac{1}{\rho},$$

where  $T = T(\rho, \varepsilon)$  is the temperature. We shall check in Chap. III, Sect. 1.1, that  $U = -\rho s$  is indeed a strictly convex entropy function (of the conservative variables  $(\rho, q_1, q_2, q_3, E)$ ) associated with the entropy fluxes

$$F_i = -\rho u_i s, \quad 1 \leq i \leq 3.$$

In fact, for a polytropic ideal gas, we have

$$T = \frac{\varepsilon}{C_v},$$

where the constant  $C_v$  is the specific heat at constant volume (see Chap. III, Sect. 1.2). Since  $p = (\gamma - 1)\rho \varepsilon$ , we obtain in this case

$$ds = C_v \left( \frac{d\varepsilon}{\varepsilon} - \frac{p}{\varepsilon} \frac{d\rho}{\rho^2} \right) = C_v \left( \frac{d\varepsilon}{\varepsilon} - (\gamma - 1) \frac{d\rho}{\rho} \right),$$

so that

$$s = s_0 + C_v \operatorname{Log} \left( \frac{\varepsilon}{\rho^{\gamma-1}} \right),$$

where  $s_0$  is an arbitrary constant. This can be equivalently written

$$s = s_0 + C_v \left( \text{Log}(E - \frac{|\mathbf{q}|^2}{2\rho}) - \gamma \text{ Log } \rho \right),$$

where  $|\mathbf{q}|^2 = q_1^2 + q_2^2 + q_3^2$ , and it is an easy but lengthy matter to check directly that the function

$$(\rho, q_1, q_2, q_3, E) \rightarrow -\rho \left( \text{Log}(E - \frac{|\mathbf{q}|^2}{2\rho}) - \gamma \text{ Log } \rho \right)$$

is strictly convex. Hence, we get that smooth solutions of (2.1) satisfy the conservation law

$$\partial_t(\rho s) + \sum_j \partial_{x_j}(\rho u_j s) = 0.$$

Now, in Lagrangian coordinates, using (2.19), this entropy conservation law writes  $\partial_t s = 0$ . This is one of the characteristics of *fluid systems* (see [414]).  $\square$

## 5.2 The Vanishing Viscosity Method

The concept of entropy will enable us to select among the weak solutions of (1.1), (1.2) the physically relevant solution. To make this point clear, we begin by introducing a viscous perturbation of the nonlinear system of conservation laws. Given a small parameter  $\varepsilon > 0$ , we associate with (1.1) the parabolic system

$$\frac{\partial \mathbf{u}_\varepsilon}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}_\varepsilon) - \varepsilon \Delta \mathbf{u}_\varepsilon = \mathbf{0}, \quad (5.7)$$

where  $-\varepsilon \Delta \mathbf{u}_\varepsilon$  can be viewed as a viscosity term. Now, assuming that the systems (5.7) (with  $\mathbf{u}_\varepsilon(\mathbf{x}, 0) = \mathbf{u}_{0\varepsilon}(\mathbf{x}) \rightarrow \mathbf{u}_0(\mathbf{x})$  as  $\varepsilon \rightarrow 0$ ) have sufficiently smooth solutions  $\mathbf{u}_\varepsilon$  (see Goodman and Xin [551] and Gisclon [517] for the existence of  $\mathbf{u}_\varepsilon$ ), we want to recover solutions of (1.1) as the limits as  $\varepsilon \rightarrow 0$  of solutions of (5.7). In that direction, we can state the following theorem.

*Theorem 5.3*

Assume that (1.1) admits an entropy  $U$  with entropy fluxes  $F_j, 1 \leq j \leq d$ . Let  $(\mathbf{u}_\varepsilon)_\varepsilon$  be a sequence of sufficiently smooth solutions of (5.7) such that

$$\|\mathbf{u}_\varepsilon\|_{\mathbf{L}^\infty(\mathbb{R}^d \times [0, +\infty[)^p} \leq C, \quad (5.8)$$

$$\mathbf{u}_\varepsilon \rightarrow \mathbf{u} \quad \text{as } \varepsilon \rightarrow 0 \quad \text{a.e. in } \mathbb{R}^d \times [0, +\infty[, \quad (5.9)$$

where  $C > 0$  is a constant independent of  $\varepsilon$ . Then  $\mathbf{u}$  is a weak solution of (1.1) and satisfies the entropy condition

$$\frac{\partial}{\partial t} U(\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}) \leq 0 \quad (5.10)$$

in the sense of distributions on  $\mathbb{R}^d \times ]0, +\infty[$ .

The inequality (5.10) means that for any function  $\varphi \in \mathbf{C}_0^\infty(\mathbb{R}^d \times ]0, +\infty[)$ ,  $\varphi \geq 0$ , we have

$$\int_0^\infty \int_{\mathbb{R}^d} \left( U(\mathbf{u}) \frac{\partial \varphi}{\partial t} + \sum_{j=1}^d F_j(\mathbf{u}) \frac{\partial \varphi}{\partial x_j} \right) d\mathbf{x} dt \geq 0. \quad (5.11)$$

*Proof.* Let  $U$  be a  $C^2$  entropy function; by applying  $U'(\mathbf{u}_\varepsilon)$  to (5.7) and taking into account the relations (5.2), we obtain

$$\frac{\partial}{\partial t} U(\mathbf{u}_\varepsilon) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}_\varepsilon) = \varepsilon U'(\mathbf{u}_\varepsilon) \Delta \mathbf{u}_\varepsilon.$$

We rewrite the right-hand side of the above equation in the following way:

$$\varepsilon U'(\mathbf{u}_\varepsilon) \Delta \mathbf{u}_\varepsilon = \varepsilon \Delta U(\mathbf{u}_\varepsilon) - \varepsilon \sum_{j=1}^d \left( \frac{\partial \mathbf{u}_\varepsilon}{\partial x_j} \right)^T U''(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_j},$$

so that by the convexity of  $U$

$$\varepsilon U'(\mathbf{u}_\varepsilon) \Delta \mathbf{u}_\varepsilon \leq \varepsilon \Delta U(\mathbf{u}_\varepsilon)$$

and therefore

$$\frac{\partial}{\partial t} U(\mathbf{u}_\varepsilon) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}_\varepsilon) \leq \varepsilon \Delta U(\mathbf{u}_\varepsilon). \quad (5.12)$$

Now, suppose that  $(\mathbf{u}_\varepsilon)_\varepsilon$  is a sequence of smooth solutions of (5.7) that satisfies the properties (5.8) and (5.9). Then

$$\mathbf{u}_\varepsilon \rightarrow \mathbf{u} \text{ as } \varepsilon \rightarrow 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d \times ]0, +\infty[)^p$$

i.e., in the sense of distributions on  $\mathbb{R}^d \times ]0, +\infty[$  so that

$$\frac{\partial \mathbf{u}_\varepsilon}{\partial t} \rightarrow \frac{\partial \mathbf{u}}{\partial t}, \quad \varepsilon \Delta \mathbf{u}_\varepsilon \rightarrow 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^d \times ]0, +\infty[)^p.$$

Next, using (5.8) and the Lebesgue dominated convergence theorem, we have

$$\mathbf{f}_j(\mathbf{u}_\varepsilon) \rightarrow \mathbf{f}_j(\mathbf{u}) \quad \text{in } \mathbf{L}_{\text{loc}}^1(\mathbb{R}^d \times ]0, +\infty[)^p.$$

Hence, by passing to the limit in (5.7), we obtain that  $\mathbf{u}$  is a solution of (1.1) in the sense of distributions on  $\mathbb{R}^d \times ]0, +\infty[$ .

Similarly, we have

$$\frac{\partial}{\partial t} U(\mathbf{u}_\varepsilon) \rightarrow \frac{\partial}{\partial t} U(\mathbf{u}), \quad \frac{\partial}{\partial x_j} F_j(\mathbf{u}_\varepsilon) \rightarrow \frac{\partial}{\partial x_j} F_j(\mathbf{u}), \quad \varepsilon \Delta U(\mathbf{u}_\varepsilon) \rightarrow 0$$

in  $\mathcal{D}'(\mathbb{R}^d \times ]0, +\infty[)$ .

Therefore, passing to the limit in (5.12) gives

$$\frac{\partial}{\partial t} U(\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}) \leq 0 \text{ in } \mathcal{D}'(\mathbb{R}^d \times ]0, +\infty[),$$

which gives the result.  $\square$

Therefore, the distributional solutions of (1.1) constructed by the method of vanishing viscosity satisfy the entropy condition (5.10) for all entropy functions  $U$ . This leads us to introduce the definition of an *entropy solution*.

*Definition 5.2*

A weak solution  $\mathbf{u}$  of (1.1), (1.2) is called an entropy solution if  $\mathbf{u}$  satisfies, for all entropy functions  $U$  of (1.1) and for all test functions  $\varphi \in \mathbf{C}_0^1(\mathbb{R}^d \times [0, \infty[)$ ,  $\varphi \geq 0$ ,

$$\begin{aligned} & \int_0^\infty \int_{\mathbb{R}^d} \left( U(\mathbf{u}) \frac{\partial \varphi}{\partial t} + \sum_{j=1}^d F_j(\mathbf{u}) \frac{\partial \varphi}{\partial x_j} \right) d\mathbf{x} dt \\ & + \int_{\mathbb{R}^d} U(\mathbf{u}_0(\mathbf{x})) \varphi(\mathbf{x}, 0) d\mathbf{x} \geq 0. \end{aligned} \quad (5.13)$$

By arguing as in the proof of Theorem 4.1, it is a simple matter to check that a piecewise  $C^1$  function  $\mathbf{u}$  is an entropy solution of (1.1), (1.2) if

- (i)  $\mathbf{u}$  is a classical solution of (1.1) in the domains where  $\mathbf{u}$  is  $C^1$  and satisfies (1.2) a.e.,
- (ii)  $\mathbf{u}$  satisfies the Rankine-Hugoniot condition (4.8),
- (iii)  $\mathbf{u}$  satisfies the jump inequality

$$n_t[U(\mathbf{u})] + \sum_{j=1}^d n_{x_j}[F_j(\mathbf{u})] \leq 0 \quad (5.14)$$

along the surfaces of discontinuity, if  $\frac{\mathbf{n}}{|\mathbf{n}|}$  is the outward (unit) normal vector pointing in the  $D_+$  direction.

Note that in the one-dimensional case, (5.14) becomes

$$[F(\mathbf{u})] \leq s[U(\mathbf{u})], \quad (5.15)$$

where  $s$  is again given by (4.10).

We shall encounter in Chap. II (Sect. 5) other criteria (which, however, coincide in the most usual cases) for selecting admissible solutions.

*Remark 5.1.* We can easily extend the definitions of weak and entropy solutions to systems with source terms (2.23): the Rankine-Hugoniot condition (4.8) remains unchanged, while inequality (5.10) should be replaced by

$$\frac{\partial}{\partial t} U(\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}) \leq U'(\mathbf{u}) \mathbf{s}(\mathbf{u}) \quad \text{in } \mathcal{D}',$$

where  $U$  and  $F_j$  satisfy the same conditions (5.2).  $\square$

*Remark 5.2.* It would be of interest to consider more general viscous perturbations of the system (1.1). In fact, for a compressible fluid, if we take into account the effect of viscosity together with heat conduction, the Euler equations of gas dynamics (2.1) are replaced by the Navier-Stokes (compressible) equations

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \sum_{j=1}^3 \frac{\partial}{\partial x_j} (\rho u_j) = 0, \\ \frac{\partial}{\partial t} (\rho u_i) + \sum_{j=1}^3 \frac{\partial}{\partial x_j} (\rho u_i u_j + p \delta_{ij} - \tau_{ij}) = 0, \quad 1 \leq i \leq 3, \\ \frac{\partial}{\partial t} (\rho e) + \sum_{j=1}^3 \frac{\partial}{\partial x_j} \left( (\rho e + p) u_j - \sum_{\ell=1}^3 \tau_{j\ell} u_\ell - \kappa \frac{\partial T}{\partial x_j} \right) = 0, \end{array} \right. \quad (5.16)$$

with

$$\begin{aligned} \tau_{ij} &= \gamma(\operatorname{div} \mathbf{u}) \delta_{ij} + 2\mu D_{ij}(\mathbf{u}), \\ D_{ij}(\mathbf{u}) &= \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right), \quad 1 \leq i, j \leq 3. \end{aligned} \quad (5.17)$$

In (5.16) and (5.17),  $T$  is the temperature,  $D(\mathbf{u}) = (D_{ij}(\mathbf{u}))$  is the deformation tensor,  $\gamma$  and  $\mu$  are Lamé coefficients of viscosity, and  $\kappa$  is the coefficient of thermal conductivity. Usually, the coefficients  $\gamma, \mu$ , and  $\kappa$  depend only on  $T$ . Now, we need to add to (5.16), (5.17) two equations of state of the form

$$p = p(\rho, \varepsilon), \quad T = T(\rho, \varepsilon). \quad (5.18)$$

Note that it may be easier to use  $\rho$  and  $T$  as the independent thermodynamic variables so that the equations of state become

$$p = p(\rho, T), \quad \varepsilon = \varepsilon(\rho, T).$$

The simplest model is obtained by taking

$$p = (\gamma - 1)\rho \varepsilon, \quad \varepsilon = C_v T$$

and assuming that the coefficients  $\gamma, \mu$ , and  $\kappa$  are constant.

By using again the notations (2.2) and setting

$$\mathbf{Q}_j \left( u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \frac{\partial u}{\partial x_3} \right) = \begin{pmatrix} 0 \\ \tau_{j1} \\ \tau_{j2} \\ \tau_{j3} \\ \sum_{\ell=1}^3 \tau_{j\ell} u_\ell + k \frac{\partial T}{\partial x_j} \end{pmatrix}, \quad 1 \leq j \leq 3,$$

the Navier-Stokes equations can be written in the form

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

Observe that we have

$$\mathbf{Q}_j = \sum_{j=1}^3 \mathbf{A}_{j\ell}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x_\ell}$$

for some  $p \times p$  matrices  $\mathbf{A}_{j\ell}(\mathbf{u})$ . Hence, (5.16) becomes

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^3 \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) - \sum_{j,\ell=1}^3 \frac{\partial}{\partial x_j} \left( \mathbf{A}_{j\ell}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x_\ell} \right) = \mathbf{0}. \quad (5.19)$$

Therefore, in the general case, a realistic situation consists in introducing instead of (5.7) the following viscous perturbation of the system (1.1):

$$\frac{\partial \mathbf{u}_\varepsilon}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}_\varepsilon) - \varepsilon \sum_{j,\ell=1}^d \frac{\partial}{\partial x_j} \left( \mathbf{A}_{j\ell}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_\ell} \right) = \mathbf{0}, \quad (5.20)$$

where the  $\mathbf{A}_{j\ell}$  are  $p \times p$  matrix-valued functions. Now, applying  $U'(\mathbf{u}_\varepsilon)$  to (5.20), we obtain

$$\frac{\partial}{\partial t} U(\mathbf{u}_\varepsilon) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}_\varepsilon) - \varepsilon \sum_{j,\ell=1}^d U'(\mathbf{u}_\varepsilon) \frac{\partial}{\partial x_j} \left( \mathbf{A}_{j\ell}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_\ell} \right) = 0.$$

We have

$$\begin{aligned} U'(\mathbf{u}_\varepsilon) \frac{\partial}{\partial x_j} \left( \mathbf{A}_{j\ell}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_\ell} \right) &= \frac{\partial}{\partial x_j} \left( U'(\mathbf{u}_\varepsilon) \mathbf{A}_{j\ell}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_\ell} \right) \\ &\quad - \left( \frac{\partial \mathbf{u}_\varepsilon}{\partial x_j} \right)^T U''(\mathbf{u}_\varepsilon) \mathbf{A}_{j\ell}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_\ell}. \end{aligned}$$

Hence, if we assume that the augmented  $dp \times dp$  matrix

$$\begin{pmatrix} U''(\mathbf{u})\mathbf{A}_{11} & \dots & U''(\mathbf{u})\mathbf{A}_{1d}(\mathbf{u}) \\ \vdots & & \vdots \\ U''(\mathbf{u})\mathbf{A}_{d1} & \dots & U''(\mathbf{u})\mathbf{A}_{dd}(\mathbf{u}) \end{pmatrix}$$

is symmetric and semi-positive-definite, we get

$$\frac{\partial}{\partial t} U(\mathbf{u}_\varepsilon) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F_j(\mathbf{u}_\varepsilon) \leq \varepsilon \sum_{j,\ell=1}^d \frac{\partial}{\partial x_j} \left( U'(\mathbf{u}_\varepsilon)\mathbf{A}_{j\ell}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_\ell} \right). \quad (5.21)$$

This property holds indeed in the case of the gas dynamics equations (see Dutt [453]). The problem is then to pass to the limit in the inequality (5.21) as  $\varepsilon$  tends to zero. This can be done in the following simple case:

- (i)  $\mathbf{A}_{j\ell} = \mathbf{D} \delta_{j\ell}$ , where the viscosity matrix  $\mathbf{D}$  is a  $p \times p$  symmetric positive-definite constant matrix (and  $\delta_{j\ell}$  the Kronecker symbol),
- (ii)  $U$  is a strictly convex positive entropy with

$$U'' \mathbf{D} \geq \alpha \quad \text{for some } \alpha > 0. \quad (5.22)$$

Then one can prove that the conclusion of Theorem 5.3 holds (see Harten [590]).  $\square$

*Remark 5.3.* For more general viscous regularization, see [563]. For general systems, without assuming that  $\mathbf{D}$  is constant, the positive-definiteness (5.22) of the matrix  $U'' \mathbf{D}$  is Mock's assumption of *admissibility*. It is also involved in the study of boundary conditions for the initial boundary value problem (see Chap. VI, Sect. 2.2), obtained with the vanishing viscosity method (see Gisclon and Serre [518] and Gisclon [517]), i.e., by studying the limit as  $\varepsilon \rightarrow 0$  of the solution  $\mathbf{u}_\varepsilon$  of the viscous system

$$\frac{\partial \mathbf{u}_\varepsilon}{\partial t} + \sum_{j=1}^3 \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}_\varepsilon) - \varepsilon \sum_{j=1}^3 \frac{\partial}{\partial x_j} \left( \mathbf{D}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x_j} \right) = \mathbf{0} \quad (5.23)$$

together with initial and boundary conditions. When the matrix  $\mathbf{D}$  satisfies (5.22), which is more precisely stated as for any compact  $K$  in  $\mathbb{R}^p$ , there exists  $\alpha_K > 0$  such that for any  $\mathbf{u} \in K$

$$(U''(\mathbf{u})\mathbf{v}, \mathbf{D}(\mathbf{u})\mathbf{v}) \geq \alpha_K |\mathbf{v}|^2, \forall \mathbf{v} \in \mathbb{R}^p,$$

it is called dissipative. Note that the so-called residual boundary conditions for the limit solution depend on the matrix  $\mathbf{D}$ .

We shall also meet a more general related assumption that is linked to the well-posedness of the viscous problem (5.23) linearized at a state  $\mathbf{u}_0$  and to the existence of viscous shock profiles (see Chap. II, Remark 5.2). It consists of the requirement that there exists a positive-definite symmetric matrix  $\mathbf{M}(\mathbf{u}_0)$

such that  $\mathbf{M}(\mathbf{u}_0)\mathbf{A}(\mathbf{u}_0)$  is symmetric and such that  $\mathbf{M}(\mathbf{u}_0)\mathbf{D}(\mathbf{u}_0)$  is positive-definite. This condition guarantees that the viscosity matrix is *strictly stable* (see Majda and Pego [843]).  $\square$

*Remark 5.4.* Assumptions (5.8), (5.9) are satisfied in the scalar case (Godlewski and Raviart [539], hereafter noted G.R., Chapter 2, Section 3). For general systems, it may be that only an  $\mathbf{L}^\infty$  estimate on  $\mathbf{u}_\varepsilon$  is available, which does not allow us to pass to the limit in the nonlinear term  $\mathbf{f}(\mathbf{u}_\varepsilon)$  when  $\varepsilon$  tends to 0, so that we cannot use Theorem 5.3. More powerful tools are provided by compensated compactness and measure-valued solutions (see Chap. V, Sect. 4.2.3 for the definition). We refer, for instance, to Murat [887], Tartar [1106, 1107], DiPerna [424, 425], Serre [1032], Chen [283].  $\square$

It is conjectured that an entropy solution of (1.1), (1.2) is necessarily unique. This is indeed proved in the case of a scalar equation ( $p = 1$ ), as stated in Theorem 5.4. For general data, the conjecture is still widely open in the case of systems even in one dimension ( $d = 1, p \geq 2$ ). We shall only prove, in Chap. II, Sect. 6, a uniqueness result for the Riemann problem (1.1), (1.3) in the case of a “convex” system and for nearby states. (We shall not consider more general data, for which we refer to the notes at the end of Chap. II). The Riemann problem is involved in many approximation schemes, which justifies a thorough study of both the theoretical aspects and some specific examples (such as the Euler system in Chap. III).

### 5.3 Existence and Uniqueness of the Entropy Solution in the Scalar Case

A thorough proof of the following theorem can be found in G.R., Chapter 2 (Section 5) [539]. Existence is obtained by the vanishing viscosity method, while uniqueness follows from Kruzhkov’s result, which states moreover that the solution has a finite domain of dependence corresponding to a finite speed of propagation.

*Theorem 5.4*

Assume that the function  $u_0$  belongs to  $\mathbf{L}^\infty(\mathbb{R}^d)$ . Then, Problem (1.1), (1.2) has a unique entropy solution  $u \in \mathbf{L}^\infty(\mathbb{R}^d \times (0, T))$ . This solution satisfies for almost all  $t \geq 0$

$$\|u(\cdot, t)\|_{\mathbf{L}^\infty(\mathbb{R}^d)} \leq \|u_0\|_{\mathbf{L}^\infty(\mathbb{R}^d)}. \quad (5.24)$$

Moreover, if  $u$  and  $v$  are the entropy solutions of (1.1) associated with the initial conditions  $u_0$  and  $v_0$ , respectively, we have

$$u_0 \geq v_0 \implies u(\cdot, t) \geq v(\cdot, t) \text{ a.e.} \quad (5.25)$$

Finally, if  $u_0$  belongs to  $\mathbf{L}^\infty(\mathbb{R}^d) \cap \mathbf{BV}(\mathbb{R}^d)$ , then  $u(\cdot, t)$  belongs to  $\mathbf{BV}(\mathbb{R}^d)$  with

$$TV(u(\cdot, t)) \leq TV(u_0). \quad (5.26)$$

*Remark 5.5.* The proof of uniqueness uses the well-known Kruzhkov's entropy pair

$$U(u) = |u - k|, \quad F_j(u) = \operatorname{sgn}(u - k)(f_j(u) - f_j(k)), \quad j = 1, \dots, d.$$

This important result extends to scalar equations with source term (2.34) assuming the dependence on  $\mathbf{x}$  is smooth (see Kruzhkov [719]).  $\square$

For what concerns the Riemann problem in the scalar one-dimensional case, the unique solution can be constructed explicitly in the general case (G.R., Chapter 2, Section 6 [539]); it is piecewise smooth, self-similar, and monotonic and is usually written in the form

$$u(x, t) = w_R\left(\frac{x}{t}; u_L, u_R\right). \quad (5.27)$$

We recall the formula for  $w_R$  only when the function  $f$  is strictly convex, since in the general case, it is difficult to exhibit a compact formula (moreover, in the case of a system, we shall assume in Chap. II, Sect. 6, that the system is *convex*, in which case the elementary waves that are involved in the solution of the Riemann problem are also rarefactions and shocks for what concerns the genuinely nonlinear fields). In this case ( $f'' > 0$ ), it is easy to prove that only decreasing shocks are admissible; moreover, the characteristics are going into the shock (see Fig. 5.1(i)). Any increasing discontinuity introduced in the initial data is immediately spread out. Then the entropy solution consists of either a rarefaction when  $u < v$  (Fig. 5.1(ii)), which is found by looking for a smooth self-similar solution of (1.5) connecting  $u$  and  $v$ , or a shock when  $u > v$  (Fig. 5.1(i)), which propagates at a constant velocity  $s$  given by the Rankine-Hugoniot condition (4.11).

The function  $\xi \mapsto w_R(\xi; u, v)$  solution of the scalar strictly convex Riemann problem is precisely defined by

if  $u < v$

$$w_R(\xi; u, v) = \begin{cases} u, & \xi \leq a(u) \\ a^{-1}(\xi), & a(u) \leq \frac{x}{t} \leq a(v), \quad a(u) = f'(u) \\ v, & \xi \geq a(v), \end{cases} \quad (5.28)$$

if  $u = v \quad w_R(\xi; u, v) = u,$

if  $u > v$

$$w_R(\xi; u, v) = \begin{cases} u & \text{if } \xi < s \\ v & \text{if } \xi > s, \end{cases}$$

where

$$s = s(u, v) = \frac{f(u) - f(v)}{u - v}$$

is the speed of propagation of the discontinuity.

## Notes

The examples presented in Sect. 1 will be examined in the next chapters, at the end of which the reader will find bibliographical notes. Let us just mention some other examples: Buckley-Leverett equations and mathematical models for oil recovery (Temple [1111], Schaeffer and Shearer [1013]; Shearer [1048] also [506]; Holden [627]; Tveito and Winther [1140]); traffic flow ([1188, Sec. 3.1], Aw-Rascle [68]); Holden and Risebro [625] and the references therein); propagation of waves in elasticity theory (Keyfitz and Kranzer [689]).

Further applications are mentioned in the Notes at the end of Chaps. IV and V. For examples as well as for most theoretical results, we refer particularly to Serre's textbook [1038, 1039].

The formation of singularities was shown with a simple example; for more general situations, see John [669]. We did not mention the “equal area” rule, which, in a single conservation law, gives a geometric way of constructing the correct position for the discontinuity ([1188, section 2.8]), nor the “transport-collapse” operator of Brenier [191]. In general, definitions or results concerning scalar equations and characteristics can be found in John's book [670], in the recent books of Taylor [1110], and as we already mentioned those of Serre [1039], also specifically in Jeffrey [650] and Taniuti and Nishihara [1105]. In Smoller's book (Chapter 16 [1066]), one finds a thoroughly different proof of Theorem 5.4 (concerning the convex case  $f'' > 0$  only). This proof involves the convergence of a finite difference scheme and uses another way of writing the entropy condition (the “spreading estimate” or one-sided Lipschitz condition) due to Oleinik [914] and a nonlinear version of the Holmgren method via the “adjoint” equation. Concerning the smoothness of the entropy solution, see Tadmor and Tassa [1095] and the references therein.

Some important papers concerned with the notion of entropy are those of Friedrichs and Lax [494], Lax [744], Dafermos [379], T.-P. Liu [822], and Bar-dos [83] and Tadmor [1091]; see also the references in Sect. 5.3.2, in Chap. II. Note that we have presented here a “mathematical” notion of entropy associated with a pair of entropy-entropy flux. We shall consider more geometrical conditions (for the characteristics) for systems in Chap. II, Sect. 5. Concerning the vanishing viscosity method, besides the scalar case, which is thoroughly studied in Godlewski and Raviart [539], the reader will find further references at the end of Chap. II.

The scalar Riemann problem is also treated in Chang and Hsiao [278], where one finds, moreover, the study of wave interaction.

*Note Added in the Second Edition*

Since the first edition, several books have been published, some of them more dedicated to theoretical analysis, among which the already cited books of D. Serre (they have been translated [1039, 1041]) [1043], C. Dafermos [384], A. Bressan [197], Ph. LeFloch [753], B. Perthame [944], also L. Tartar [1109], P.-L. Lions [810]; to the mathematical aspects of fluid dynamics in the Handbook edited by S. Friedlander and D. Serre [493]; others are more oriented toward numerical approximation and applications, B. Cockburn et al. [320], J.W. Thomas [1120], R. Eymard and coauthors in the Handbook of numerical analysis [466], also the recent Handbook of numerical methods for hyperbolic problems [19, 20], R. LeVeque [777], E. Toro [1127], F. Bouchut [163], J.A. Trangenstein [1131]; B. Després [416, 417], D. Kröner [713], V.D. Sharma [1046], J.S. Hesthaven [612]; see also T. Barth and M. Ohlberger [90]; L. Gosse [553] and with D. Amadori [33] for approximation of balance laws, H. Holden and N.H. Risebro [626] for wave front tracking technique; others to modeling, for instance, V. Giovangigli [515] for multicomponent flow, M. Garavello and B. Piccoli [507] for traffic flow; then Y. Zheng [1222] for two-dimensional Riemann problems, the list is certainly not exhaustive. There are also many interesting papers concerning theoretical results in the scalar case [926, 927], some linked to the kinetic representation [812], and it is out of scope to cite them all. More references can be found at the end of each chapter.



## II

# Nonlinear Hyperbolic Systems in One Space Dimension

The main goal of this chapter is to study the Riemann problem for a general nonlinear hyperbolic system of conservation laws in one space dimension. We begin by considering the case of a linear hyperbolic system with constant coefficients for which the Riemann problem is easily solved. Next, in the nonlinear case, we introduce the notions of rarefaction waves, shock waves, and contact discontinuities, which play an essential role in the explicit construction of the solution of the Riemann problem. These notions are illustrated on the examples of the  $p$ -system and the gas dynamics equations. Then, we prove the local existence of an entropy solution of the Riemann problem for a general system in the sense that the initial states are sufficiently close. In fact, in the case of the  $p$ -system, we are able to prove that the Riemann problem is always solvable. We shall show in the next chapter that this is also true for the gas dynamics equations. These two basic chapters present the results with detailed proofs, which could be easily shortened for a reader already familiar with the subject.

## 1 Linear Hyperbolic Systems with Constant Coefficients

We begin by considering the first-order linear system

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, \quad x \in \mathbb{R}, \quad t > 0, \quad (1.1)$$

where  $\mathbf{u} = (u_1, \dots, u_p)^T$  is a  $p$ -vector (all the vectors of  $\mathbb{R}^p$  will be considered as column vectors) and  $\mathbf{A}$  is a  $p \times p$  constant matrix. We assume that the system is strictly hyperbolic, i.e., the matrix  $\mathbf{A}$  has  $p$  distinct real eigenvalues arranged in increasing order

$$\lambda_1 < \lambda_2 < \dots < \lambda_p.$$

With each eigenvalue  $\lambda_k$ , we associate a right eigenvector  $\mathbf{r}_k \in \mathbb{R}^p$

$$\mathbf{A}\mathbf{r}_k = \lambda_k \mathbf{r}_k \quad (1.2)$$

and a “left” eigenvector  $\mathbf{l}_k^T$

$$\mathbf{l}_k^T \mathbf{A} = \lambda_k \mathbf{l}_k^T \quad (1.3)$$

i.e.,  $\mathbf{l}_k$  is an eigenvector of  $\mathbf{A}^T$ . Since the eigenvalues are distinct, the eigenvectors  $\mathbf{r}_k, 1 \leq k \leq p$ , form a basis of  $\mathbb{R}^p$  and we have

$$\mathbf{l}_j^T \mathbf{r}_k = \mathbf{l}_j \cdot \mathbf{r}_k = 0, \quad j \neq k$$

(we use either the dot  $\cdot$  for the scalar product of the two vectors or  $\mathbf{l}_j^T \mathbf{r}_k$ , which denotes the product of the  $1 \times p$  and  $p \times 1$  matrices). Moreover, we can normalize the vectors  $\mathbf{l}_k^T$  in such a way that

$$\mathbf{l}_k^T \mathbf{r}_k = 1.$$

Hence, using the Kronecker delta symbol, we obtain

$$\mathbf{l}_j^T \mathbf{r}_k = \delta_j^k, \quad 1 \leq j, k \leq p. \quad (1.4)$$

Now, setting

$$\mathbf{u} = \sum_{k=1}^p \alpha_k \mathbf{r}_k, \quad \alpha_k = \mathbf{l}_k^T \mathbf{u},$$

we have

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \sum_{k=1}^p \left( \frac{\partial \alpha_k}{\partial t} + \lambda_k \frac{\partial \alpha_k}{\partial x} \right) \mathbf{r}_k$$

We can write the formulas in a more compact form by introducing the matrix  $\mathbf{T}$  with columns the eigenvectors  $\mathbf{r}_k$ ’s, the rows of  $\mathbf{T}^{-1}$  are the  $\mathbf{l}_k^T$ ’s, so that if we note  $\mathbf{v}$  the vector with component the  $\alpha_k$ ’s, we have  $\mathbf{u} = \mathbf{T}\mathbf{v}$ . Now  $\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \Lambda$  is a diagonal matrix with eigenvalues  $\lambda_k$ . Then multiplying (1.1) on the left by  $\mathbf{T}^{-1}$  gives

$$\mathbf{T}^{-1} \frac{\partial}{\partial t} \mathbf{u} + \mathbf{T}^{-1} \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \frac{\partial \mathbf{v}}{\partial t} + \Lambda \frac{\partial \mathbf{v}}{\partial x} = \mathbf{0}.$$

Anyway we get that the first-order system (1.1) is equivalent to the  $p$  independent scalar first-order equations

$$\frac{\partial \alpha_k}{\partial t} + \lambda_k \frac{\partial \alpha_k}{\partial x} = 0, \quad 1 \leq k \leq p.$$

Thus, we can derive an explicit expression for the solution  $\mathbf{u}$  of the Cauchy problem

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, & x \in \mathbb{R}, \quad t > 0, \\ \mathbf{u}(x, 0) = \mathbf{u}_0(x). \end{cases} \quad (1.5)$$

Indeed, setting

$$\alpha_{k0}(x) = \mathbf{l}_k^T \mathbf{u}_0(x),$$

equivalently written  $\mathbf{v}_0(x) = \mathbf{T}^{-1}\mathbf{u}_0(x)$ , we obtain

$$\alpha_k(x, t) = \alpha_{k0}(x - \lambda_k t) = \mathbf{l}_k^T \mathbf{u}_0(x - \lambda_k t), \quad 1 \leq k \leq p,$$

and therefore

$$\mathbf{u}(x, t) = \sum_{k=1}^p \mathbf{l}_k^T \mathbf{u}_0(x - \lambda_k t) \mathbf{r}_k. \quad (1.6)$$

Consider in particular the Riemann problem for the system (1.1) corresponding to the initial condition

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0. \end{cases} \quad (1.7)$$

If we define  $\alpha_{kR}$  and  $\alpha_{kL}$ ,  $1 \leq k \leq p$ , by

$$\mathbf{u}_L = \sum_{k=1}^p \alpha_{kL} \mathbf{r}_k, \quad \mathbf{u}_R = \sum_{k=1}^p \alpha_{kR} \mathbf{r}_k,$$

we obtain

$$\alpha_k(x, t) = \begin{cases} \alpha_{kL}, & x < \lambda_k t, \\ \alpha_{kR}, & x > \lambda_k t, \end{cases}$$

so that the solution  $\mathbf{u}$  of the Riemann problem (1.5), (1.7) is self-similar, i.e., of the form

$$\mathbf{u}(x, t) = \mathbf{w}_R \left( \frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R \right). \quad (1.8)$$

Moreover, for  $\lambda_m < \frac{x}{t} < \lambda_{m+1}$ ,  $\mathbf{u}$  takes the constant value

$$\mathbf{w}_m = \sum_{k=1}^m \alpha_{kR} \mathbf{r}_k + \sum_{k=m+1}^p \alpha_{kL} \mathbf{r}_k, \quad 0 \leq m \leq p \quad (1.9)$$

with the convention  $\lambda_0 = -\infty, \lambda_{p+1} = +\infty$ . Hence, we have

$$\mathbf{w}_R \left( \frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R \right) = \begin{cases} \mathbf{w}_0 = \mathbf{u}_L, & \frac{x}{t} < \lambda_1, \\ \mathbf{w}_1, & \lambda_1 < \frac{x}{t} < \lambda_2, \\ \vdots \\ \mathbf{w}_{p-1}, & \lambda_{p-1} < \frac{x}{t} < \lambda_p, \\ \mathbf{w}_p = \mathbf{u}_R, & \frac{x}{t} > \lambda_p, \end{cases} \quad (1.10)$$

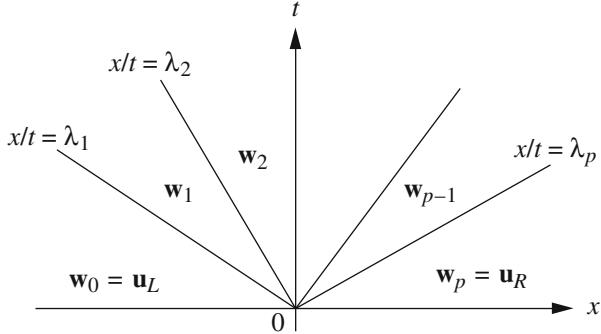
which shows that, in general, the initial discontinuity breaks up into  $p$  discontinuity waves, which propagate with the characteristic speeds  $\lambda_k$ ,  $1 \leq k \leq p$  (see Fig. 1.1). Note that the intermediate states  $\mathbf{w}_m$  satisfy

$$\mathbf{w}_m - \mathbf{w}_{m-1} = (\alpha_{mR} - \alpha_{mL}) \mathbf{r}_m$$

and therefore

$$\mathbf{A}(\mathbf{w}_m - \mathbf{w}_{m-1}) = \lambda_m(\mathbf{w}_m - \mathbf{w}_{m-1}).$$

Thus, across the line of discontinuity  $x = \lambda_m t$ , the Rankine-Hugoniot jump condition (4.11) of the Chap. I is indeed satisfied.



**Fig. 1.1** Solution of the Riemann problem for a linear system

*Remark 1.1.* We have assumed that the eigenvalues are distinct to simplify the presentation. The result can be easily extended provided we assume that the matrix  $\mathbf{A}$  has  $p$  real eigenvalues and a basis of eigenvectors. If, for instance, two eigenvalues coincide, we have only  $p$  constant states in formula (1.10) and  $p - 1$  discontinuity waves. Just to give an example, assume  $\lambda_1 = \lambda_2 < \lambda_3 < \dots < \lambda_p$ . Then the intermediate states correspond to those denoted  $\mathbf{w}_2, \dots, \mathbf{w}_{p-1}$ , and in the lines which follow, the only change is in  $\mathbf{w}_2 - \mathbf{u}_g = (\alpha_{1,R} - \alpha_{1,L})\mathbf{r}_1 + (\alpha_{2,R} - \alpha_{2,L})\mathbf{r}_2$ , and the Rankine-Hugoniot condition is still satisfied.  $\square$

## 2 The Nonlinear Case, Definitions and Examples

We turn now to nonlinear hyperbolic systems. Let  $\Omega$  be an open subset of  $\mathbb{R}^p$  and  $f : \Omega \rightarrow \mathbb{R}^p$  be a sufficiently smooth function (of class  $C^2$  at least). We consider the nonlinear system of conservation laws

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = \mathbf{0}, \quad x \in \mathbb{R}, \quad t > 0, \quad (2.1)$$

where  $\mathbf{u} = (u_1, \dots, u_p)^T$  and  $\mathbf{f}(\mathbf{u}) = (f_1(\mathbf{u}), \dots, f_p(\mathbf{u}))^T$ . Let us assume for simplicity that the system (2.1) is strictly hyperbolic, i.e., for any state  $\mathbf{u} \in \Omega$ , the  $p \times p$  Jacobian matrix

$$\mathbf{A}(\mathbf{u}) = \left( \frac{\partial f_i}{\partial u_j}(\mathbf{u}) \right)_{1 \leq i, j \leq p}$$

has  $p$  distinct real eigenvalues

$$\lambda_1(\mathbf{u}) < \lambda_2(\mathbf{u}) < \cdots < \lambda_p(\mathbf{u}).$$

With each eigenvalue  $\lambda_k(\mathbf{u})$ , we associate a right eigenvector  $\mathbf{r}_k(\mathbf{u})$

$$\mathbf{A}(\mathbf{u})\mathbf{r}_k(\mathbf{u}) = \lambda_k(\mathbf{u})\mathbf{r}_k(\mathbf{u}) \quad (2.2)$$

and a “left eigenvector”  $\mathbf{l}_k(\mathbf{u})$

$$\mathbf{l}_k(\mathbf{u})^T \mathbf{A}(\mathbf{u}) = \lambda_k(\mathbf{u}) \mathbf{l}_k(\mathbf{u})^T \quad (2.3)$$

i.e.,  $\mathbf{l}_k(\mathbf{u})$  is an eigenvector of  $\mathbf{A}(\mathbf{u})^T$ . Since the eigenvalues are distinct,  $(\mathbf{l}_k(\mathbf{u}))_k$  is a dual basis of  $(\mathbf{r}_k(\mathbf{u}))_k$  and

$$\mathbf{l}_j(\mathbf{u})^T \mathbf{r}_k(\mathbf{u}) = \mathbf{l}_j(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0, \quad k \neq j. \quad (2.4)$$

In fact, most of the results of the following sections hold for a hyperbolic system, provided the eigenvectors are complete and the eigenvalues have constant multiplicity (see Remark 6.1). If we assume that  $\mathbf{f}$  is a  $C^m$  function, we obtain that  $\lambda_k, \mathbf{r}_k, \mathbf{l}_k$  are  $C^{m-1}$  functions of  $\mathbf{u}$  ( $C^1$  functions at least). The pair  $(\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u}))$  is called the  $k$ th characteristic field.

*Definition 2.1*

The  $k$ th characteristic field is said to be genuinely nonlinear if

$$D\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \neq 0, \quad \forall \mathbf{u} \in \Omega. \quad (2.5)$$

The  $k$ -th characteristic field is said to be linearly degenerate if

$$D\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0, \quad \forall \mathbf{u} \in \Omega. \quad (2.6)$$

In (2.5) and (2.6),  $D\lambda_k(\mathbf{u}) = \lambda'_k(\mathbf{u}) \in \mathcal{L}(\mathbb{R}^p; \mathbb{R})$  denotes the derivative of  $\lambda_k(\mathbf{u})$  and can be identified with a vector  $\nabla \lambda_k(\mathbf{u})$  of  $\mathbb{R}^p$ .

*Example 2.1.* Consider first the case  $p = 1$ , i.e., (2.1) is a scalar equation. Then, we have

$$\lambda(u) = f'(u),$$

which we note  $a(u)$ . Hence, we are in the genuinely nonlinear case if and only if  $a(u)$  does not vanish, i.e., if and only if  $f$  is a strictly convex or a strictly concave function. On the other hand, we are in the linearly degenerate case if and only if  $a(u) = a$  does not depend on  $u$ , which corresponds to the case of a linear hyperbolic equation with constant coefficients.  $\square$

*Example 2.2. The p-system.* Let us consider again the  $p$ -system introduced in the Chap. I, Example 1.3:

$$\frac{\partial \mathbf{w}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{w}) = \mathbf{0}, \quad \mathbf{w} = \begin{pmatrix} v \\ u \end{pmatrix}, \quad \mathbf{f}(\mathbf{w}) = \begin{pmatrix} -u \\ p(v) \end{pmatrix}. \quad (2.7)$$

We have already noticed that the  $p$ -system is strictly hyperbolic provided that we assume  $p'(v) < 0$ . In that case, since the Jacobian matrix of  $\mathbf{f}$  depends only on  $v$ , we can note  $\mathbf{A}(\mathbf{w}) = \mathbf{A}(v)$ , and

$$\mathbf{A}(v) = \begin{pmatrix} 0 & -1 \\ p'(v) & 0 \end{pmatrix}$$

has two real distinct eigenvalues

$$\lambda_1(v) = -\sqrt{-p'(v)} < 0 < \lambda_2(v) = \sqrt{-p'(v)}. \quad (2.8)$$

Moreover, the corresponding right eigenvectors can be chosen as

$$\mathbf{r}_1(v) = \begin{pmatrix} 1 \\ \sqrt{-p'(v)} \end{pmatrix}, \quad \mathbf{r}_2(v) = \begin{pmatrix} 1 \\ -\sqrt{-p'(v)} \end{pmatrix}. \quad (2.9)$$

Hence, we obtain

$$\nabla \lambda_1(v) = \begin{pmatrix} \frac{p''(v)}{2\sqrt{-p'(v)}} \\ 0 \end{pmatrix}, \quad \nabla \lambda_2(v) = \begin{pmatrix} -\frac{p''(v)}{2\sqrt{-p'(v)}} \\ 0 \end{pmatrix}$$

so that

$$\begin{aligned} D\lambda_1(v) \cdot \mathbf{r}_1(v) &= \nabla \lambda_1(v)^T \mathbf{r}_1(v) = \frac{p''(v)}{2\sqrt{-p'(v)}}, \\ D\lambda_2(v) \cdot \mathbf{r}_2(v) &= \nabla \lambda_2(v)^T \mathbf{r}_2(v) = -\frac{p''(v)}{2\sqrt{-p'(v)}}. \end{aligned}$$

If we suppose  $p''(v) > 0$ , we obtain that the two characteristic fields are genuinely nonlinear.  $\square$

## 2.1 Change of Variables, Change of Frame

### 2.1.1 Nonconservative Variables

In fact, in many applications, it can be easier when studying the characteristic fields to work on a nonconservative form of the nonlinear system (2.1). Indeed,

let  $\theta$  be a  $C^1$  diffeomorphism of an open subset  $\vartheta \subset \mathbb{R}^p$  onto  $\Omega$ . By making the change of dependent variables

$$\mathbf{u} = \theta(\mathbf{v}), \quad (2.10)$$

the differential system

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0} \quad (2.11)$$

becomes

$$D\theta(\mathbf{v}) \frac{\partial \mathbf{v}}{\partial t} + \mathbf{A}(\theta(\mathbf{v})) D\theta(\mathbf{v}) \frac{\partial \mathbf{v}}{\partial x} = \mathbf{0},$$

where  $D\theta(\mathbf{v}) = \theta'(\mathbf{v})$  denotes the Jacobian matrix of  $\theta(\mathbf{v})$  (we will use both notations  $D\theta$  or  $\theta'$  in an equivalent way), and therefore

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{B}(\mathbf{v}) \frac{\partial \mathbf{v}}{\partial x} = \mathbf{0} \quad (2.12)$$

with

$$\mathbf{B}(\mathbf{v}) = (D\theta(\mathbf{v}))^{-1} \mathbf{A}(\theta(\mathbf{v})) D\theta(\mathbf{v}).$$

Denote by  $\mu_k(\mathbf{v})$  and  $\mathbf{s}_k(\mathbf{v})$ ,  $1 \leq k \leq p$ , the eigenvalues and the corresponding right eigenvectors of the  $p \times p$  matrix  $\mathbf{B}(\mathbf{v})$

$$\mathbf{B}(\mathbf{v}) \mathbf{s}_k(\mathbf{v}) = \mu_k(\mathbf{v}) \mathbf{s}_k(\mathbf{v}).$$

By the similarity of the matrices  $\mathbf{A}(\theta(\mathbf{v}))$  and  $\mathbf{B}(\mathbf{v})$ , we obtain

$$\mu_k(\mathbf{v}) = \lambda_k(\theta(\mathbf{v})). \quad (2.13)$$

Moreover, we can take

$$\mathbf{s}_k(\mathbf{v}) = (D\theta(\mathbf{v}))^{-1} \mathbf{r}_k(\theta(\mathbf{v})); \quad (2.14)$$

indeed

$$\begin{aligned} \mathbf{B}(\mathbf{v}) \mathbf{s}_k(\mathbf{v}) &= (D\theta(\mathbf{v}))^{-1} \mathbf{A}(\theta(\mathbf{v})) D\theta(\mathbf{v}) (D\theta(\mathbf{v}))^{-1} \mathbf{r}_k(\theta(\mathbf{v})) \\ &= (D\theta(\mathbf{v}))^{-1} \mathbf{A}(\theta(\mathbf{v})) \mathbf{r}_k(\theta(\mathbf{v})) \\ &= \lambda_k(\theta(\mathbf{v})) (D\theta(\mathbf{v}))^{-1} \mathbf{r}_k(\theta(\mathbf{v})) = \mu_k(\mathbf{v}) \mathbf{s}_k(\mathbf{v}). \end{aligned}$$

Next, we note that

$$D\mu_k(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}) = D\lambda_k(\theta(\mathbf{v})) \cdot D\theta(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}) = D\lambda_k(\theta(\mathbf{v})) \cdot \mathbf{r}_k(\theta(\mathbf{v})),$$

i.e.,

$$D\mu_k(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}) = D\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}), \quad \mathbf{u} = \theta(\mathbf{v}). \quad (2.15)$$

Therefore, the notions of genuinely nonlinear or linearly degenerate characteristic fields do not depend on the chosen conservative or nonconservative form of the nonlinear hyperbolic system (2.1). This remark will enable us

to simplify greatly the analysis in the examples. In the sequel, we will not necessarily distinguish in the notations, and  $\lambda_k(\mathbf{u})$  and  $\mu_k(\mathbf{v})$  will be simply noted  $\lambda_k$ .

However, let us emphasize that if the nonconservative forms (2.11) and (2.12) are equivalent for smooth solutions of (2.1), for weak discontinuous solutions, only (2.11) is meaningful.

### 2.1.2 Characteristic Fields in the Eulerian and Lagrangian Frames

Let us go back to systems of the form

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho \Phi) + \frac{\partial}{\partial x}(\rho \Phi u + \mathbf{g}(\rho, \Phi)) = \mathbf{0}. \end{cases} \quad (2.16)$$

considered in the Chap. I (see (2.8), Sect. 2). In Lagrangian coordinates, (2.16) becomes

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial \Phi}{\partial t} + \frac{\partial}{\partial m} \mathbf{g}\left(\frac{1}{\tau}, \Phi\right) = \mathbf{0}, \end{cases} \quad (2.17)$$

where  $\tau = \frac{1}{\rho}$ . As expected, the two above systems (2.16) and (2.17) have strongly related mathematical properties. It will be convenient to set here

$$\mathbf{U} = \begin{pmatrix} \rho \\ \rho \Phi \end{pmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} \rho u \\ \rho \Phi u + \mathbf{g}(\rho, \Phi) \end{pmatrix} \quad (2.18)$$

and

$$\mathbf{V} = \begin{pmatrix} \tau \\ \Phi \end{pmatrix}, \quad \mathbf{G}(\mathbf{V}) = \begin{pmatrix} -u \\ \mathbf{g}\left(\frac{1}{\tau}, \Phi\right) \end{pmatrix}. \quad (2.19)$$

We first check that if  $\mathbf{U}$  lies in the set of states  $\Omega \subset \mathbb{R}^p$ , then  $\mathbf{V}$  belongs to  $\Omega$  too. Then, denote by  $\mathbf{A}(\mathbf{U})$  (resp.  $\mathbf{B}(\mathbf{V})$ ) the Jacobian matrix of  $\mathbf{F}(\mathbf{U})$  (resp.  $\mathbf{G}(\mathbf{V})$ ), we can state the following easy result.

*Lemma 2.1*

*The mapping  $\mathbf{U} \mapsto \mathbf{V}(\mathbf{U})$  is one-to-one for  $\rho > 0$ , and the Jacobian matrix  $D\mathbf{V}(\mathbf{U})$  and its inverse  $D\mathbf{U}(\mathbf{V})$  are given by*

$$D\mathbf{V}(\mathbf{U}) = \begin{pmatrix} -\frac{1}{\rho^2} & \mathbf{0} \\ -\frac{\Phi}{\rho} & \mathbf{I} \end{pmatrix}, \quad D\mathbf{U}(\mathbf{V}) = \begin{pmatrix} -\frac{1}{\tau^2} & \mathbf{0} \\ -\frac{\Phi}{\tau^2} & \mathbf{I} \end{pmatrix} \quad (2.20)$$

where  $\mathbf{I}$  is the  $(p-1) \times (p-1)$  identity matrix. Moreover, we have

$$\mathbf{B}(\mathbf{V}) = \frac{1}{\tau} \{ D\mathbf{V}(\mathbf{U}) \cdot \mathbf{A}(\mathbf{U}) \cdot D\mathbf{U}(\mathbf{V}) - u\mathbf{I} \}, \quad \mathbf{U} = \mathbf{U}(\mathbf{V}). \quad (2.21)$$

*Proof.* Differentiating

$$\mathbf{V}(\mathbf{U}) = \begin{pmatrix} \frac{1}{\rho} \\ \frac{\rho}{\rho\Phi} \\ \frac{\rho\Phi}{\rho} \end{pmatrix}$$

gives the expression (2.20) for  $D\mathbf{V}(\mathbf{U})$ . As a consequence, we obtain that the mapping  $\mathbf{U} \mapsto \mathbf{V}(\mathbf{U})$  is indeed one-to-one for  $\rho > 0$ . Inverting the matrix  $D\mathbf{V}(\mathbf{U})$  gives the expression (2.20) for  $D\mathbf{U}(\mathbf{V})$ . In fact, we may also notice that both mappings  $\mathbf{U} \mapsto \mathbf{V}(\mathbf{U})$  and  $\mathbf{V} \mapsto \mathbf{U}(\mathbf{V})$  can be written in the form  $(w_1, \bar{\mathbf{w}})^T \mapsto (\frac{1}{w_1^2}, \frac{\bar{\mathbf{w}}}{w_1})^T$  which explains why the Jacobian matrices have the same expressions.

On the other hand, it follows from (2.18) that

$$\mathbf{A}(\mathbf{U}) \doteq \mathbf{F}'(\mathbf{U}) = u\mathbf{I} + \mathbf{U} \cdot Du(\mathbf{U}) + \begin{pmatrix} \mathbf{0} \\ D\tilde{\mathbf{g}}(\mathbf{U}) \end{pmatrix}$$

where  $Du(\mathbf{U})$  is identified with the row matrix  $(\frac{\partial u}{\partial U_j})$  and  $\tilde{\mathbf{g}}(\mathbf{U}) = \mathbf{g}(\rho, \frac{\rho\Phi}{\rho})$ . Hence, we have

$$D\mathbf{V}(\mathbf{U}) \cdot \mathbf{A}(\mathbf{U}) = u D\mathbf{V}(\mathbf{U}) + D\mathbf{V}(\mathbf{U}) \cdot (\mathbf{U} \cdot Du(\mathbf{U})) + D\mathbf{V}(\mathbf{U}) \begin{pmatrix} \mathbf{0} \\ D\tilde{\mathbf{g}}(\mathbf{U}) \end{pmatrix}.$$

Using the expression (2.20) of  $D\mathbf{V}(\mathbf{U})$ , we obtain

$$D\mathbf{V}(\mathbf{U}) \cdot \mathbf{A}(\mathbf{U}) = u D\mathbf{V}(\mathbf{U}) + \frac{1}{\rho} \begin{pmatrix} -Du(\mathbf{U}) \\ D\tilde{\mathbf{f}}(\mathbf{U}) \end{pmatrix}$$

where

$$\tilde{\mathbf{f}}(\mathbf{U}) = \rho\Phi u + \tilde{\mathbf{g}}(\mathbf{U}).$$

Since

$$\mathbf{B}(\mathbf{V}) = D\mathbf{G}(\mathbf{V}) = \begin{pmatrix} -Du(\mathbf{U}) \\ D\tilde{\mathbf{f}}(\mathbf{U}) \end{pmatrix} \cdot D\mathbf{U}(\mathbf{V}),$$

we find

$$\mathbf{B}(\mathbf{V}) = \rho D\mathbf{V}(\mathbf{U}) \cdot (\mathbf{A}(\mathbf{U}) - u\mathbf{I}) \cdot D\mathbf{U}(\mathbf{V})$$

which yields (2.21).  $\square$

Denoting by  $(\lambda_k(\mathbf{U}), \mathbf{r}_k(\mathbf{U}))$  and  $(\mu_k(\mathbf{V}), \mathbf{s}_k(\mathbf{V}))$  the eigenpairs of the Jacobian matrices  $\mathbf{A}(\mathbf{U})$  and  $\mathbf{B}(\mathbf{V})$ , respectively, i.e.,

$$\mathbf{A}(\mathbf{U})\mathbf{r}_k(\mathbf{U}) = \lambda_k(\mathbf{U})\mathbf{r}_k(\mathbf{U}), \quad \mathbf{B}(\mathbf{V})\mathbf{s}_k(\mathbf{V}) = \mu_k(\mathbf{V})\mathbf{s}_k(\mathbf{V}),$$

we can now precisely confirm the correspondence between the characteristic fields induced by the change of frame.

*Theorem 2.1*

We have the relations

$$\begin{cases} \mu_k = \frac{1}{\tau}(\lambda_k(\mathbf{U}) - u), \\ \mathbf{s}_k(\mathbf{V}) = D\mathbf{V}(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U}) \end{cases} \quad (2.22)$$

where  $\mathbf{U} = \mathbf{U}(\mathbf{V})$ . If we assume in addition that  $U_2 = \Phi_1 = \rho u$ , then

$$D\mu_k(\mathbf{V}) \cdot \mathbf{s}_k(\mathbf{V}) = \frac{1}{\tau} D\lambda_k(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U}). \quad (2.23)$$

*Proof.* The first relation (2.22) follows immediately from (2.21). On the other hand, if  $\mathbf{s}_k(\mathbf{V})$  is defined as in the second Eq. (2.22), we have

$$\begin{aligned} \mathbf{B}(\mathbf{V}) \cdot \mathbf{s}_k(\mathbf{V}) &= \frac{1}{\tau} (D\mathbf{V}(\mathbf{U}) \cdot (\mathbf{A}(\mathbf{U}) \cdot D\mathbf{U}(\mathbf{V}) - u\mathbf{I}) \cdot D\mathbf{V}(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U})) = \\ &= \frac{1}{\tau} D\mathbf{V}(\mathbf{U}) \cdot (\mathbf{A}(\mathbf{U}) - u\mathbf{I}) \cdot \mathbf{r}_k(\mathbf{U}) = \frac{1}{\tau} (\lambda_k(\mathbf{U}) - u) D\mathbf{V}(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U}) = \\ &= \mu_k(\mathbf{V}) \mathbf{s}_k(\mathbf{V}). \end{aligned}$$

Next differentiating the first Eq. (2.22) gives

$$D\mu_k(\mathbf{V}) = D(\rho(\lambda_k - u))(\mathbf{U}) \cdot D\mathbf{U}(\mathbf{V})$$

so that

$$\begin{aligned} D\mu_k(\mathbf{V}) \cdot \mathbf{s}_k(\mathbf{V}) &= D(\rho(\lambda_k - u))(\mathbf{U}) \cdot D\mathbf{U}(\mathbf{V}) \cdot D\mathbf{V}(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U}) = \\ &= D(\rho(\lambda_k - u))(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U}). \end{aligned}$$

Now, we write

$$D(\rho(\lambda_k - u)) = \rho D\lambda_k + \lambda_k D\rho - D(\rho u).$$

If we assume  $U_1 = \rho$ ,  $U_2 = \rho u$ , we have

$$\lambda_k D\rho - D(\rho u) = (\lambda_k, -1, 0, \dots, 0)$$

which coincides with the first row of the matrix  $\lambda_k \mathbf{I} - \mathbf{A}$ . Hence, we get

$$(\lambda_k D\rho - D(\rho u)) \cdot \mathbf{r}_k(\mathbf{U}) = \mathbf{0}$$

and

$$D\mu_k(\mathbf{V}) \cdot \mathbf{s}_k(\mathbf{V}) = \rho D\lambda_k(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U}),$$

which gives (2.23).  $\square$

It follows from Theorem 2.1 that the system (2.16) is hyperbolic (resp. strictly hyperbolic if and only if the system (2.17) is hyperbolic (resp. strictly hyperbolic)). Moreover, the nature of a characteristic field does not depend on the Eulerian or Lagrangian form of the system of conservation laws: the  $k$ -characteristic fields are simultaneously either genuinely nonlinear or linearly degenerate.

Let us now compare the entropies of the two systems (2.16) and (2.17).

*Theorem 2.2*

*A pair  $(\eta, q)$  is an entropy pair for the system of conservation laws (2.17) written in Lagrangian coordinates if and only if the pair  $(\rho\eta, \rho\eta + q)$  is an entropy pair for the system (2.16) written in Eulerian coordinates.*

*Proof.* Let  $(\eta, q) = (\eta(\mathbf{V}), q(\mathbf{V}))$  be an entropy pair for the system (2.17), i.e., the functions  $\eta$  and  $q$  satisfy

$$Dq(\mathbf{V}) = D\eta(\mathbf{V}) \cdot \mathbf{B}(\mathbf{V}) \quad (2.24)$$

and moreover  $\eta$  is a convex function of  $\mathbf{V}$ . Recall that (2.24) means that, for any smooth solution  $\mathbf{V}$  of (2.17), we have

$$\frac{\partial}{\partial t}\eta(\mathbf{V}) + \frac{\partial}{\partial m}q(\mathbf{V}) = 0.$$

Using (2.21) and (2.24), we obtain

$$D\eta(\mathbf{V}) \cdot D\mathbf{V}(\mathbf{U}) \cdot \mathbf{A}(\mathbf{U}) = uD\eta(\mathbf{V}) \cdot D\mathbf{V}(\mathbf{U}) + \frac{1}{\rho}Dq(\mathbf{V}) \cdot D\mathbf{V}(\mathbf{U})$$

where  $\mathbf{V} = \mathbf{V}(\mathbf{U})$ . Now, given any smooth solution  $\mathbf{U}$  of (2.16), we can write

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial x} = \mathbf{0}$$

and therefore

$$D\eta(\mathbf{V}) \cdot D\mathbf{V}(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial t} + uD\eta(\mathbf{V}) \cdot D\mathbf{V}(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial x} + \frac{1}{\rho}Dq(\mathbf{V}) \cdot D\mathbf{V}(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial x} = 0$$

or equivalently

$$\frac{\partial}{\partial t}\eta(\mathbf{V}(\mathbf{U})) + u\frac{\partial}{\partial x}\eta(\mathbf{V}(\mathbf{U})) + \frac{1}{\rho}\frac{\partial}{\partial x}q(\mathbf{V}(\mathbf{U})) = 0.$$

In short, we obtain

$$\frac{\partial\eta}{\partial t} + u\frac{\partial\eta}{\partial x} + \frac{1}{\rho}\frac{\partial q}{\partial x} = 0.$$

Hence, by the mass conservation equation,  $\mathbf{U}$  satisfies the additional conservation law

$$\frac{\partial}{\partial t}(\rho\eta) + \frac{\partial}{\partial x}(\rho u\eta + q) = 0.$$

On the other hand, using Lemma 1.3 of Chap. III below, one can check that the function  $\mathbf{V} \rightarrow \eta(\mathbf{V}) = \eta(\tau, \Phi)$  is (strictly) convex if and only if the function  $\mathbf{U} \rightarrow \rho\eta(\mathbf{V}(\mathbf{U})) = \rho\eta\left(\frac{1}{\rho}, \frac{\rho\Phi}{\rho}\right)$  is (strictly) convex. This proves that  $(\rho\eta, \rho u\eta + q)$  is indeed an entropy pair for the system (2.16). The converse property is proved in exactly the same way.  $\square$

The change of frame, and the easy formulas deduced in the above computations from the correspondence between the two frames, is very much used in the derivation of numerical schemes (for instance, see, [236]). We will use this correspondence again later on. An important fact is that the equivalence exists not only between strong solutions but also between weak solutions of the associated Cauchy problems as proved by Wagner in [1175] (see also Peng [939]).

## 2.2 The Gas Dynamics Equations

*Example 2.3. The gas dynamics equations in Lagrangian coordinates.* Let us consider again the gas dynamics equations in slab symmetry written in Lagrangian coordinates that were introduced in the Chap. I, Example 2.3:

$$\begin{cases} \frac{\partial\tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial}{\partial m}(pu) = 0. \end{cases} \quad (2.25)$$

Again  $\tau = \frac{1}{\rho}$  is the specific volume,  $u$  the velocity,  $p$  the pressure,  $e$  the specific internal energy, and  $e = \varepsilon + \frac{u^2}{2}$  the specific total energy; the independent variable  $m$  stands for the mass variable. We supplement (2.25) with an equation of state, which can be chosen in the (incomplete) form

$$p = p(\tau, \varepsilon). \quad (2.26)$$

This equation of state must satisfy various conditions of a thermodynamic nature which will be explicitly given later on. Here we just recall some results from thermodynamics that will be made precise in Chap. III. In order to simplify the computations, it is convenient to make a change of dependent variables. We introduce the specific entropy  $s$  defined via the second law of thermodynamics

$$Tds = d\varepsilon + p d\tau, \quad (2.27)$$

where  $T$  denote the temperature. We have

$$T \frac{\partial s}{\partial t} = \frac{\partial \varepsilon}{\partial t} + p \frac{\partial \tau}{\partial t}.$$

If we assume that the dependent variables are sufficiently smooth functions of  $x$  and  $t$ , it follows from (2.16) that

$$T \frac{\partial s}{\partial t} = \frac{\partial e}{\partial t} - u \frac{\partial u}{\partial t} + p \frac{\partial \tau}{\partial t} = 0.$$

Since  $T$  is  $> 0$ , we obtain

$$\frac{\partial s}{\partial t} = 0,$$

which expresses the conversation of entropy for smooth flows. In fact, we shall prove in Chap. III that  $s$  is a strictly convex function of the triple  $(\tau, u, e)$  so that  $(-s, 0)$  is an entropy pair for the system (2.25). Now, we observe that the mapping

$$\boldsymbol{\theta} : \begin{pmatrix} \tau \\ u \\ s \end{pmatrix} \mapsto \begin{pmatrix} \tau \\ u \\ e \end{pmatrix}$$

is smooth and one-to-one. Indeed, from

$$e = \varepsilon(\tau, s) + \frac{u^2}{2}$$

we get

$$D\boldsymbol{\theta} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -p & u & T \end{pmatrix},$$

and  $T$  is  $> 0$ . Hence, in the case of smooth flows, we may equivalently write the system (2.16) in another conservative form (which is not equivalent for discontinuous solutions)

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial s}{\partial t} = 0. \end{cases} \quad (2.28)$$

On the other hand, we know from thermodynamics that given any two of the thermodynamic variables  $\tau, p, \varepsilon, T, s$ , we can determine the remaining three variables. Thus, we may assume that  $p, \varepsilon$ , and  $T$  are given functions of  $\tau$  and  $s$  so that the equation of state becomes

$$p = p(\tau, s) \quad (2.29)$$

(with some abuse of notations since we have noted in the same way the two different mappings  $p(\tau, \varepsilon)$  and  $p(\tau, s)$ ) with the following properties:

$$\frac{\partial p}{\partial \tau}(\tau, s) = -\frac{c^2}{\tau^2} < 0, \quad \frac{\partial^2 p}{\partial \tau^2}(\tau, s) > 0. \quad (2.30)$$

In (2.30)  $c$  is the local sound speed,  $c/\tau$  the Lagrangian sound speed. Recall that for a polytropic ideal gas, we have (see Chap. I, in Example 2.1,  $p = (\gamma - 1) \exp(\frac{(s-s_0)}{C_v}) \rho^\gamma$ )

$$p = C \exp\left(\frac{s}{C_v}\right) \tau^{-\gamma}, \quad (2.31)$$

where  $C_v$  is the specific heat at constant volume. The system (2.28), (2.29) now has a form suitable for the study of the characteristic fields. Indeed, the Jacobian matrix of the system (2.28) is given by

$$\begin{pmatrix} 0 & -1 & 0 \\ p_\tau & 0 & p_s \\ 0 & 0 & 0 \end{pmatrix},$$

with  $p = p(\tau, s)$  and where

$$p_\tau = \frac{\partial p(\tau, s)}{\partial \tau}, \quad p_s = \frac{\partial p(\tau, s)}{\partial s}.$$

Since the corresponding characteristic polynomial is

$$\lambda \left( \lambda^2 + \frac{\partial p}{\partial \tau} \right) = 0,$$

we obtain the distinct real eigenvalues

$$\lambda_1 = -\sqrt{-\frac{\partial p}{\partial \tau}} = -\frac{c}{\tau} < \lambda_2 = 0 < \lambda_3 = \sqrt{-\frac{\partial p}{\partial \tau}} = \frac{c}{\tau}. \quad (2.32)$$

The associated eigenvectors can be taken as

$$\mathbf{r}_1 = \begin{pmatrix} \tau \\ c \\ 0 \end{pmatrix}, \quad \mathbf{r}_2 = \begin{pmatrix} p_s \\ 0 \\ -p_\tau \end{pmatrix}, \quad \mathbf{r}_3 = \begin{pmatrix} \tau \\ -c \\ 0 \end{pmatrix}, \quad (2.33)$$

Then, we have

$$\nabla \lambda_1 = \begin{pmatrix} (\frac{\tau}{2c})p_{\tau\tau} \\ 0 \\ (\frac{\tau}{2c})p_{\tau s} \end{pmatrix}, \quad \nabla \lambda_2 = \mathbf{0}, \quad \nabla \lambda_3 = \begin{pmatrix} (-\frac{\tau}{2c})p_{\tau\tau} \\ 0 \\ (-\frac{\tau}{2c})p_{\tau s} \end{pmatrix}$$

so that

$$\begin{aligned} \nabla \lambda_1^T \cdot \mathbf{r}_1 &= \left( \frac{\tau^2}{2c} \right) \frac{\partial^2 p}{\partial \tau^2}, \\ \nabla \lambda_2^T \cdot \mathbf{r}_2 &= 0, \\ \nabla \lambda_3^T \cdot \mathbf{r}_3 &= -\left( \frac{\tau^2}{2c} \right) \frac{\partial^2 p}{\partial \tau^2}. \end{aligned}$$

Using (2.30), we find that the first and the third characteristic fields are genuinely nonlinear, while the second characteristic field is linearly degenerate.

The above results can be easily extended to the system

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial v}{\partial t} = 0, \\ \frac{\partial w}{\partial t} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial(pu)}{\partial m} = 0 \end{cases} \quad (2.34)$$

where we have taken into account the transverse velocity  $(v, w)^T$  and  $e = \varepsilon + \frac{1}{2}(u^2 + v^2 + w^2)$ . Again  $(-s, 0)$  is an entropy pair for the system (2.34). The eigenvalues of its Jacobian matrix are

$$\lambda_1 = -\frac{c}{\tau} < \lambda_2 = \lambda_3 = \lambda_4 = 0 < \lambda_5 = \frac{c}{\tau}.$$

The eigenspace associated with the triple eigenvalue  $\lambda = 0$  has dimension 3 so that the system is indeed hyperbolic. Moreover, the first and fifth characteristic fields are genuinely nonlinear, while the other characteristic fields are linearly degenerate.  $\square$

*Example 2.4. The gas dynamics equations in Eulerian coordinates.* We turn to the gas dynamics equations written in Eulerian coordinates that were introduced in the Chap. I, Example 2.1:

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}(\rho e + p)u = 0, \end{cases} \quad (2.35)$$

with  $e = \varepsilon + \frac{u^2}{2}$  (total specific energy), and the equation of state

$$p = p(\rho, \varepsilon). \quad (2.36)$$

Note that we have again noted in the same way the pressure though the mappings  $p(\tau, \varepsilon)$  and  $p(\rho, \varepsilon)$  differ. It may be necessary to use different notations, for instance, in some context, we may note  $p = \tilde{p}(\tau, .) = p(\rho, .)$ , with  $\tau = \frac{1}{\rho}$ .

Using Theorems 2.1 and 2.2 with

$$\Phi = \begin{pmatrix} u \\ e \end{pmatrix}, \quad \mathbf{g}(\rho, \Phi) = \begin{pmatrix} p \\ pu \end{pmatrix},$$

we shall exhibit in Chap. IV the Jacobian matrix of the system (2.35) together with its eigenvectors when we shall study Roe's method. At the present time, it is an easy matter to derive the properties of the system (2.35) from those of the system (2.25). First, it follows from Theorem 2.1 that the eigenvalues of the Jacobian matrix of (2.35) are

$$\lambda_1 = u - c < \lambda_2 = u < \lambda_3 = u + c, \quad (2.37)$$

where  $c^2 = -\tau^2 \frac{\partial \tilde{p}(\tau, s)}{\partial \tau} = \frac{\partial p(\rho, s)}{\partial \rho}$ , and the characteristic fields associated with  $\lambda_1$  and  $\lambda_3$  are genuinely nonlinear, while the characteristic field associated with  $\lambda_2$  is linearly degenerate. On the other hand, it follows from Theorem 2.2 that  $(-\rho s, -\rho su)$  is an entropy pair for the system (2.35), i.e.,  $-\rho s$  is a strictly convex function of the triple  $(\rho, \rho u, \rho e)$  and we have for

smooth flows

$$\frac{\partial}{\partial t}(\rho s) + \frac{\partial}{\partial x}(\rho su) = 0.$$

Though the results have already been obtained through the general equivalence theorem between the Lagrangian and Eulerian frames, let us detail some direct computations for the reader's convenience since the Euler system in Eulerian coordinates is very frequently used. We notice that in the case of smooth flows, the system (2.35) can be equivalently written in the nonconservative form

$$\begin{cases} \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + \rho \frac{\partial u}{\partial x} = 0, \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{1}{\rho} \frac{\partial p}{\partial x} = 0, \\ \frac{\partial \varepsilon}{\partial t} + u \frac{\partial \varepsilon}{\partial x} + \frac{p}{\rho} \frac{\partial u}{\partial x} = 0. \end{cases} \quad (2.38)$$

If we again introduce the specific entropy  $s$ , which by (2.27) satisfies

$$Tds = d\varepsilon - \frac{p}{\rho^2} d\rho,$$

we obtain

$$T \left( \frac{\partial s}{\partial t} + u \frac{\partial s}{\partial x} \right) = \frac{\partial \varepsilon}{\partial t} + u \frac{\partial \varepsilon}{\partial x} - \frac{p}{\rho^2} \left( \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} \right) = 0,$$

and therefore

$$\frac{\partial s}{\partial t} + u \frac{\partial s}{\partial x} = 0.$$

Next, we observe that the mapping

$$\boldsymbol{\theta} : \begin{pmatrix} \rho \\ u \\ s \end{pmatrix} \mapsto \begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix}$$

is smooth and one-to-one with Jacobian

$$D\boldsymbol{\theta} = \begin{pmatrix} 1 & 0 & 0 \\ u & \rho & 0 \\ e + \frac{p}{\rho} & \rho u & \rho T \end{pmatrix},$$

so that for smooth flows an equivalent form of (2.27) is given by

$$\begin{cases} \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + \rho \frac{\partial u}{\partial x} = 0, \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{1}{\rho} \frac{\partial p}{\partial x} = 0, \\ \frac{\partial s}{\partial t} + u \frac{\partial s}{\partial x} = 0. \end{cases} \quad (2.39)$$

Choosing  $\rho$  and  $s$  as the two independent thermodynamic variables, we supplement the system (2.39) with the equation of state

$$p = p(\rho, s). \quad (2.40)$$

Note that the properties (2.30) become here

$$\begin{cases} \frac{\partial p}{\partial \rho}(\rho, s) = c^2 > 0, \\ \frac{\partial^2 p}{\partial \rho^2} + \frac{2}{\rho} \frac{\partial p}{\partial \rho} = \frac{2c}{\rho} \left( \rho \frac{\partial c}{\partial \rho} + c \right) > 0. \end{cases} \quad (2.41)$$

System (2.39) is in the form (2.12) with  $\mathbf{v} = (\rho, u, s)^T$  and a matrix  $\mathbf{B}$  which has the simple form

$$\mathbf{B} = \begin{pmatrix} u & \rho & 0 \\ \frac{1}{\rho} \frac{\partial p}{\partial \rho} & u & \frac{1}{\rho} \frac{\partial p}{\partial s} \\ 0 & 0 & u \end{pmatrix}.$$

In the preceding and following formulas, we omit the dependence on  $\mathbf{v}$  to lighten the notations. Since the characteristic equation is given by

$$(u - \lambda) \left( (u - \lambda)^2 - \frac{\partial p}{\partial \rho} \right) = 0,$$

we obtain again the real distinct eigenvalues

$$\lambda_1 = u - c < \lambda_2 = u < \lambda_3 = u + c. \quad (2.42)$$

The associated eigenvectors of  $\mathbf{B}$  can be chosen as

$$\mathbf{r}_1 = \begin{pmatrix} \rho \\ -c \\ 0 \end{pmatrix}, \quad \mathbf{r}_2 = \begin{pmatrix} \frac{\partial p}{\partial s} \\ 0 \\ -c^2 \end{pmatrix}, \quad \mathbf{r}_3 = \begin{pmatrix} \rho \\ c \\ 0 \end{pmatrix}. \quad (2.43)$$

Thus, we have

$$\nabla \lambda_1 = \begin{pmatrix} -\frac{\partial c}{\partial \rho} \\ 1 \\ -\frac{\partial c}{\partial s} \end{pmatrix}, \quad \nabla \lambda_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \nabla \lambda_3 = \begin{pmatrix} \frac{\partial c}{\partial \rho} \\ 1 \\ \frac{\partial c}{\partial s} \end{pmatrix}$$

and

$$\nabla \lambda_1^T \mathbf{r}_1 = - \left( c + \rho \frac{\partial c}{\partial \rho} \right), \quad \nabla \lambda_2^T \mathbf{r}_2 = 0, \quad \nabla \lambda_3^T \mathbf{r}_3 = c + \rho \frac{\partial c}{\partial \rho}.$$

Using (2.41), we obtain again that the first and the third characteristic fields are genuinely nonlinear, while the second characteristic field is linearly degenerate.

As expected, we get the same results in Lagrangian and Eulerian coordinates.  $\square$

*Remark 2.1.* For the sake of completeness, let us also give the expression of the matrices that “diagonalize” the above matrix  $\mathbf{B}$ . Considering the matrix  $\mathbf{T}$  with columns  $(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)$ , then  $\mathbf{T}^{-1}$  has for rows the “left eigenvectors”  $\mathbf{l}_k(\mathbf{u})^T$  (see (2.3)), which gives, setting  $p_s = \frac{\partial p(\rho, s)}{\partial s}$ ,

$$\mathbf{T}^{-1} = \begin{pmatrix} \frac{1}{2\rho} & -\frac{1}{2c} & \frac{p_s}{2pc^2} \\ 0 & 0 & -\frac{1}{c^2} \\ \frac{1}{2\rho} & \frac{1}{2c} & \frac{p_s}{2pc^2} \end{pmatrix},$$

and  $\mathbf{T}^{-1}\mathbf{B}\mathbf{T} = \text{diag}(\lambda_i)$ . These expressions are used in many numerical schemes and also when one linearizes the system (in the variables  $(\rho, u, s)$ ) about a constant state  $\bar{\mathbf{V}}$ . Then, setting

$$\mathbf{V} = (\rho, u, s)^T, \mathbf{W} = \mathbf{T}^{-1}(\bar{\mathbf{V}})\mathbf{V},$$

we get a system of  $p$  decoupled equations in the components  $\alpha_k$  of  $\mathbf{W}$ , which are called “characteristic variables.” Note that we can equivalently write

$$\mathbf{V} = \sum \alpha_k \mathbf{r}_k(\bar{\mathbf{V}}),$$

and if we introduce a linearized pressure by

$$p = \frac{\partial p}{\partial \rho}(\bar{\mathbf{V}})\rho + \frac{\partial p}{\partial s}(\bar{\mathbf{V}})s = \bar{c}^2 \rho + \frac{\partial p}{\partial s}(\bar{\mathbf{V}})s,$$

the characteristic variables  $\alpha_k$  take the simple form

$$\begin{aligned} \alpha_1 &= \frac{(p - \bar{p} \bar{c} u)}{2\bar{c}^2 \bar{\rho}}, \\ \alpha_2 &= -\frac{s}{\bar{c}^2}, \\ \alpha_3 &= \frac{(p + \bar{p} \bar{c} u)}{2\bar{c}^2 \bar{\rho}}, \end{aligned}$$

which will be used in numerical schemes.  $\square$

*Remark 2.2.* We can also write the system in a still different nonconservative form using the “primitive” variables  $\mathbf{V}' = (\rho, u, p)^T$ . The equation for  $p$  is obtained by combining the mass and entropy conservation equations together with (2.41) and a classical chain rule, writing  $\partial_t p = \partial_\rho p \partial_t \rho + \partial_s p \partial_t s$  with a similar formula for  $\partial_x p$ . Note that the mapping  $\varphi : (\rho, u, p)^T \mapsto (\rho, \rho u, \rho e)^T$  is one-to-one, with, for a  $\gamma$ -law  $p = (\gamma - 1)\rho\varepsilon$ ,

$$D\varphi = \begin{pmatrix} 1 & 0 & 0 \\ u & \rho & 0 \\ \frac{u^2}{2} & \rho u & \frac{1}{(\gamma-1)} \end{pmatrix},$$

and

$$(D\varphi)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{-u}{\rho} & \frac{1}{\rho} & 0 \\ \frac{(\gamma-1)u^2}{2} & -(\gamma-1)u & \gamma-1 \end{pmatrix}.$$

Then the matrix of the corresponding nonconservative system (see (2.12)) defined by  $\mathbf{B}' = (D\varphi)^{-1} \mathbf{A} D\varphi$  (here and in the following lines of this subsection, the notation “prime” does not mean derivation) is

$$\mathbf{B}' = \begin{pmatrix} u & \rho & 0 \\ 0 & u & \frac{1}{\rho} \\ 0 & \rho c^2 & u \end{pmatrix};$$

the eigenvectors associated with  $\lambda_1 = u - c$ ,  $\lambda_2 = u$ ,  $\lambda_3 = u + c$  can be chosen as

$$\mathbf{r}'_1 = \begin{pmatrix} 1 \\ -c/\rho \\ c^2 \end{pmatrix}, \quad \mathbf{r}'_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{r}'_3 = \begin{pmatrix} 1 \\ c/\rho \\ c^2 \end{pmatrix}.$$

If  $\mathbf{T}'$  is the matrix with columns  $(\mathbf{r}'_1, \mathbf{r}'_2, \mathbf{r}'_3)$ , then  $\mathbf{T}'^{-1} \mathbf{B}' \mathbf{T}' = \text{diag}(\lambda_i)$  and  $\mathbf{T}'^{-1}$  is given by

$$\mathbf{T}'^{-1} = \begin{pmatrix} 0 & -\frac{\rho}{2c} & \frac{1}{2c^2} \\ 1 & 0 & -\frac{1}{c^2} \\ 0 & \frac{\rho}{2c} & \frac{1}{2c^2} \end{pmatrix}.$$

In fact, the corresponding nonconservative system is not equivalent to (2.35) for discontinuous solutions, and neither is (2.38).  $\square$

### 2.3 Ideal MHD

*Example 2.5.* The ideal MHD equations in Lagrangian coordinates. We pass to the ideal magnetohydrodynamics equations already considered in the Chap. I, Example 2.4:

$$\left\{ \begin{array}{l} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p^*}{\partial m} = 0, \\ \frac{\partial v}{\partial t} - \frac{\partial}{\partial m} \left( \frac{B_x}{\mu} B_y \right) = 0, \\ \frac{\partial w}{\partial t} - \frac{\partial}{\partial m} \left( \frac{B_x}{\mu} B_z \right) = 0, \\ \frac{\partial}{\partial t} (\tau B_y) - \frac{\partial}{\partial m} (B_x v) = 0, \\ \frac{\partial}{\partial t} (\tau B_z) - \frac{\partial}{\partial m} (B_x w) = 0, \\ \frac{\partial e^*}{\partial t} + \frac{\partial}{\partial m} \left( p^* u - \frac{B_x}{\mu} (B_x u + B_y v + B_z w) \right) = 0. \end{array} \right. \quad (2.44)$$

In (2.44),  $\mathbf{B} = (B_x, B_y, B_z)^T$  is the magnetic field,  $B_x$  is a constant,  $\mu$  is the magnetic permeability (assumed to be a constant), and

$$e^* = \varepsilon + \frac{1}{2}(u^2 + v^2 + w^2) + \frac{|\mathbf{B}|^2}{2\mu}, \quad (2.45)$$

$$p^* = p(\tau, \varepsilon) + \frac{|\mathbf{B}|^2}{2\mu} \quad (2.46)$$

where  $p(\tau, \varepsilon)$  is an (incomplete) equation of state of the magnetized fluid.

Again, we introduce the specific entropy  $s$  as in (2.27). Let us then check that any sufficiently smooth solution of (2.44) satisfies

$$\frac{\partial s}{\partial t} = 0.$$

Indeed, using (2.27), (2.45), and (2.46), we have

$$\left\{ \begin{array}{l} T \frac{\partial s}{\partial t} = \frac{\partial e^*}{\partial t} - (u \frac{\partial u}{\partial t} + v \frac{\partial v}{\partial t} + w \frac{\partial w}{\partial t}) - \\ - \frac{\tau}{\mu} (B_y \frac{\partial B_y}{\partial t} + B_z \frac{\partial B_z}{\partial t}) + (p - \frac{|\mathbf{B}|^2}{2\mu}) \frac{\partial \tau}{\partial t}. \end{array} \right.$$

Now the first, fifth, and sixth Eqs. (2.44) yield

$$\tau(B_y \frac{\partial B_y}{\partial t} + B_z \frac{\partial B_z}{\partial t}) = B_x(B_x \frac{\partial u}{\partial m} + B_y \frac{\partial v}{\partial m} + B_z \frac{\partial w}{\partial m}) - |\mathbf{B}|^2 \frac{\partial u}{\partial m}$$

while the second, third, and fourth Eqs. (2.44) give

$$u \frac{\partial u}{\partial t} + v \frac{\partial v}{\partial t} + w \frac{\partial w}{\partial t} = -u \frac{\partial p^*}{\partial m} + \frac{1}{\mu} B_x(v \frac{\partial B_y}{\partial m} + w \frac{\partial B_z}{\partial m}).$$

Hence, we obtain

$$\begin{cases} u \frac{\partial u}{\partial t} + v \frac{\partial v}{\partial t} + w \frac{\partial w}{\partial t} + \frac{\tau}{\mu}(B_y \frac{\partial B_y}{\partial t} + B_z \frac{\partial B_z}{\partial t}) = \\ = \frac{1}{\mu} B_x \frac{\partial}{\partial m}(B_x u + B_y v + B_z w) - u \frac{\partial p^*}{\partial m} - \frac{|\mathbf{B}|^2}{\mu} \frac{\partial u}{\partial m} \end{cases}$$

and therefore by the first and last Eqs. (2.44)

$$\begin{aligned} T \frac{\partial s}{\partial t} &= -\frac{\partial}{\partial m}(p^* u) + u \frac{\partial p^*}{\partial m} + \frac{|\mathbf{B}|^2}{\mu} \frac{\partial u}{\partial m} + (p - \frac{|\mathbf{B}|^2}{2\mu}) \frac{\partial u}{\partial m} = \\ &= (p + \frac{|\mathbf{B}|^2}{2\mu} - p^*) \frac{\partial u}{\partial m} = 0 \end{aligned}$$

which proves our assertion.

Indeed, we shall prove in Chap. III that  $-s$  is a strictly convex function of  $(\tau, u, v, w, \tau, B_y, \tau B_z, e^*)$  so that  $(-s, 0)$  is an entropy pair for (2.44). Next, as in Example 2.3, we observe that the mapping

$$\boldsymbol{\theta} : (\tau, u, v, w, B_y, B_z, s)^T \rightarrow (\tau, u, v, w, \tau B_y, \tau B_z, e^*)^T$$

is smooth for  $\tau > 0$  and one-to-one. Therefore, for smooth solutions, we may equivalently write the system (2.44) in the nonconservative form

$$\left\{ \begin{array}{l} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p^*}{\partial m} = 0, \\ \frac{\partial v}{\partial t} - \frac{B_x}{\mu} \frac{\partial}{\partial m} B_y = 0, \\ \frac{\partial w}{\partial t} - \frac{B_x}{\mu} \frac{\partial}{\partial m} B_z = 0, \\ \frac{\partial}{\partial t} B_y + \frac{B_y}{\tau} \frac{\partial u}{\partial m} - \frac{B_x}{\tau} \frac{\partial v}{\partial m} = 0, \\ \frac{\partial}{\partial t} B_z + \frac{B_z}{\tau} \frac{\partial u}{\partial m} - \frac{B_x}{\tau} \frac{\partial w}{\partial m} = 0, \\ \frac{\partial s}{\partial t} = 0. \end{array} \right. \quad (2.47)$$

with

$$p^* = p(\tau, s) + \frac{|\mathbf{B}|^2}{2\mu}. \quad (2.48)$$

The corresponding matrix (see (2.12)) is then given by

$$\begin{pmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ \frac{\partial p}{\partial \tau} & 0 & 0 & 0 & & \frac{B_z}{\mu} & \frac{\partial p}{\partial s} \\ 0 & 0 & 0 & 0 & -\frac{B_x}{\mu} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{B_x}{\mu} & 0 \\ 0 & \frac{B_y}{\mu} & -\frac{B_x}{\tau} & 0 & 0 & 0 & 0 \\ 0 & \frac{B_z}{\tau} & 0 & -\frac{B_x}{\tau} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Through a simple but lengthy calculation, one can check that its characteristic polynomial reads

$$\lambda(\lambda^2 - \frac{B_x^2}{\mu\tau})(\lambda^4 + \lambda^2(\frac{\partial p}{\partial \tau} - \frac{|\mathbf{B}|^2}{\mu\tau}) - \frac{\partial p}{\partial \tau}\frac{B_x^2}{\mu\tau}).$$

Setting

$$b_x = \sqrt{\frac{B_x^2\tau}{\mu}}, \quad b = \sqrt{\frac{|\mathbf{B}|^2\tau}{\mu}}, \quad (c^*)^2 = c^2 + b^2, \quad c^2 = -\tau^2 \frac{\partial p}{\partial \tau} \quad (2.49)$$

and

$$\begin{cases} c_a = b_x, \\ c_s = \left\{ \frac{1}{2}((c^*)^2 - \sqrt{(c^*)^4 - 4c^2b_x^2}) \right\}^{1/2}, \\ c_f = \left\{ \frac{1}{2}((c^*)^2 + \sqrt{(c^*)^4 - 4c^2b_x^2}) \right\}^{1/2}, \end{cases} \quad (2.50)$$

the eigenvalues of the Jacobian matrix are given by

$$-\frac{c_f}{\tau} \leq -\frac{c_a}{\tau} \leq -\frac{c_s}{\tau} \leq 0 \leq \frac{c_s}{\tau} \leq -\frac{c_a}{\tau} \leq -\frac{c_f}{\tau} \quad (2.51)$$

where  $c_s$ ,  $c_a$ , and  $c_f$  are the slow, the Alfvén, and the fast characteristic speeds, respectively.

We notice that these eigenvalues are simple for  $\mathbf{B}_\perp \neq \mathbf{0}$  where  $\mathbf{B}_\perp = (0, B_y, B_z)^T$  is the transverse magnetic field. Indeed, the eigenvalues are simple unless

$$c_a = c_s \text{ or } c_f.$$

This latter case occurs if and only if

$$(c^*)^2 \pm \sqrt{(c^*)^4 - 4c^2 b_x^2} = 2b_x^2$$

i.e., if and only if

$$(c^*)^2 - c^2 b_x^2 = 2b_x^2 \iff b^2 = b_x^2 \iff \mathbf{B}_\perp = \mathbf{0}.$$

Hence, the ideal MHD system is strictly hyperbolic for  $\mathbf{B}_\perp \neq \mathbf{0}$ . Then, denoting by  $\mathbf{r}_0, \mathbf{r}_{\pm c_a}$ , and  $\mathbf{r}_{\pm c_\alpha}$  the eigenvectors associated with the eigenvalues  $\lambda = 0$ ,  $\lambda = \pm \frac{c_a}{\tau}$ , and  $\lambda = \pm \frac{c_\alpha}{\tau}$ ,  $\alpha = s, f$ , respectively, one can check that these eigenvectors may be chosen as

$$\mathbf{r}_0 = \begin{pmatrix} \frac{\partial p}{\partial s} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -\frac{\partial p}{\partial \tau} \end{pmatrix}, \quad \mathbf{r}_{\pm c_a} = \begin{pmatrix} 0 \\ 0 \\ \mp \operatorname{sgn}(B_x) B_z \\ \pm \operatorname{sgn}(B_x) B_y \\ \sqrt{\frac{\mu}{\tau}} B_z \\ -\sqrt{\frac{\mu}{\tau}} B_y \\ 0 \end{pmatrix}, \quad \mathbf{r}_{\pm c_\alpha} = \begin{pmatrix} \tau \\ \mp c_\alpha \\ \mp \frac{\tau c_\alpha B_x B_y}{\mu(b_x^2 - c_\alpha^2)} \\ \mp \frac{\tau c_\alpha B_x B_z}{\mu(b_x^2 - c_\alpha^2)} \\ \frac{c_\alpha^2 B_y}{b_x^2 - c_\alpha^2} \\ \frac{c_\alpha^2 B_z}{b_x^2 - c_\alpha^2} \\ 0 \end{pmatrix}. \quad (2.52)$$

On the other hand, when  $\mathbf{B}_\perp = \mathbf{0}$ , we have

$$c_s^2 = c^2 + b_x^2 - |c^2 - b_x^2|, \quad c_f^2 = c^2 + b_x^2 + |c^2 - b_x^2|.$$

Hence, one of the three following cases occurs:

- (i)  $c < b_x \Rightarrow c_f = c_a$ : two double eigenvalues;
- (ii)  $c = b_x \Rightarrow c_f = c_s = c_a$ : two triple eigenvalues;
- (iii)  $c > b_x \Rightarrow c_s = c_a$ : two double eigenvalues.

In the case of a double (respectively triple) eigenvalue, we obtain that the corresponding eigenspace is spanned by the two first (resp. three) vectors

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ -\frac{\lambda B_x}{\mu} \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ -\frac{\lambda B_x}{\mu} \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -\lambda \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

In each case, the dimension of the eigenspace is equal to the multiplicity of the eigenvalue. This proves that the ideal MHD system is hyperbolic in the whole state space.

Concerning the nature of the characteristic fields, we observe that those associated with the eigenvalues  $\lambda = 0$  and  $\lambda = \pm \frac{c_a}{\tau}$  are linearly degenerate, and the associated waves are respectively called material wave and Alfvén waves. The linear degeneracy is obvious for  $\lambda = 0$ . For  $\lambda = \pm \frac{c_a}{\tau}$ , it follows immediately from the expression (2.52) of  $\mathbf{r}_{\pm c_a}$  that  $\nabla \lambda \cdot \mathbf{r} = 0$ . On the other hand, the characteristic fields associated with  $\lambda = \pm \frac{c_\alpha}{\tau}$ ,  $\alpha = s, f$ , are neither genuinely nonlinear nor linearly degenerate. Indeed if, for simplicity, we restrict ourselves to the case  $\mathbf{B}_\perp = 0$ , the eigenvalues  $\lambda = \pm \frac{c_\alpha}{\tau}$ ,  $\alpha = s, f$ , are the solutions of

$$\lambda^4 + \lambda^2 \left( \frac{\partial p}{\partial \tau} - \frac{B_x^2}{\mu \tau} \right) - \frac{\partial p}{\partial \tau} \frac{B_x^2}{\mu \tau} = 0.$$

Hence, we obtain on the one hand

$$c_s^2 = \begin{cases} b_x^2, & c^2 > b_x^2 \\ c^2, & c^2 < b_x^2 \end{cases}, \quad c_f^2 = \begin{cases} c^2, & c^2 > b_x^2 \\ b_x^2, & c^2 < b_x^2 \end{cases}$$

and on the other hand

$$(2\lambda^2 + \frac{\partial p}{\partial \tau} - \frac{B_x^2}{\mu \tau}) \nabla \lambda^2 = \lambda^2 \nabla \left( \frac{\partial p}{\partial \tau} - \frac{B_x^2}{\mu \tau} \right) - \nabla \left( \frac{\partial p}{\partial \tau} \frac{B_x^2}{\mu \tau} \right).$$

Assume, for instance,  $\lambda = \pm \frac{c_f}{\tau}$ . We first observe that

$$2\lambda^2 + \frac{\partial p}{\partial \tau} - \frac{B_x^2}{\mu \tau} = \frac{1}{\tau^2} (2c_f^2 - c^2 - b_x^2) = \frac{1}{\tau^2} \begin{cases} c^2 - b_x^2, & c^2 > b_x^2 \\ b_x^2 - c^2, & c^2 < b_x^2 \end{cases}$$

and therefore the above expression keeps a constant sign (+ in this case) as  $c$  varies. Next, using (2.52), we have

$$\left\{ \lambda^2 \nabla \left( \frac{\partial p}{\partial \tau} - \frac{B_x^2}{\mu \tau} \right) - \nabla \left( \frac{\partial p}{\partial \tau} \frac{B_x^2}{\mu \tau} \right) \right\} \cdot \mathbf{r} = \tau \left\{ \frac{\partial^2 p}{\partial \tau^2} (\lambda^2 - \frac{B_x^2}{\mu \tau}) + (\lambda^2 + \frac{\partial p}{\partial \tau}) \frac{B_x^2}{\mu \tau^2} \right\}$$

$$= \frac{1}{\tau} \frac{\partial^2 p}{\partial \tau^2} (c_f^2 - b_x^2) + \frac{1}{\tau^3} (c_f^2 - c^2) b_x^2 = \begin{cases} \frac{1}{\tau} \frac{\partial^2 p}{\partial \tau^2} (c^2 - b_x^2) > 0, & b_x^2 < c^2, \\ \frac{1}{\tau^3} (b_x^2 - c^2) b_x^2 < 0, & b_x^2 > c^2. \end{cases}$$

Hence,  $\nabla \lambda \cdot \mathbf{r}$  changes of sign as  $c$  crosses  $b_x$ . A similar conclusion holds with  $\lambda = \pm \frac{c_s}{\tau}$ . The associated waves are called slow/fast magnetosonic waves. The system is nonstrictly hyperbolic, and the Riemann problem is complicated to solve [989].  $\square$

We can state similar results for the ideal MHD equations in Eulerian coordinates, thanks to the correspondence between the two frames.

Let us go back to the general system of conservation laws (2.1). If the  $k$ th characteristic field is genuinely nonlinear, it is convenient to normalize the right and left eigenvectors  $\mathbf{r}_k(\mathbf{u})$  and  $\mathbf{l}_k(\mathbf{u})^T$  in such a way that

$$\begin{cases} D\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 1, \\ \mathbf{l}_k(\mathbf{u})^T \cdot \mathbf{r}_k(\mathbf{u}) = 1. \end{cases} \quad (2.53)$$

Similarly, if the  $k$ th characteristic field is linearly degenerate, i.e.,

$$D\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0,$$

one normalizes the vectors  $\mathbf{r}_k(u)$  and  $\mathbf{l}_k(u)$  so that

$$\mathbf{l}_k(\mathbf{u})^T \cdot \mathbf{r}_k(\mathbf{u}) = 1. \quad (2.54)$$

### 3 Simple Waves and Riemann Invariants

#### 3.1 Rarefaction Waves

Let  $\mathbf{u}_L$  and  $\mathbf{u}_R$  be two states of  $\Omega \subset \mathbb{R}^p$ ; in this section, we are looking for piecewise smooth continuous functions  $\mathbf{u} : (x, t) \rightarrow \mathbf{u}(x, t)$ , solutions of (2.1) that connect  $\mathbf{u}_L$ , and  $\mathbf{u}_R$ :

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = 0, \\ \mathbf{u}(x, 0) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0. \end{cases} \end{cases} \quad (3.1)$$

At first, we restrict ourselves to *self-similar solutions* of (3.1), i.e., solutions of the form

$$\mathbf{u}(x, t) = \mathbf{v}\left(\frac{x}{t}\right). \quad (3.2)$$

We begin by considering *classical* self-similar solutions of (3.1); these solutions satisfy the equation

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}$$

in the classical sense. We must have

$$-\left(\frac{x}{t^2}\right)\mathbf{v}'\left(\frac{x}{t}\right) + \left(\frac{1}{t}\right)\mathbf{A}\left(\mathbf{v}\left(\frac{x}{t}\right)\right)\mathbf{v}'\left(\frac{x}{t}\right) = \mathbf{0},$$

so that by setting  $\xi = \frac{x}{t}$

$$(\mathbf{A}(\mathbf{v}(\xi)) - \xi \mathbf{I}) \cdot \mathbf{v}'(\xi) = \mathbf{0}.$$

Hence, either we obtain

$$\mathbf{v}'(\xi) = \mathbf{0}$$

or there exists an index  $k \in 1, \dots, p$  such that

$$\mathbf{v}'(\xi) = \alpha(\xi)\mathbf{r}_k(\mathbf{v}(\xi)), \quad \lambda_k(\mathbf{v}(\xi)) = \xi.$$

If  $\mathbf{v}'(\xi)$  is nonzero on an interval, since the eigenvalues are distinct, the index  $k$  does not depend on  $\xi$  in that interval. If we differentiate the second equation with respect to  $\xi$ , we get

$$D\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi) = 1,$$

and using the first equation

$$\alpha(\xi)D\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{r}_k(\mathbf{v}(\xi)) = 1. \quad (3.3)$$

Equation (3.3) cannot be solved if the  $k$ th characteristic field is linearly degenerate. But if the  $k$ th field is genuinely nonlinear, we get, with the normalization (2.53),

$$\alpha(\xi) = 1.$$

Hence, we find either

$$\mathbf{v}'(\xi) = \mathbf{0}$$

or

$$\begin{cases} \mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi)), \\ \lambda_k(\mathbf{v}(\xi)) = \xi, \end{cases} \quad (3.4)$$

and  $\mathbf{v}$  is therefore an integral curve of the field  $\mathbf{r}_k$ . Thus, assume that the  $k$ th characteristic field is genuinely nonlinear and that the function  $\mathbf{v}$  is a solution of (3.3) with

$$\mathbf{v}(\lambda_k(\mathbf{u}_L)) = \mathbf{u}_L, \quad \mathbf{v}(\lambda_k(\mathbf{u}_R)) = \mathbf{u}_R$$

(this presupposes that  $\mathbf{u}_L$  and  $\mathbf{u}_R$  are on the same integral curve of  $\mathbf{r}_k$  and that  $\lambda_k$  increases from  $\mathbf{u}_L$  to  $\mathbf{u}_R$  along this curve). Then, it follows from the above analysis that the function

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_L, & \frac{x}{t} \leq \lambda_k(\mathbf{u}_L), \\ \mathbf{v}\left(\frac{x}{t}\right), & \lambda_k(\mathbf{u}_L) \leq \frac{x}{t} \leq \lambda_k(\mathbf{u}_R), \\ \mathbf{u}_R, & \frac{x}{t} \geq \lambda_k(\mathbf{u}_R) \end{cases} \quad (3.5)$$

is a continuous self-similar weak solution of (2.1).

*Definition 3.1*

Such a self-similar weak solution (3.5) of (2.1) is called a  $k$ -centered simple wave or a  $k$ -rarefaction wave connecting the states  $\mathbf{u}_L$  and  $\mathbf{u}_R$ .

We shall see in Sect. 3.2 that the straight lines that form the rarefaction fan in Fig. 3.1 are the characteristic curves of the  $k$ th field.

Concerning the existence of  $k$ -simple waves connecting two states, we have the following local result.

*Theorem 3.1*

Assume that the  $k$ th characteristic field is genuinely nonlinear with the normalization (2.53). Given a state  $\mathbf{u}_L \in \Omega$ , there exists a curve  $\mathcal{R}_k(\mathbf{u}_L)$  of states of  $\Omega$  that can be connected to  $\mathbf{u}_L$  on the right by a  $k$ -simple wave. Moreover, there exists a parametrization of  $\mathcal{R}_k(\mathbf{u}_L) : \varepsilon \mapsto \Phi_k(\varepsilon)$  defined for  $0 \leq \varepsilon \leq \varepsilon_0$ ,  $\varepsilon_0$  small enough, such that

$$\Phi_k(\varepsilon) = \mathbf{u}_L + \varepsilon \mathbf{r}_k(\mathbf{u}_L) + \frac{\varepsilon^2}{2} D\mathbf{r}_k(\mathbf{u}_L) \cdot \mathbf{r}_k(\mathbf{u}_L) + \mathcal{O}(\varepsilon^3). \quad (3.6)$$

*Proof.* Let  $\mathbf{v} : \xi \mapsto \mathbf{v}(\xi)$  be the solution of the differential system

$$\mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi)), \quad \xi > \lambda_k(\mathbf{u}_L), \quad (3.7)$$

$$\mathbf{v}(\lambda_k(\mathbf{u}_L)) = \mathbf{u}_L. \quad (3.8)$$

The function  $\mathbf{v}$  exists for  $\lambda_k(\mathbf{u}_L) \leq \xi \leq \lambda_k(\mathbf{u}_L) + \varepsilon_0$ ,  $\varepsilon_0 > 0$  small enough. Using (3.7), we have

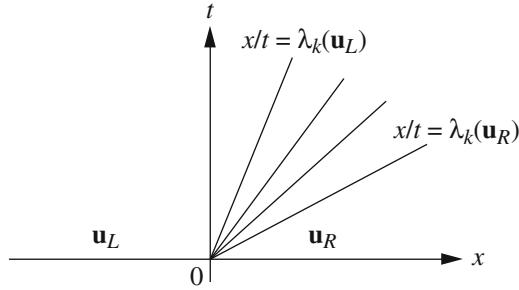
$$\frac{d}{d\xi} \lambda_k(\mathbf{v}(\xi)) = D\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi) = D\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{r}_k(\mathbf{v}(\xi))$$

and by (2.53)

$$\frac{d}{d\xi} \lambda_k(\mathbf{v}(\xi)) = 1.$$

Thus,  $\mathbf{v}$  satisfies

$$\lambda_k(\mathbf{v}(\xi)) - \lambda_k(\mathbf{v}(\lambda_k(\mathbf{u}_L))) = \xi - \lambda_k(\mathbf{u}_L)$$



**Fig. 3.1** Rarefaction fan

so that by (3.8)

$$\lambda_k(\mathbf{v}(\xi)) = \xi.$$

Hence, the function  $\mathbf{v}$  is indeed the solution of (3.4) for which (3.8) holds.

Next, we define

$$\mathcal{R}_k(\mathbf{u}_L) = \{\mathbf{v}(\xi); \lambda_k(\mathbf{u}_L) \leq \xi \leq \lambda_k(\mathbf{u}_L) + \varepsilon_0\}.$$

The curve  $\mathcal{R}_k(\mathbf{u}_L)$  is therefore the set of all states of  $\Omega$  that can be connected to  $\mathbf{u}_L$  on the right by a  $k$ -simple wave. Setting

$$\Phi_k(\varepsilon) = \mathbf{v}(\lambda_k(\mathbf{u}_L) + \varepsilon), \quad 0 \leq \varepsilon \leq \varepsilon_0,$$

we have

$$\Phi_k(0) = \mathbf{u}_L$$

and

$$\Phi'_k(0) = \mathbf{v}'(\lambda_k(\mathbf{u}_L)) = \mathbf{r}_k(\mathbf{v}(\lambda_k(\mathbf{u}_L))) = \mathbf{r}_k(\mathbf{u}_L).$$

Finally, since

$$\mathbf{v}''(\varepsilon) = D\mathbf{r}_k(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi) = D\mathbf{r}_k(\mathbf{v}(\xi)) \cdot \mathbf{r}_k(\mathbf{v}(\xi)),$$

we find

$$\Phi''_k(0) = D\mathbf{r}_k(\mathbf{u}_L) \cdot \mathbf{r}_k(\mathbf{u}_L).$$

This proves the expansion (3.6).  $\square$

The curve  $\mathcal{R}_k(\mathbf{u}_L)$  is called a *k-rarefaction curve*. It is an *integral curve of*  $\mathbf{r}_k$ , which is thus tangent to  $\mathbf{r}_k(\mathbf{u}_L)$  at the point  $\mathbf{u}_L$ .

To go on with our study of elementary waves associated with one specific characteristic family (for which we no longer assume that it is genuinely nonlinear), let us introduce the Riemann invariants.

### 3.2 Riemann Invariants

*Definition 3.2*

A smooth function  $w : \Omega \rightarrow \mathbb{R}$  is called a  $k$ -Riemann invariant if it satisfies

$$Dw(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0, \quad \forall \mathbf{u} \in \Omega. \quad (3.9)$$

A  $k$ -Riemann invariant  $w$  is constant on a curve  $\mathbf{v} : \xi \in \mathbb{R} \mapsto \mathbf{v}(\xi) \in \mathbb{R}^p$  iff

$$\frac{d}{d\xi} w(\mathbf{v}(\xi)) = Dw(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi) = 0,$$

which holds if  $\mathbf{v}$  is an integral curve of  $\mathbf{r}_k$

$$\mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi)). \quad (3.10)$$

This means that a  $k$ -Riemann invariant is constant along the trajectories of the vector field  $\mathbf{r}_k$ . Note that (3.9) is a first-order linear differential equation in  $\mathbb{R}^p$ . It can usually be integrated explicitly, as we shall see.

*Remark 3.1.* When the  $k$ th field is linearly degenerate,  $\lambda_k$  is a  $k$ -Riemann invariant (see (2.6)).  $\square$

Let us show that there exist locally  $(p - 1)$   $k$ -Riemann invariants whose gradients are linearly independent. In general, we can solve (3.9) locally by the method of characteristics as soon as  $w$  is given on some initial surface that is not characteristic for the vector field  $\mathbf{r}_k$ . Recall that a surface  $\mathcal{S} = \{\mathbf{u} \in \mathbb{R}^p, \varphi(\mathbf{u}) = 0\}$  is characteristic if

$$D\varphi(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0 \quad \forall \mathbf{u} \in \mathcal{S},$$

i.e., if  $\mathbf{r}_k(\mathbf{u})$  is tangent to  $\mathcal{S}$  for  $\mathbf{u} \in \mathcal{S}$ . For  $p = 2$ , we thus take a curve that is not tangent to  $\mathbf{r}_k(\mathbf{u})$  at any point, and we assign arbitrary values for  $w$  along it. For simplicity, we prove the result in the case where the hyperplane

$$\mathcal{S}_p = \{\mathbf{u} \in \mathbb{R}^p, \mathbf{u}_p = 0\}$$

is not characteristic, i.e.,

$$\mathbf{r}_k(\mathbf{u}) \notin \mathcal{S}_p \quad \text{for } \mathbf{u} \in \mathcal{S}_p \cap \Omega.$$

*Lemma 3.1*

Assume that the hyperplane  $\mathcal{S}_p = \{\mathbf{u} \in \mathbb{R}^p, \mathbf{u}_p = 0\}$  is not characteristic for  $\mathbf{r}_k$ . Then, there exists a smooth change of variables  $\mathbf{u} = \boldsymbol{\theta}(\mathbf{v})$  defined in a neighborhood of  $\mathcal{S}_p$  such that (3.9) is equivalent to

$$\frac{\partial z}{\partial v_p} = 0, \quad z = w \circ \boldsymbol{\theta}. \quad (3.11)$$

Moreover, the  $(p - 1)$  functions

$$w_j = z_j \circ \boldsymbol{\theta}^{-1}(\mathbf{v}), \quad 1 \leq j \leq p-1,$$

where

$$z_j(\mathbf{v}) = v_j, \quad 1 \leq j \leq p-1,$$

are  $k$ -Riemann invariants whose gradients are linearly independent.

*Proof.* Let  $\mathbf{e}_p = (0, 0, \dots, 1)^T \in \mathbb{R}^p$  be the  $p$ th canonical basis vector. Since by definition

$$\frac{\partial z}{\partial v_p}(\mathbf{v}) = Dz(\mathbf{v}) \cdot \mathbf{e}_p$$

and

$$Dz(\mathbf{v}) \cdot \mathbf{e}_p = D(w \circ \boldsymbol{\theta})(\mathbf{v}) \cdot \mathbf{e}_p = Dw(\boldsymbol{\theta}(\mathbf{v})) \cdot D\boldsymbol{\theta}(\mathbf{v}) \cdot \mathbf{e}_p = Dw(\mathbf{u}) \cdot \frac{\partial \boldsymbol{\theta}}{\partial v_p}(\mathbf{v}),$$

we see that (3.11) is equivalent to (3.9) iff  $\boldsymbol{\theta}$  satisfies

$$\frac{\partial \boldsymbol{\theta}}{\partial v_p}(\mathbf{v}) = \mathbf{r}_k(\boldsymbol{\theta}(\mathbf{v})). \quad (3.12)$$

Given the values of  $\boldsymbol{\theta}$  on the surface  $\mathcal{S}_p$ , for instance,

$$\boldsymbol{\theta}(v_1, \dots, v_{p-1}, 0) = (v_1, \dots, v_{p-1}, 0),$$

for each  $(v_1, \dots, v_{p-1})$ , (3.12) has a smooth solution that is an integral curve of  $\mathbf{r}_k$  defined for  $v_p$  small enough. Since by assumption

$$\frac{\partial \boldsymbol{\theta}}{\partial v_p}(v_1, \dots, v_{p-1}, 0) = \mathbf{r}_k(v_1, \dots, v_{p-1}, 0) \notin \mathcal{S}_p$$

and

$$\frac{\partial \boldsymbol{\theta}}{\partial v_i}(v_1, \dots, v_{p-1}, 0) = \mathbf{e}_i, \quad i = 1, \dots, p-1,$$

this enables us to define a mapping  $\mathbf{v} \mapsto \boldsymbol{\theta}(\mathbf{v})$  which is one-to-one in a neighborhood of  $\mathcal{S}_p$ . Then, the coordinate functions  $z_j(\mathbf{v}) = v_j$ ,  $1 \leq j \leq p-1$  are solutions of (3.11), and their gradients are obviously linearly independent. The result follows easily.  $\square$

We have thus proven that locally there exist  $(p-1)$   $k$ -Riemann invariants whose gradients are linearly independent. Now, given  $w_1, \dots, w_j$ , a set of  $j$   $k$ -Riemann invariants, and  $Z : \mathbb{R}^j \rightarrow \mathbb{R}$ ,  $Z(w_1, \dots, w_j)$  is clearly a  $k$ -Riemann invariant and vice versa. We have the following result.

*Lemma 3.2*

*Given  $(p-1)$   $k$ -Riemann invariants  $w_1, \dots, w_{p-1}$  whose gradients are linearly independent, there exists (locally), for any  $k$ -Riemann invariant  $w$ , a function  $Z = \mathbb{R}^{p-1} \rightarrow \mathbb{R}$  such that*

$$w = Z(w_1, \dots, w_{p-1}).$$

*Proof.* By Lemma 3.1, the proof is obvious since, after a smooth change of variables, the  $(p-1)$   $k$ -Riemann invariants are the coordinates  $z_j(\mathbf{v}) = v_j$ ,  $1 \leq j \leq p-1$ . The general case follows in a similar way. Let us complete the set  $w_1, \dots, w_{p-1}$  by a smooth function  $w_p : \Omega \rightarrow \mathbb{R}$  so that the  $p$  gradients  $\nabla w_1, \dots, \nabla w_{p-1}, \nabla w_p$  are linearly independent. For instance, we can find locally a function  $w_p$  such that  $\nabla w_p(\mathbf{u})$  is collinear to  $\mathbf{r}_k(\mathbf{u})$ , since by (3.9)  $\nabla w_1, \dots, \nabla w_{p-1}$  are all orthogonal to  $\mathbf{r}_k$ . Then, the mapping  $\mathbb{R}^p \rightarrow \mathbb{R}^p$  defined by

$$\mathbf{u} = (u_1, \dots, u_p)^T \mapsto (w_1, \dots, w_p)^T(\mathbf{u}) = \mathbf{v} \quad (3.13)$$

is a diffeomorphism of  $\mathbb{R}^p$ . Denote by  $\boldsymbol{\theta}$  the inverse diffeomorphism so that  $\mathbf{u} = \boldsymbol{\theta}(\mathbf{v})$ , and we can write

$$w(\mathbf{u}) = w \circ \boldsymbol{\theta} \circ \boldsymbol{\theta}^{-1}(\mathbf{u}).$$

If we prove that

$$\frac{\partial(w \circ \boldsymbol{\theta})}{\partial w_p}(\mathbf{v}) = 0, \quad (3.14)$$

the result will follow by setting  $Z = w \circ \boldsymbol{\theta}$  since then

$$w(\mathbf{u}) = w \circ \boldsymbol{\theta}(w_1, \dots, w_{p-1})(\mathbf{u}) = Z(w_1, \dots, w_{p-1})(\mathbf{u}).$$

It remains to check (3.14). We write

$$\frac{\partial(w \circ \boldsymbol{\theta})}{\partial w_p}(\mathbf{v}) = D(w \circ \boldsymbol{\theta})(\mathbf{v}) \cdot \mathbf{e}_p = Dw(\mathbf{u}) \cdot \boldsymbol{\theta}'(\mathbf{v}) \cdot \mathbf{e}_p,$$

where  $\mathbf{u} = \boldsymbol{\theta}(\mathbf{v})$ . The vector

$$\mathbf{f}_p = \boldsymbol{\theta}'(\mathbf{v}) \cdot \mathbf{e}_p = ((\boldsymbol{\theta}^{-1})'(\mathbf{u}))^{-1} \cdot \mathbf{e}_p$$

obviously satisfies

$$(\boldsymbol{\theta}^{-1})'(\mathbf{u}) \cdot \mathbf{f}_p = \mathbf{e}_p.$$

By the definition of  $\boldsymbol{\theta}$ , this yields

$$Dw_k(\mathbf{u}) \cdot \mathbf{f}_p = 0, \quad k = 1, \dots, p-1.$$

Now, since  $w$  is a  $k$ -Riemann invariant,  $\nabla w(\mathbf{u})$  is orthogonal to  $\mathbf{r}_k(\mathbf{u})$  so that in the basis  $\nabla w_1, \dots, \nabla w_{p-1}, \nabla w_p$ , we have

$$\nabla w(\mathbf{u}) = \sum_{k=1}^{p-1} \alpha_k(\mathbf{u}) \nabla w_k(\mathbf{u}),$$

which implies

$$Dw(\mathbf{u}) \cdot \mathbf{f}_p = 0$$

and proves the desired result.  $\square$

*Example 3.1.* The  $p$ -system (Example 2.2 revisited). Consider once more the  $p$ -system (2.7). Since we have a system of two equations, we are looking for one 1-Riemann invariant  $w_1$  and one 2-Riemann invariant  $w_2$ . By (2.9) we have

$$\nabla w_1(\mathbf{w}) \cdot \mathbf{r}_1(v) = \frac{\partial w_1}{\partial v} + \sqrt{-p'(v)} \frac{\partial w_1}{\partial u} = 0.$$

Hence,  $w_1(\mathbf{w}) = w_1(v, u)$  is given by

$$w_1(v, u) = u - \int^v \sqrt{-p'(y)} dy. \quad (3.15)$$

Similarly, we get

$$\nabla w_2(\mathbf{w}) \cdot \mathbf{r}_2(v) = \frac{\partial w_2}{\partial v} - \sqrt{-p'(v)} \frac{\partial w_2}{\partial u} = 0,$$

so that  $w_2$  is given by

$$w_2(v, u) = u + \int^v \sqrt{-p'(y)} dy. \quad (3.15b)$$

We have thus obtained global Riemann invariants. As a consequence, any  $k$ -Riemann invariant is globally defined as a function of  $w_k$ , for  $k = 1, 2$ .  $\square$

Again, it may be more convenient to work on a nonconservative form of the system of conservation laws (2.1). The above computations for Lemma 3.2 show that the notion of Riemann invariant is independent of the chosen conservative or nonconservative form of (2.1). Indeed, by using the change of dependent variables (2.10) and setting

$$z(\mathbf{v}) = w(\boldsymbol{\theta}(\mathbf{v})),$$

we obtain

$$Dz(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}) = Dw(\boldsymbol{\theta}(\mathbf{v})) \cdot \boldsymbol{\theta}'(\mathbf{v})^{-1} \mathbf{r}_k(\boldsymbol{\theta}(\mathbf{v})) = Dw(\boldsymbol{\theta}(\mathbf{v})) \cdot \mathbf{r}_k(\boldsymbol{\theta}(\mathbf{v})).$$

Hence, a  $k$ -Riemann invariant may be equivalently defined by

$$Dz(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}) = 0.$$

*Example 3.2.* The gas dynamics equations in Lagrangian coordinates. We consider again the Eqs. (2.25). Let us check that the three pairs of Riemann invariants can be taken as

$$(u + \ell, s), (u, p), (u - \ell, s), \quad (3.16)$$

where the function  $\ell = \ell(\tau, s)$  is defined up to an additive function of  $s$  by

$$\frac{\partial \ell}{\partial \tau} = -\frac{c}{\tau}. \quad (3.17)$$

As we have just noticed, we may work on the other conservative form (2.28) of the gas dynamics equations. Since (see (2.33))  $\mathbf{r}_1 = (1, \frac{c}{\tau}, 0)^T$ , the 1-Riemann invariants  $w = w(\tau, u, s)$  are solutions of the equation

$$\frac{\partial w}{\partial \tau} + \frac{c}{\tau} \frac{\partial w}{\partial u} = 0.$$

Hence,  $u + \ell$  and  $s$  are indeed two 1-Riemann invariants whose derivatives

$$\nabla(u + \ell) = \left( \frac{\partial \ell}{\partial \tau}, 1, \frac{\partial \ell}{\partial s} \right)^T, \quad \nabla s = (0, 0, 1)^T$$

are clearly linearly independent. Here again, any 1-Riemann invariant is globally of the form  $Z(u + \ell, s)$ . We proceed in a similar way for the 2- and 3-Riemann invariants.  $\square$

*Example 3.3. The gas dynamics equations in Eulerian coordinates.* Let us consider finally the Eqs. (2.35). Observe that the function  $\ell = \ell(\rho, s)$  defined by (3.17) may be equivalently defined by

$$\frac{\partial \ell}{\partial \rho}(\rho, s) = \frac{c}{\rho}. \quad (3.18)$$

Then using the nonconservative form (2.38) of Eqs. (2.35), it is an easy matter to check that the three pairs of Riemann invariants can be again taken as in (3.16).  $\square$

*Remark 3.2.* The fact that the Riemann invariants are the same in Eulerian and Lagrangian coordinates is a direct consequence of the general result linking the two frames (see Sect. 2.1.2). More precisely, let us note  $\tilde{w}(\mathbf{U}) = w(\mathbf{V}(\mathbf{U}))$ . Then assuming that  $Dw(\mathbf{V}) \cdot \mathbf{s}_k(\mathbf{V}) = 0$ , we get from Theorem 2.1

$$Dw(\mathbf{V}(\mathbf{U})) \cdot \mathbf{s}_k(\mathbf{V}(\mathbf{U})) = Dw(\mathbf{V}(\mathbf{U})) \cdot D\mathbf{V}(\mathbf{U})\mathbf{r}_k(\mathbf{U}) = 0$$

and thus  $D\tilde{w}(\mathbf{U}) \cdot \mathbf{r}_k(\mathbf{U}) = 0$ .  $\square$

Let us check the following simple useful property of Riemann invariants.

*Theorem 3.2*

*On a  $k$ -rarefaction wave, all  $k$ -Riemann invariants are constant.*

*Proof.* Let  $\mathbf{u}$  be a  $k$ -rarefaction wave of the form (3.5), and let  $w$  be a  $k$ -Riemann invariant. The function  $w(\mathbf{u}) : (x, t) \mapsto w(\mathbf{u}(x, t))$  is continuous for  $t > 0$ . First,  $w(\mathbf{u})$  is constant for  $\frac{x}{t} \leq \lambda_k(\mathbf{u}_L)$  and  $\frac{x}{t} \geq \lambda_k(\mathbf{u}_R)$ . Next,

for  $\lambda_k(\mathbf{u}_L) \leq \frac{x}{t} \leq \lambda_k(\mathbf{u}_R)$ ,  $\mathbf{u}$  is an integral curve of  $\mathbf{r}_k$ , which proves the result.  $\square$

More generally, let us define a wider class of smooth  $k$ -waves.

*Definition 3.3*

A smooth solution  $\mathbf{u}(x, t)$  of (2.1) defined on a domain  $D$  of  $\mathbb{R} \times \mathbb{R}_+$  is called a  $k$ -simple wave if  $w(\mathbf{u}(x, t))$  is constant in  $D$  for any  $k$ -Riemann invariant  $w$ .

*Example 3.4. (3.3 Revisited).* For the gas dynamics equations, a 1-simple wave is one for which  $u + \ell$  and  $s$  are constant. We shall use this fact in Chap. III, Sect. 3, and also in Chap. IV for Osher's scheme.  $\square$

*Proposition 3.1*

If  $\mathbf{u}$  is a  $k$ -simple wave, the field of values of  $\mathbf{u}$  is restricted to only one integral curve of  $\mathbf{r}_k$  in  $\mathbb{R}^p$ .

*Proof.* We can find at least one  $k$ -Riemann invariant whose constant values on two integral curves are distinct. Indeed, in the case  $p = 2$ , we have seen that given a curve  $\mathcal{C}$  that is not tangent to  $\mathbf{r}_k$ , we can assign arbitrary values for  $w$ , for instance, strictly increasing along  $\mathcal{C}$ . The value on each integral curve of  $\mathbf{r}_k$  intersecting  $\mathcal{C}$  is then determined, and  $w$  takes distinct values along distinct integral curves. The general case follows similarly.  $\square$

Hence, we look for a function

$$\mathbf{u}(x, t) = \mathbf{v}(\varphi(x, t)),$$

where  $\mathbf{v}$  is solution of (3.10) with some initial value  $\mathbf{v}_0$

$$\begin{cases} \mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi)) \\ \mathbf{v}(\xi_0) = \mathbf{v}_0. \end{cases} \quad (3.19)$$

Such a solution  $\mathbf{v}$  exists at least locally (i.e., for  $\xi$  close to  $\xi_0$ ) as we have already observed. The equation for  $\varphi$  is obtained by writing

$$\begin{aligned} \mathbf{0} &= \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) \\ &= \mathbf{v}'(\varphi) \frac{\partial \varphi}{\partial t} + \mathbf{A}(\mathbf{u}) \mathbf{v}'(\varphi) \frac{\partial \varphi}{\partial x} \\ &= \mathbf{r}_k(\mathbf{v}(\varphi)) \frac{\partial \varphi}{\partial t} + \mathbf{A}(\mathbf{u}) \mathbf{r}_k(\mathbf{v}(\varphi)) \frac{\partial \varphi}{\partial x} \\ &= \mathbf{r}_k(\mathbf{v}(\varphi)) \left\{ \frac{\partial \varphi}{\partial t} + \lambda_k(\mathbf{v}(\varphi)) \frac{\partial \varphi}{\partial x} \right\}. \end{aligned}$$

Thus, we look for  $\varphi$  as a smooth solution of a quasilinear scalar equation

$$\frac{\partial \varphi}{\partial t} + \lambda_k(\mathbf{v}(\varphi)) \frac{\partial \varphi}{\partial x} = 0. \quad (3.20)$$

We have already studied such equations in the Chap. I, Sect. 2. The characteristics, i.e., the integral curves of the differential equation

$$\frac{dx}{dt} = \lambda_k(\mathbf{v}(\varphi(x, t))), \quad (3.21)$$

are straight lines along which  $\varphi$  is constant. We deduce the implicit formula for  $\varphi$  by integrating along the characteristics

$$\varphi(x, t) = \varphi_0 \left( x - \lambda_k(\mathbf{v}(\varphi(x, t)))t \right) \quad \text{if } \varphi(\cdot, 0) = \varphi_0, \quad (3.22)$$

which is valid as long as  $\varphi$  remains smooth. This gives the expression of a  $k$ -simple wave

$$u(x, t) = \mathbf{v}(\varphi_0(x - \lambda_k(\mathbf{v}(\varphi))t)).$$

Let us introduce the *characteristic curves*  $C_k$  of the  $k$ th field. They are defined as the integral curves of the differential system

$$\frac{dx}{dt} = \lambda_k(\mathbf{u}(x, t)). \quad (3.23)$$

When  $\mathbf{u}$  is a  $k$ -simple wave, we have thus proved the following theorem.

*Theorem 3.3*

*Let  $\mathbf{u}$  be a  $k$ -simple wave. Then, the characteristics of the  $k$ th field are straight lines along which  $\mathbf{u}$  is constant.*

We can determine some of the  $k$ -simple waves when the  $k$ th field is either genuinely nonlinear or linearly degenerate.

*Example 3.5. Simple waves for a genuinely nonlinear field.* Assume that the  $k$ th field is genuinely nonlinear. With the normalization (2.53) and (3.19), we have

$$\frac{d}{d\varphi} \left( \lambda_k(\mathbf{v}(\varphi)) \right) = D\lambda_k(\mathbf{v}(\varphi)) \cdot \mathbf{v}'(\varphi) = 1.$$

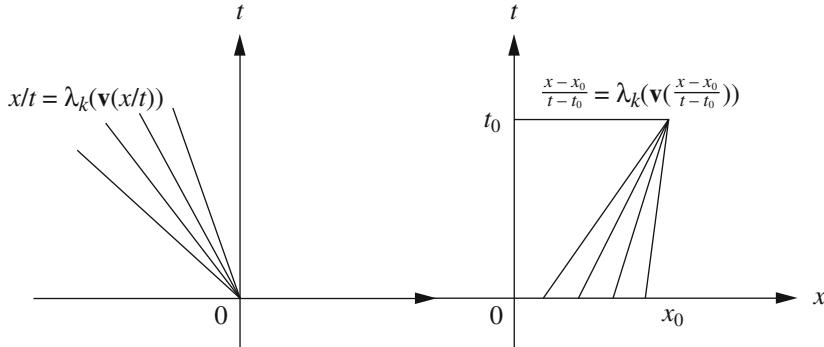
Hence, after possibly changing the function  $\varphi_0$  by an additive constant, we get

$$\lambda_k(\mathbf{v}(\varphi)) = \varphi,$$

and (3.20) becomes

$$\frac{\partial \varphi}{\partial t} + \varphi \frac{\partial \varphi}{\partial x} = 0,$$

which means that  $\varphi$  is a solution of the Burgers equation. As a first example, we find the self-similar centered simple  $k$ -waves previously defined in (3.5)



**Fig. 3.2** Rarefaction and compression centered waves

that correspond to  $\varphi(x, t) = \frac{x}{t}$ , or more generally to  $\frac{(x-x_0)}{(t-t_0)}$ . They are called rarefaction (centered) waves because the characteristics form a fan centered at point  $(x_0, t_0)$  that diverges in opposition to compression centered waves, for which the fan converges, and that generates singularities (see Fig. 3.2). We shall also find compression waves for  $t \geq t_0$  ( $\lambda_k$  decreases from right to left in the fan) in Chap. IV, when studying Osher's scheme. We have also encountered non-self-similar simple waves when resolving the G.R.P. (generalized Riemann problem) in G.R., Chapter 4, Section 3.3 [539]. In the case of a  $k$ -rarefaction wave (3.5) connecting two states  $\mathbf{u}_L$  and  $\mathbf{u}_R$ , the fan that we have already depicted in Sect. 3.1 (see Fig. 3.1) is thus composed of the characteristics  $C_k$  which are straight lines and is bordered by the lines  $\frac{x}{t} = \lambda_k(\mathbf{u}_L)$  and  $\frac{x}{t} = \lambda_k(\mathbf{u}_R)$ .  $\square$

*Example 3.6. Simple waves for a linearly degenerate field.* Assume now that the  $k$ th field is linearly degenerate. We cannot find centered simple waves, as we have already observed, because (3.3) never holds in that case. Nonetheless, we know (see Remark 3.1) that  $\lambda_k$  is a  $k$ -Riemann invariant, so that  $\lambda_k$  is constant on a  $k$ -simple wave. Thus, the functions  $\mathbf{u}$  of the form

$$\mathbf{u}(x, t) = \mathbf{v}(\varphi_0(x - \bar{\lambda}_k t)), \quad \bar{\lambda}_k = \lambda_k(\mathbf{v}), \quad (3.24)$$

with  $\mathbf{v}$  satisfying (3.19), are  $k$ -simple waves, and the characteristics  $C_k$  are now the parallel lines  $x - \bar{\lambda}_k t = C$ . We cannot connect two states  $\mathbf{u}_L$  and  $\mathbf{u}_R$  by a continuous  $k$ -wave, and we must consider a discontinuity wave. This will be the object of the next section.  $\square$

## 4 Shock Waves and Contact Discontinuities

Given two states  $\mathbf{u}_L$  and  $\mathbf{u}_R \in \Omega$ , we are now looking for piecewise constant discontinuous solutions of (2.1) that connect  $\mathbf{u}_L$  and  $\mathbf{u}_R$ . Let us recall that along a line of discontinuity  $x = \xi(t)$  of a weak solution  $\mathbf{u}$  of (2.1),  $\mathbf{u}$  satisfies the Rankine-Hugoniot jump conditions

$$\sigma[\mathbf{u}] = [\mathbf{f}(\mathbf{u})], \quad (4.1)$$

where  $\sigma = \xi'(t)$  is the speed of propagation of the discontinuity (see Chap. I, Sect. 4). Therefore, the function

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_L, & x < \sigma t, \\ \mathbf{u}_R, & x > \sigma t \end{cases} \quad (4.2)$$

is a weak solution of (2.1) provided that the real number  $\sigma$  satisfies the Rankine-Hugoniot condition

$$\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) = \sigma(\mathbf{u}_R - \mathbf{u}_L). \quad (4.3)$$

Such a solution (4.1), (4.2) of the nonlinear hyperbolic system (2.1) is called a *discontinuity wave*. Given a state  $\mathbf{u}_L \in \Omega$ , we want to determine all the states  $\mathbf{u}_R \in \Omega$  to which  $\mathbf{u}_L$  can be connected on the right by a discontinuity wave. Thus, we introduce the following definition.

*Definition 4.1*

The Rankine-Hugoniot set of  $\mathbf{u}_0$  is the set of all states  $\mathbf{u} \in \Omega$  such that there exists  $\sigma(\mathbf{u}_0, \mathbf{u}) \in \mathbb{R}$  with

$$\sigma(\mathbf{u}_0, \mathbf{u})(\mathbf{u} - \mathbf{u}_0) = \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{u}_0). \quad (4.4)$$

The structure of the Rankine-Hugoniot set of  $\mathbf{u}_0$  is given by the following theorem.

*Theorem 4.1*

Let  $\mathbf{u}_0$  be in  $\Omega$ . The Rankine-Hugoniot set of  $\mathbf{u}_0$  is locally made of  $p$  smooth curves  $\mathcal{S}_k(\mathbf{u}_0)$ ,  $1 \leq k \leq p$ . Moreover, for all  $k$ , there exists a parametrization of  $\mathcal{S}_k(\mathbf{u}_0) : \varepsilon \rightarrow \Psi_k(\varepsilon)$  defined for  $|\varepsilon| \leq \varepsilon_1$ ,  $\varepsilon_1$  small enough, such that

$$\Psi_k(\varepsilon) = \mathbf{u}_0 + \varepsilon \mathbf{r}_k(\mathbf{u}_0) + \frac{\varepsilon^2}{2} D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \mathcal{O}(\varepsilon^3) \quad (4.5)$$

and

$$\sigma(\mathbf{u}_0, \Psi_k(\varepsilon)) = \lambda_k(\mathbf{u}_0) + \frac{\varepsilon}{2} D\lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \mathcal{O}(\varepsilon^2). \quad (4.6)$$

*Proof.* We write

$$\begin{aligned}\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{u}_0) &= \int_0^1 \frac{d}{ds} \mathbf{f}(\mathbf{u}_0 + s(\mathbf{u} - \mathbf{u}_0)) ds \\ &= \left( \int_0^1 \mathbf{A}(\mathbf{u}_0 + s(\mathbf{u} - \mathbf{u}_0)) ds \right) (\mathbf{u} - \mathbf{u}_0).\end{aligned}$$

Hence, setting

$$\mathbf{A}(\mathbf{u}, \mathbf{v}) = \int_0^1 \mathbf{A}(\mathbf{u} + s(\mathbf{v} - \mathbf{u})) ds,$$

the jump condition (4.4) becomes

$$(\mathbf{A}(\mathbf{u}_0, \mathbf{u}) - \sigma(\mathbf{u}_0, \mathbf{u}))(\mathbf{u} - \mathbf{u}_0) = \mathbf{0}. \quad (4.7)$$

Note that the  $p \times p$  matrix  $\mathbf{A}(\mathbf{u}_0, \mathbf{u}_0) = \mathbf{A}(\mathbf{u}_0)$  has  $p$  distinct eigenvalues  $\lambda_1(\mathbf{u}_0) < \dots < \lambda_p(\mathbf{u}_0)$ , and the function  $\mathbf{u} \mapsto \mathbf{A}(\mathbf{u}_0, \mathbf{u})$  is continuous. Thus, using a continuity argument, there exists a neighborhood  $\mathcal{N}$  of  $\mathbf{u}_0$  in  $\Omega$  and  $p$  real functions  $\mathbf{u} \mapsto \lambda_k(\mathbf{u}_0, \mathbf{u}), 1 \leq k \leq p$ , defined in  $\mathcal{N}$  such that  $\lambda_k(\mathbf{u}_0, \mathbf{u}), 1 \leq k \leq p$ , are the  $p$  distinct real eigenvalues of  $\mathbf{A}(\mathbf{u}_0, \mathbf{u})$  with

$$\lambda_k(\mathbf{u}_0, \mathbf{u}_0) = \lambda_k(\mathbf{u}_0).$$

We denote by  $\mathbf{l}_k(\mathbf{u}_0, \mathbf{u})^T, 1 \leq k \leq p$ , the “left eigenvectors” of the matrix  $\mathbf{A}(\mathbf{u}_0, \mathbf{u})$ , i.e.,

$$\mathbf{l}_k(\mathbf{u}_0, \mathbf{u})^T \mathbf{A}(\mathbf{u}_0, \mathbf{u}) = \lambda_k(\mathbf{u}_0, \mathbf{u}) \mathbf{l}_k(\mathbf{u}_0, \mathbf{u})^T.$$

Now, using (4.7), a state  $\mathbf{u} \in \mathcal{N}$  belongs to the Rankine-Hugoniot set of  $\mathbf{u}_0$  if and only if there exists an index  $k \in \{1, \dots, p\}$  such that

$$\sigma(\mathbf{u}_0, \mathbf{u}) = \lambda_k(\mathbf{u}_0, \mathbf{u}) \quad (4.8)$$

and

$\mathbf{u} - \mathbf{u}_0$  is a right eigenvector of  $\mathbf{A}(\mathbf{u}_0, \mathbf{u})$  associated with  $\lambda_k(\mathbf{u}_0, \mathbf{u})$ . (4.9)

On the other hand, (4.9) holds if and only if

$$\mathbf{l}_j(\mathbf{u}_0, \mathbf{u})^T (\mathbf{u} - \mathbf{u}_0) = 0, \quad j \neq k. \quad (4.10)$$

This gives a system of  $(p - 1)$  equations in  $p$  unknowns which can be written in the form

$$\mathbf{G}_k(\mathbf{u}) = \mathbf{M}_k(\mathbf{u})(\mathbf{u} - \mathbf{u}_0) = \mathbf{0},$$

where

$$\mathbf{M}_k(\mathbf{u}) = \begin{pmatrix} \mathbf{l}_1(\mathbf{u}_0, \mathbf{u})^T \\ \vdots \\ \mathbf{l}_{k-1}(\mathbf{u}_0, \mathbf{u})^T \\ \mathbf{l}_{k+1}(\mathbf{u}_0, \mathbf{u})^T \\ \vdots \\ \mathbf{l}_p(\mathbf{u}_0, \mathbf{u})^T \end{pmatrix}.$$

We have

$$\mathbf{G}_k(\mathbf{u}_0) = 0, \quad D\mathbf{G}_k(\mathbf{u}_0) = \mathbf{M}_k(\mathbf{u}_0).$$

Moreover, since the vectors  $\mathbf{l}_j(\mathbf{u}_0, \mathbf{u}_0) = \mathbf{l}_j(\mathbf{u}_0)$ ,  $1 \leq j \leq p$ , are linearly independent, the  $(p-1) \times p$  matrix  $\mathbf{M}_k(\mathbf{u}_0)$  has rank  $p-1$ . Therefore, by the implicit function theorem, there exists a one-parameter family  $\mathcal{S}_k(\mathbf{u}_0)$  of solutions of (4.10),  $\theta \mapsto \Psi_k(\theta)$ ,  $|\theta| \leq \theta_1$  small enough, with

$$\Psi_k(0) = \mathbf{u}_0. \quad (4.11)$$

In addition, we have

$$\begin{cases} \sigma(\mathbf{u}_0, \Psi_k(\theta)) = \sigma_k(\mathbf{u}_0, \Psi_k(\theta)), \\ \sigma(\mathbf{u}_0, \Psi_k(0)) = \lambda_k(\mathbf{u}_0). \end{cases} \quad (4.12)$$

Hence, we have proved that the Rankine-Hugoniot set of  $\mathbf{u}_0$  is locally made of  $p$  curves  $\mathcal{S}_k(\mathbf{u}_0)$ . It remains to check the expansions (4.5) and (4.6). First, it follows from (4.10) and (4.11) that

$$\begin{aligned} 0 &= \lim_{\theta \rightarrow 0} \mathbf{l}_j(\mathbf{u}_0, \Psi_k(\theta))^T \left( \frac{\Psi_k(\theta) - \Psi_k(0)}{\theta} \right) \\ &= \mathbf{l}_j(\mathbf{u}_0)^T \Psi'_k(0), \quad j \neq k, \end{aligned}$$

so that  $\Psi'_k(0)$  is collinear to  $\mathbf{r}_k(\mathbf{u}_0)$ . Hence, we can change our parametrization in order to get

$$\Psi'_k(0) = \mathbf{r}_k(\mathbf{u}_0). \quad (4.13)$$

Next, for the sake of simplicity, we set

$$\begin{aligned} \mathbf{A}_k(\theta) &= \mathbf{A}(\Psi_k(\theta)), \\ \sigma_k(\theta) &= \lambda_k(\mathbf{u}_0, \Psi_k(\theta)). \end{aligned}$$

Then, differentiating the condition (4.4)

$$\sigma_k(\Psi_k - \mathbf{u}_0) = \mathbf{f}(\Psi_k) - \mathbf{f}(\mathbf{u}_0),$$

we obtain

$$\sigma'_k(\Psi_k - \mathbf{u}_0) + \sigma_k \Psi'_k = \mathbf{A}_k \Psi'_k$$

and

$$\sigma''_k(\Psi_k - \mathbf{u}_0) + 2\sigma'_k \Psi'_k + \sigma_k \Psi''_k = \mathbf{A}'_k \Psi'_k + \mathbf{A}_k \Psi''_k,$$

which gives for  $\theta = 0$

$$2\sigma'_k(0)\mathbf{r}_k(\mathbf{u}_0) + \lambda_k(\mathbf{u}_0)\Psi''_k(0) = \mathbf{A}'_k(0)\mathbf{r}_k(\mathbf{u}_0) + \mathbf{A}(\mathbf{u}_0)\Psi''_k(0)$$

or equivalently

$$(\mathbf{A}(\mathbf{u}_0) - \lambda_k(\mathbf{u}_0))\Psi''_k(0) + \mathbf{A}'_k(0)\mathbf{r}_k(\mathbf{u}_0) = 2\sigma'_k(0)\mathbf{r}_k(\mathbf{u}_0). \quad (4.14)$$

On the other hand, differentiating

$$\mathbf{A}_k \mathbf{r}_k(\Psi_k) = \lambda_k(\Psi_k) \mathbf{r}_k(\Psi_k)$$

gives

$$\mathbf{A}'_k \mathbf{r}_k(\Psi_k) + \mathbf{A}_k D\mathbf{r}_k(\Psi_k) \cdot \Psi'_k = D\lambda_k(\Psi_k) \cdot \Psi'_k \mathbf{r}_k(\Psi_k) + \lambda_k(\Psi_k) D\mathbf{r}_k(\Psi_k) \cdot \Psi'_k$$

and for  $\theta = 0$

$$\begin{cases} (\mathbf{A}(\mathbf{u}_0) - \lambda_k(\mathbf{u}_0))D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) \\ + \mathbf{A}'_k(0)\mathbf{r}_k(\mathbf{u}_0) - D\lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) = 0. \end{cases} \quad (4.15)$$

Hence, by subtracting (4.15) from (4.14), we get

$$\begin{cases} (\mathbf{A}(\mathbf{u}_0) - \lambda_k(\mathbf{u}_0))(\Psi''_k(0) - D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0)) \\ + D\lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) = 2\sigma'_k(0)\mathbf{r}_k(\mathbf{u}_0). \end{cases} \quad (4.16)$$

Now, multiplying (4.16) by  $\mathbf{l}_k(\mathbf{u}_0)^T$  on the left and using the normalization

$$\mathbf{l}_k(\mathbf{u}_0)^T \mathbf{r}_k(\mathbf{u}_0) = 1,$$

we find

$$\sigma'_k(0) = \frac{1}{2} D\lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0). \quad (4.17)$$

Substituting  $\sigma'_k(0)$  in (4.16) by its expression (4.17) gives

$$(\mathbf{A}(\mathbf{u}_0) - \lambda_k(\mathbf{u}_0))(\Psi''_k(0) - D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0)) = 0.$$

Therefore, there exists a real number  $\beta$  such that

$$\Psi''_k(0) = D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \beta \mathbf{r}_k(\mathbf{u}_0). \quad (4.18)$$

Again, we change our parametrization by setting

$$\theta = \varepsilon - \frac{1}{2}\beta\varepsilon^2.$$

Then, we have by (4.12) and (4.17)

$$\begin{aligned}\sigma(\mathbf{u}_0, \Psi_k(\varepsilon)) &= \lambda_k(\mathbf{u}_0) + \frac{\theta}{2} D\lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \mathcal{O}(\theta^2) \\ &= \lambda_k(\mathbf{u}_0) + \frac{\varepsilon}{2} D\lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \mathcal{O}(\varepsilon^2),\end{aligned}$$

which gives (4.6). Next, using (4.11), (4.13), and (4.18), we obtain

$$\begin{aligned}\Psi_k(\varepsilon) &= \mathbf{u}_0 + \theta \mathbf{r}_k(\mathbf{u}_0) + \frac{\theta^2}{2} (D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \beta \mathbf{r}_k(\mathbf{u}_0)) + \mathcal{O}(\theta^3) \\ &= \mathbf{u}_0 + \varepsilon \mathbf{r}_k(\mathbf{u}_0) + \frac{\varepsilon^2}{2} D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \mathcal{O}(\varepsilon^3)\end{aligned}$$

i.e., the expansion (4.5).  $\square$

*Remark 4.1.* From (4.5), we have

$$\lambda_k(\Psi_k(\varepsilon)) = \lambda_k(\mathbf{u}_0) + \varepsilon D\lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \mathcal{O}(\varepsilon^2),$$

and together with (4.6), we get

$$\sigma(\mathbf{u}_0, \Psi_k(\varepsilon)) = \frac{1}{2} (\lambda_k(\mathbf{u}_0) + \lambda_k(\Psi_k(\varepsilon))) + \mathcal{O}(\varepsilon^2).$$

Thus, the speed of propagation of the discontinuity is approximated at the order  $\mathcal{O}(\varepsilon^2)$  by the mean value of the characteristic speeds on both sides.  $\square$

*Corollary 4.1*

For any  $k$ -Riemann invariant  $w$ , we have

$$w(\Psi_k(\varepsilon)) = w(\mathbf{u}_0) + \mathcal{O}(\varepsilon^3). \quad (4.19)$$

*Proof.* Let  $w$  be a  $k$ -Riemann invariant. By differentiating the relation (3.9), we find

$$D^2 w(\mathbf{u}) \cdot (\mathbf{r}_k(\mathbf{u}), \mathbf{v}) + Dw(\mathbf{u}) \cdot D\mathbf{r}_k(\mathbf{u}) \cdot \mathbf{v} = 0, \quad (4.20)$$

where  $D^j w(\mathbf{u}) \in \mathcal{L}_j(\mathbb{R}^p; \mathbb{R})$  denotes the  $j$ th Frechet derivative of  $w$  at the point  $\mathbf{u}$ . Now, using (4.5) gives

$$\begin{aligned}w(\Psi_k(\varepsilon)) &= w(\mathbf{u}_0 + \varepsilon \mathbf{r}_k(\mathbf{u}_0) + \frac{\varepsilon^2}{2} D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \mathcal{O}(\varepsilon^3)) \\ &= w(\mathbf{u}_0) + \varepsilon Dw(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + \frac{\varepsilon^2}{2} \{Dw(\mathbf{u}_0) \cdot D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) \\ &\quad + D^2 w(\mathbf{u}_0) \cdot (\mathbf{r}_k(\mathbf{u}_0), \mathbf{r}_k(\mathbf{u}_0))\} + \mathcal{O}(\varepsilon^3).\end{aligned}$$

Hence, (4.19) follows from (3.9) and (4.20).  $\square$

Consider the case where the  $k$ th characteristic field is *genuinely nonlinear*. The curve  $S_k(\mathbf{u}_0)$  is then called a  *$k$ -shock curve*. Moreover, using the

normalization (2.53), (4.6) can be written

$$\sigma(\mathbf{u}_0, \Psi_k(\varepsilon)) = \lambda_k(\mathbf{u}_0) + \frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^2). \quad (4.21)$$

If  $\mathbf{u}_R$  belongs to the  $k$ -shock curve  $S_k(\mathbf{u}_L)$ , or equivalently at this stage if  $\mathbf{u}_L$  belongs to the  $k$ -shock curve  $S_k(\mathbf{u}_R)$ , a weak solution of (2.1) of the form (4.2), (4.3) is called a  *$k$ -shock wave*. In fact, we shall see in the next section that not all the states  $\mathbf{u}$  of the  $k$ -shock curve  $S_k(\mathbf{u}_0)$  are admissible but only those that correspond to  $\varepsilon < 0$  (for the normalization (2.53)).

Now, if we consider  $|\varepsilon|$  as a measure of the strength of the  $k$ -shock connecting  $\mathbf{u}_0$  and  $\mathbf{u} = \Psi_k(\varepsilon)$ , it follows from the corollary of Theorem 4.1 that *across a weak  $k$ -shock, the change in any  $k$ -Riemann invariant is of order 3 in  $\varepsilon$*  (the term *weak* shock refers to the fact that  $\varepsilon$  is small, i.e., we consider nearby states).

Let us next turn to the case where the  $k$ th characteristic field is *linearly degenerate*. Then, we can state the following result.

*Theorem 4.2*

*If the  $k$ th characteristic field is linearly degenerate, the curve  $S_k(\mathbf{u}_0)$  given by Theorem 4.1 is an integral curve of the vector field  $\mathbf{r}_k$ , and we have*

$$\sigma(\mathbf{u}_0, \Psi_k(\varepsilon)) = \lambda_k(\Psi_k(\varepsilon)) = \lambda_k(\mathbf{u}_0). \quad (4.22)$$

Moreover, we have for any  $k$ -Riemann invariant  $w$

$$w(\Psi_k(\varepsilon)) = w(\mathbf{u}_0). \quad (4.23)$$

*Proof.* Let us consider the integral curve of the vector field  $\mathbf{r}_k$  passing through the point  $\mathbf{u}_0$ , i.e., the solution  $\xi \rightarrow \mathbf{v}(\xi)$  of

$$\begin{cases} \mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi)), \\ \mathbf{v}(0) = \mathbf{u}_0. \end{cases}$$

Let us check that the Rankine-Hugoniot jump condition (4.4) holds along this integral curve with constant speed  $\sigma(\mathbf{u}_0, \mathbf{u}) = \lambda_k(\mathbf{u}_0)$  if the  $k$ th characteristic field is linearly degenerate. Indeed, we have

$$\begin{aligned} \frac{d}{d\xi} \{ \mathbf{f}(\mathbf{v}(\xi)) - \mathbf{f}(\mathbf{u}_0) - \lambda_k(\mathbf{v}(\xi))(\mathbf{v}(\xi) - \mathbf{u}_0) \} \\ = (\mathbf{A}(\mathbf{v}(\xi)) - \lambda_k(\mathbf{v}(\xi)))\mathbf{v}'(\xi) - D\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi)(\mathbf{v}(\xi) - \mathbf{u}_0) \\ = (\mathbf{A}(\mathbf{v}(\xi)) - \lambda_k(\mathbf{v}(\xi)))\mathbf{r}_k(\mathbf{v}(\xi)) - D\lambda_k(\mathbf{v}(\xi)) \cdot \mathbf{r}_k(\mathbf{v}(\xi))(\mathbf{v}(\xi) - \mathbf{u}_0) \\ = \mathbf{0} \end{aligned}$$

and therefore

$$\mathbf{f}(\mathbf{v}(\xi)) - \mathbf{f}(\mathbf{u}_0) = \lambda_k(\mathbf{v}(\xi))(\mathbf{v}(\xi) - \mathbf{u}_0).$$

Hence, the integral curve coincides with  $\mathcal{S}_k(\mathbf{u}_0)$ , and furthermore

$$\sigma(\mathbf{u}_0, \mathbf{v}(\xi)) = \lambda_k(\mathbf{v}(\xi)).$$

On the other hand, let  $w$  be a  $k$ -Riemann invariant. As we have already observed,  $w$  is constant on an integral curve of  $\mathbf{r}_k$ ; indeed, we have by (3.9)

$$\frac{d}{d\xi} w(\mathbf{v}(\xi)) = Dw(\mathbf{v}(\xi)) \cdot \mathbf{v}'(\xi) = Dw(\mathbf{v}(\xi)) \cdot \mathbf{r}_k(\mathbf{v}(\xi)) = 0$$

so that

$$w(\mathbf{v}(\xi)) = w(\mathbf{u}_0).$$

The theorem follows since  $\lambda_k$  is a  $k$ -Riemann invariant when the  $k$ th characteristic field is linearly degenerate.  $\square$

Thus, assume that the  $k$ th characteristic field is linearly degenerate and that  $\mathbf{u}_R \in \mathcal{S}_k(\mathbf{u}_L)$  or, equivalently,  $\mathbf{u}_L \in \mathcal{S}_k(\mathbf{u}_R)$ . Then, a weak solution  $\mathbf{u}$  of (2.1) of the form (4.2), (4.3), where

$$\sigma = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}_R) = \bar{\lambda}_k$$

i.e.,

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_L, & x < \bar{\lambda}_k t, \\ \mathbf{u}_R, & x > \bar{\lambda}_k t, \end{cases} \quad (4.24)$$

is called a  $k$ -contact discontinuity (see Fig. 4.1).

*Remark 4.2.* It is noteworthy that the function (4.24) is the limit of  $k$ -simple (noncentered) waves (3.24). Indeed, let  $\varphi_0^\varepsilon, \varepsilon > 0$  be a smooth increasing function such that

$$\varphi_0^\varepsilon(x) = \begin{cases} \xi_0, & x \leq 0, \\ \xi_R, & x > \varepsilon, \end{cases}$$

where the solution  $\mathbf{v}$  of (3.19) with  $\mathbf{u}_0 = \mathbf{u}_L$  satisfies  $\mathbf{v}(\xi_R) = \mathbf{u}_R$ . The corresponding  $k$ -simple wave (3.24) is such that

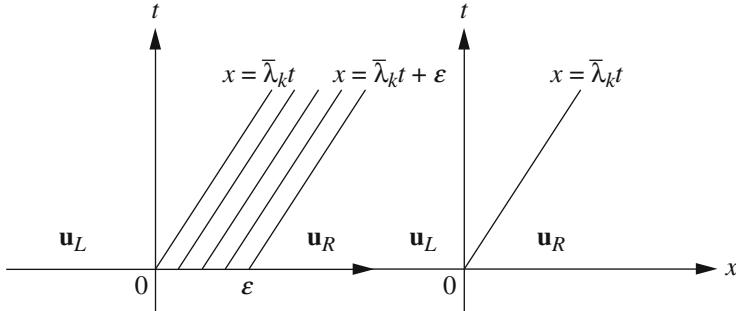
$$\mathbf{u}^\varepsilon(x, t) = \begin{cases} \mathbf{u}_L, & x \leq \bar{\lambda}_k t, \\ \mathbf{u}_R, & x > \bar{\lambda}_k t + \varepsilon \end{cases}$$

and thus approaches (4.24) as  $\varepsilon \rightarrow 0$ .

As we have noticed in Example 3.5, the characteristics of the  $k$ -simple wave are the parallel lines  $x = \bar{\lambda}_k t + C$ .  $\square$

The above results can be readily applied to the  $p$ -system; this will be done in detail in Sect. 7 of this chapter, where we shall indeed obtain a global

description of the rarefaction curves and shock curves. In many applications, however, in order to get the parametrization of the wave curves, as we have already noticed, it could be more convenient to use nonconservative variables. Let us check that the results of Theorems 4.1 and 4.2 are still valid when we use a nonconservative form (2.12) of the nonlinear hyperbolic system (2.1). We introduce new dependent variables  $\mathbf{v}$  defined by (2.10), and we look for a parametrization of the curve  $\mathcal{S}_k(\mathbf{u}_0)$  of the form



**Fig. 4.1** Contact discontinuity

$$\mathbf{v}(\varepsilon) = \mathbf{v}_0 + \varepsilon \mathbf{v}_1 + \varepsilon^2 \mathbf{v}_2 + \mathcal{O}(\varepsilon^3).$$

Setting

$$\mathbf{u}(\varepsilon) = \Psi_k(\varepsilon),$$

we have

$$\begin{aligned} \mathbf{u}(\varepsilon) &= \theta(\mathbf{v}(\varepsilon)) = \theta(\mathbf{v}_0) + \varepsilon D\theta(\mathbf{v}_0) \cdot \mathbf{v}_1 + \varepsilon^2 \{D\theta(\mathbf{v}_0) \cdot \mathbf{v}_2 \\ &\quad + \frac{1}{2} D^2\theta(\mathbf{v}_0) \cdot (\mathbf{v}_1, \mathbf{v}_1)\} + \mathcal{O}(\varepsilon^3). \end{aligned}$$

Comparing this with the expansion (4.5) gives

$$\begin{aligned} \mathbf{u}_0 &= \theta(\mathbf{v}_0), \\ \mathbf{r}_k(\mathbf{u}_0) &= D\theta(\mathbf{v}_0) \cdot \mathbf{v}_1, \end{aligned}$$

$$D\mathbf{r}_k(\mathbf{u}_0) = 2D\theta(\mathbf{v}_0) \cdot \mathbf{v}_2 + D^2\theta(\mathbf{v}_0) \cdot (\mathbf{v}_1, \mathbf{v}_1), \quad (4.25)$$

and so on. Hence, we find

$$\mathbf{v}_0 = \theta^{-1}(\mathbf{u}_0)$$

and, by (2.14),

$$\mathbf{v}_1 = D\theta(\mathbf{v}_0)^{-1} \mathbf{r}_k(\theta(\mathbf{v}_0)) = \mathbf{s}_k(\mathbf{v}_0).$$

Moreover, by differentiating

$$\mathbf{r}_k(\boldsymbol{\theta}(\mathbf{v})) = D\boldsymbol{\theta}(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}),$$

we obtain

$$D\mathbf{r}_k(\boldsymbol{\theta}(\mathbf{v})) \cdot D\boldsymbol{\theta}(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}) = D^2\boldsymbol{\theta}(\mathbf{v}) \cdot (\mathbf{s}_k(\mathbf{v}), \mathbf{s}_k(\mathbf{v})) + D\boldsymbol{\theta}(\mathbf{v}) \cdot D\mathbf{s}_k(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v}).$$

Thus, we have

$$D\mathbf{r}_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = D^2\boldsymbol{\theta}(\mathbf{v}) \cdot (\mathbf{s}_k(\mathbf{v}), \mathbf{s}_k(\mathbf{v})) + D\boldsymbol{\theta}(\mathbf{v}) \cdot D\mathbf{s}_k(\mathbf{v}) \cdot \mathbf{s}_k(\mathbf{v})$$

and, since  $\mathbf{v}_1 = \mathbf{s}_k(\mathbf{v}_0)$ ,

$$D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) = D^2\boldsymbol{\theta}(\mathbf{v}_0) \cdot (\mathbf{v}_1, \mathbf{v}_1) + D\boldsymbol{\theta}(\mathbf{v}_0) \cdot D\mathbf{s}_k(\mathbf{v}_0) \cdot \mathbf{s}_k(\mathbf{v}_0),$$

so that by (4.25)

$$\mathbf{v}_2 = \frac{1}{2}D\mathbf{s}_k(\mathbf{v}_0) \cdot \mathbf{s}_k(\mathbf{v}_0).$$

Therefore, we get

$$\mathbf{v}(\varepsilon) = \mathbf{v}_0 + \varepsilon \mathbf{s}_k(\mathbf{v}_0) + \frac{\varepsilon^2}{2} D\mathbf{s}_k(\mathbf{v}_0) \cdot \mathbf{s}_k(\mathbf{v}_0) + \mathcal{O}(\varepsilon^3). \quad (4.26)$$

On the other hand, setting  $\sigma(\varepsilon) = \sigma(\mathbf{u}, \boldsymbol{\Psi}_k(\varepsilon))$  and using (2.13), (2.15) together with (4.6), we have

$$\sigma(\varepsilon) = \mu_k(\mathbf{v}_0) + \frac{\varepsilon}{2} D\mu_k(\mathbf{v}_0) \cdot \mathbf{s}_k(\mathbf{v}_0) + \mathcal{O}(\varepsilon^2). \quad (4.27)$$

Finally, when the  $k$ th characteristic field is linearly degenerate, we know that

$$\mathbf{u}'(\varepsilon) = \mathbf{r}_k(\mathbf{u}(\varepsilon)),$$

so that

$$D\boldsymbol{\theta}(\mathbf{v}(\varepsilon)) \cdot \mathbf{v}'(\varepsilon) = D\boldsymbol{\theta}(\mathbf{v}(\varepsilon)) \cdot \mathbf{s}_k(\mathbf{v}(\varepsilon)).$$

Hence

$$\begin{aligned} \mathbf{v}'(\varepsilon) &= \mathbf{s}_k(\mathbf{v}(\varepsilon)), \\ \mathbf{v}(0) &= \mathbf{v}_0, \end{aligned}$$

$\mathcal{S}_k(\mathbf{u}_0)$  is an integral curve of the vector field  $\mathbf{s}_k(\mathbf{v})$ , and moreover,

$$\mu_k(\mathbf{v}(\varepsilon)) = \mu_k(\mathbf{v}_0) = \lambda_k(\mathbf{u}_0). \quad (4.28)$$

*Example 4.1.* Consider the gas dynamics equations in Lagrangian coordinates written in the form (2.28), i.e., using the nonconservative variables  $(\tau, u, s)$ .

It follows from (2.33) that

$$D\mathbf{r}_1 \cdot \mathbf{r}_1 = \begin{pmatrix} \tau \\ \tau \frac{\partial c}{\partial \tau} \\ 0 \end{pmatrix}, \quad D\mathbf{r}_3 \cdot \mathbf{r}_3 = \begin{pmatrix} \tau \\ -\tau \frac{\partial c}{\partial \tau} \\ 0 \end{pmatrix}.$$

Hence, the 1-shock curve passing through the point  $(\tau_0, u_0, s_0)$  has a parametrization of the form

$$\begin{cases} \tau(\varepsilon) = \tau_0 \left(1 + \varepsilon + \frac{\varepsilon^2}{2}\right) + \mathcal{O}(\varepsilon^3), \\ u(\varepsilon) = u_0 + \varepsilon c_0 + \frac{\varepsilon^2}{2} \tau_0 \left(\frac{\partial c}{\partial \tau}\right)_0 + \mathcal{O}(\varepsilon^3), \\ s(\varepsilon) = s_0 + \mathcal{O}(\varepsilon^3). \end{cases} \quad (4.29)$$

We find that the change in entropy  $s$  is of third order in  $\varepsilon$ , as was expected from Corollary 4.1 of Theorem 4.1 since  $s$  is a 1-Riemann invariant (see Example 3.2). On the other hand, for such a 1-shock, we have

$$\begin{aligned} p(\varepsilon) &= p(\tau(\varepsilon), s(\varepsilon)) = p(\tau_0 + \varepsilon \tau_0 + \dots, s_0 + \dots) \\ &= p(\tau_0, s_0) + \varepsilon \tau_0 \frac{\partial p}{\partial \tau}(\tau_0, s_0) + \mathcal{O}(\varepsilon^2) \end{aligned}$$

and therefore, by (2.21),

$$p(\varepsilon) = p_0 - \varepsilon \frac{c_0^2}{\tau_0} + \mathcal{O}(\varepsilon^2). \quad (4.30)$$

Similarly, the 3-shock curve passing through the point  $(\tau_0, u_0, s_0)$  has a parametrization of the form

$$\begin{cases} \tau(\varepsilon) = \tau_0 \left(1 + \varepsilon + \frac{\varepsilon^2}{2}\right) + \mathcal{O}(\varepsilon^3), \\ u(\varepsilon) = u_0 - \varepsilon c_0 - \frac{\varepsilon^2}{2} \tau_0 \left(\frac{\partial c}{\partial \tau}\right)_0 + \mathcal{O}(\varepsilon^3), \\ s(\varepsilon) = s_0 + \mathcal{O}(\varepsilon^3) \end{cases} \quad (4.31)$$

and

$$p(\varepsilon) = p_0 - \varepsilon \frac{c_0^2}{\tau_0} + \mathcal{O}(\varepsilon^2). \quad (4.32)$$

Finally, since  $u, p$  are 2-Riemann invariants, the 2-contact discontinuity curve passing through  $(\tau_0, u_0, s_0)$  satisfies, by Theorem 4.2,

$$u(\varepsilon) = u_0, \quad p(\varepsilon) = p_0, \quad (4.33)$$

and we have

$$\sigma(\varepsilon) = 0 \quad (4.34)$$

for this 2-contact discontinuity.  $\square$

*Example 4.2.* We turn to the gas dynamics equations in Eulerian coordinates written in the nonconservative form (2.39). Using (2.43), we have

$$D\mathbf{r}_1 \cdot \mathbf{r}_1 = \begin{pmatrix} \rho \\ -\rho \frac{\partial c}{\partial \rho} \\ 0 \end{pmatrix}, \quad D\mathbf{r}_3 \cdot \mathbf{r}_3 = \begin{pmatrix} \rho \\ \rho \frac{\partial c}{\partial \rho} \\ 0 \end{pmatrix}.$$

And the 1-shock curve passing through the point  $(\rho_0, u_0, s_0)$  has a parametrization of the form

$$\begin{cases} \rho(\varepsilon) = \rho_0 \left( 1 + \varepsilon + \frac{\varepsilon^2}{2} \right) + \mathcal{O}(\varepsilon^3), \\ u(\varepsilon) = u_0 - \varepsilon c_0 - \frac{\varepsilon^2}{2} \rho_0 \left( \frac{\partial c}{\partial \rho} \right)_0 + \mathcal{O}(\varepsilon^3), \\ s(\varepsilon) = s_0 + \mathcal{O}(\varepsilon^3), \end{cases} \quad (4.35)$$

while the 3-shock curve passing through this point is given by

$$\begin{cases} \rho(\varepsilon) = \rho_0 \left( 1 + \varepsilon + \frac{\varepsilon^2}{2} \right) + \mathcal{O}(\varepsilon^3), \\ u(\varepsilon) = u_0 + \varepsilon c_0 + \frac{\varepsilon^2}{2} \rho_0 \left( \frac{\partial c}{\partial \rho} \right)_0 + \mathcal{O}(\varepsilon^3), \\ s(\varepsilon) = s_0 + \mathcal{O}(\varepsilon^3). \end{cases} \quad (4.36)$$

Note that for both shock curves, we have

$$p(\varepsilon) = p_0 + \varepsilon \rho_0 c_0^2 + \mathcal{O}(\varepsilon^2). \quad (4.37)$$

On the other hand, the 2-contact discontinuity curve passing through the point  $(\rho_0, u_0, s_0)$  again satisfies

$$u(\varepsilon) = u_0, \quad p(\varepsilon) = p_0 \quad (4.38)$$

with

$$\sigma(\varepsilon) = u_0, \quad (4.39)$$

which characterizes the physical contact discontinuities.  $\square$

## 5 Characteristic Curves and Entropy Conditions

### 5.1 Characteristic Curves

Assume that  $\mathbf{u}$  is a classical solution of the system (2.1) in a domain  $D$  so that

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0} \quad \text{in } D$$

or equivalently, by (2.3),

$$\mathbf{l}_k(\mathbf{u})^T \left( \frac{\partial \mathbf{u}}{\partial t} + \lambda_k(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} \right) = 0, \quad 1 \leq k \leq p. \quad (5.1)$$

Each Eq. (5.1) involves differentiation in only one direction. Introducing as in (3.21) the characteristic curves  $C_k$  as the integral curves of the differential system

$$\frac{dx}{dt} = \lambda_k(\mathbf{u}(x, t)), \quad (5.2)$$

and denoting by  $s_k \mapsto (x(s_k), t(s_k))$  a parametric representation of  $C_k$ , each Eq. (5.1) becomes

$$\mathbf{l}_k(\mathbf{u})^T \frac{d\mathbf{u}}{ds_k} = 0, \quad 1 \leq k \leq p.$$

Equations (5.1) are therefore called the *characteristic equations*.

Note that the change (2.10) of dependent variables does not affect the eigenvalues  $\lambda_k$  and thus the characteristic curves.

*Example 5.1.* Consider first the case  $p = 2$ . For ease of notation, we denote by  $\lambda$  and  $\mu$  the eigenvalues of the  $2 \times 2$  matrix  $\mathbf{A}(\mathbf{u})$ , by  $\mathbf{r}_\lambda$  and  $\mathbf{r}_\mu$  (resp.  $\mathbf{l}_\lambda^T$  and  $\mathbf{l}_\mu^T$ ) the corresponding right (resp. left) eigenvectors. Next, we introduce a Riemann invariant  $w$  (resp.  $z$ ) associated with the eigenvalue  $\lambda$  (resp.  $\mu$ ) which satisfies by definition

$$\nabla w^T \mathbf{r}_\lambda = 0 \quad (\text{resp. } \nabla z^T \mathbf{r}_\mu = 0).$$

Recall that we have obtained in Example 3.1 the global existence of such Riemann invariants. Since  $\mathbf{l}_\mu^T \mathbf{r}_\lambda = 0$ , it follows that  $\nabla w$  is collinear to  $\mathbf{l}_\mu$  so that

$$\nabla w^T \mathbf{A} = \mu \nabla w^T.$$

Hence, we obtain

$$\frac{\partial w}{\partial t} + \mu \frac{\partial w}{\partial x} = \nabla w^T \left( \frac{\partial \mathbf{u}}{\partial t} + \mu \frac{\partial \mathbf{u}}{\partial x} \right) = \nabla w^T \left( \frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} \right) = 0.$$

Similarly, we find

$$\frac{\partial z}{\partial z} + \lambda \frac{\partial z}{\partial x} = \nabla z^T \left( \frac{\partial \mathbf{u}}{\partial t} + \lambda \frac{\partial \mathbf{u}}{\partial x} \right) = \nabla z^T \left( \frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} \right) = 0.$$

Hence, the system of characteristic equations is given here by

$$\begin{cases} \frac{\partial z}{\partial t} + \lambda \frac{\partial z}{\partial x} = 0, \\ \frac{\partial w}{\partial t} + \mu \frac{\partial w}{\partial x} = 0. \end{cases} \quad (5.3)$$

If we denote by  $C_\lambda$  (resp.  $C_\mu$ ) the characteristic curves associated with the eigenvalue  $\lambda$  (resp.  $\mu$ ), we find that  $w$  is constant along the characteristics  $C_\mu$ , while  $z$  is constant along the characteristics  $C_\lambda$ .

If we turn to the  $p$ -system (2.7), we have (see Example 3.1)

$$\begin{aligned} \lambda &= -\sqrt{-p'(v)}, \quad \mu = \sqrt{-p'(v)}, \\ w &= u - \int^v \sqrt{-p'(y)} dy, \quad z = u + \int^v \sqrt{-p'(y)} dy. \end{aligned}$$

As a consequence, if instead of  $C_\lambda$  and  $C_\mu$  we denote by  $C_-$  and  $C_+$ , respectively, the characteristic curves, we find that  $u \pm \int^v \sqrt{-p'(y)} dy$  is constant on the  $C_\mp$  characteristics. Then, if by some device we have determined the characteristic curves, we can obtain the solution  $(v, u)$  of the  $p$ -system provided that this solution is smooth. The theory for systems of two equations is made simpler by the use of the two Riemann invariants and the simple form of the characteristic equations (see Lax [745]); more details concerning “reducible” hyperbolic systems can be found, for instance, in Li Ta-tsien [794].  $\square$

*Example 5.2.* Again, we consider the gas dynamics equations in Lagrangian coordinates (Example 3.2 revisited). Let us derive the system of characteristic equations. We start from the other conservative form (2.28). Since

$$dp = -\left(\frac{c^2}{\tau^2}\right)d\tau + \frac{\partial p}{\partial s}ds,$$

it follows from the third Eq. (2.28) that

$$\frac{\partial p}{\partial t} = -\frac{c^2}{\tau^2} \frac{\partial \tau}{\partial t},$$

so that by the first Eq. (2.28)

$$\frac{\partial p}{\partial t} + \frac{c^2}{\tau^2} \frac{\partial u}{\partial x} = 0.$$

Hence, another nonconservative form of the gas dynamics equations in Lagrangian coordinates is given by

$$\begin{cases} \frac{\partial p}{\partial t} + \frac{c^2}{\tau^2} \frac{\partial u}{\partial x} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial x} = 0, \\ \frac{\partial s}{\partial t} = 0. \end{cases} \quad (5.4)$$

By multiplying the second Eq. (5.4) by  $\pm \frac{c}{\tau}$  and adding to the first equation, we obtain the characteristic equations

$$\begin{cases} \left\{ \frac{\partial p}{\partial t} - \frac{c}{\tau} \frac{\partial p}{\partial x} \right\} - \frac{c}{\tau} \left\{ \frac{\partial u}{\partial t} - \frac{c}{\tau} \frac{\partial u}{\partial x} \right\} = 0, \\ \frac{\partial s}{\partial t} = 0, \\ \left\{ \frac{\partial p}{\partial t} + \frac{c}{\tau} \frac{\partial p}{\partial x} \right\} + \frac{c}{\tau} \left\{ \frac{\partial u}{\partial t} + \frac{c}{\tau} \frac{\partial u}{\partial x} \right\} = 0. \end{cases} \quad (5.5)$$

Note that the second Eq. (5.5) may be equivalently replaced by

$$\frac{\partial p}{\partial t} + \frac{c^2}{\tau^2} \frac{\partial \tau}{\partial t} = 0.$$

Therefore, the corresponding characteristic curves are given by

$$\begin{cases} \frac{dx}{dt} = -\frac{c}{\tau} & (C_-), \\ \frac{dx}{dt} = 0 & (C_0), \\ \frac{dx}{dt} = \frac{c}{\tau} & (C_+). \end{cases} \quad (5.6)$$

We find that the characteristics  $C_0$  are straight lines along which the entropy  $s$  is constant.

We have

$$\begin{cases} dp - \frac{c}{\tau} du = 0 & \text{along the } C_- \text{ characteristics,} \\ ds = 0 \text{ or } dp + \left( \frac{c^2}{\tau^2} \right) d\tau = 0 & \text{along the } C_0 \text{ characteristics,} \\ dp + \frac{c}{\tau} du = 0 & \text{along the } C_+ \text{ characteristics,} \end{cases} \quad (5.7)$$

which is another way of writing the characteristic equations.  $\square$

*Example 5.3.* We turn to the gas dynamics equations written in Eulerian coordinates (Example 3.3 revisited). To derive the characteristic equations, we use the nonconservative equations of Remark 2.2 together with the equation of state and (2.41) and we explicit the computations. Since

$$dp = c^2 d\rho + \frac{\partial p}{\partial s} ds,$$

we obtain from the third Eq. (2.39)

$$\left( \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} \right) - c^2 \left( \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} \right) = 0.$$

Together with the first Eq. (2.39), this gives

$$\frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + \rho c^2 \frac{\partial u}{\partial x} = 0. \quad (5.8)$$

Multiplying the second Eq. (2.39) by  $\pm \rho c$  and adding to the above equation, we find

$$\begin{cases} \left\{ \frac{\partial p}{\partial t} + (u - c) \frac{\partial p}{\partial x} \right\} - \rho c \left\{ \frac{\partial u}{\partial t} + (u - c) \frac{\partial u}{\partial x} \right\} = 0, \\ \frac{\partial s}{\partial t} + u \frac{\partial s}{\partial x} = 0, \\ \left\{ \frac{\partial p}{\partial t} + (u + c) \frac{\partial p}{\partial x} \right\} + \rho c \left\{ \frac{\partial u}{\partial t} + (u + c) \frac{\partial u}{\partial x} \right\} = 0, \end{cases} \quad (5.9)$$

which is the desired characteristic system, while the associated characteristic curves are

$$\begin{cases} \frac{dx}{dt} = u - c & (C_-), \\ \frac{dx}{dt} = u & (C_0), \\ \frac{dx}{dt} = u + c & (C_+), \end{cases} \quad (5.10)$$

Again, the characteristic equations (5.9) may be equivalently written

$$\begin{cases} dp - \rho c du = 0 & \text{along the } C_- \text{ characteristics,} \\ ds = 0 \text{ or } dp - c^2 d\rho = 0 & \text{along the } C_0 \text{ characteristics,} \\ dp + \rho c du = 0 & \text{along the } C_+ \text{ characteristics} \end{cases} \quad (5.11)$$

(compared with (5.7)). □

In the general case, we obtain the following result, which appears to be a consequence of Theorem 3.3.

*Theorem 5.1*

Assume that  $\mathbf{u}$  is a  $k$ -rarefaction wave in  $D$ . Then the characteristic curves  $C_k$  are straight lines along which  $\mathbf{u}$  is constant.

*Proof.* We give a direct proof. Let  $w_j, 1 \leq j \leq p-1$ , be  $(p-1)$   $k$ -Riemann invariants, with derivatives  $Dw_j$  linearly independent in  $D$ . Since by Theorem 3.2,  $w_j(\mathbf{u})$  is constant in  $D$ , we have

$$\nabla w_j(\mathbf{u})^T \left( \frac{\partial \mathbf{u}}{\partial t} + \lambda_k \frac{\partial \mathbf{u}}{\partial x} \right) = \frac{d}{ds_k} w_j(\mathbf{u}) = 0.$$

Together with (5.1), this gives

$$\mathbf{N}_k(\mathbf{u}) \left( \frac{\partial \mathbf{u}}{\partial t} + \lambda_k \frac{\partial \mathbf{u}}{\partial x} \right) = \mathbf{0},$$

where  $\mathbf{N}_k(\mathbf{u})$  is the  $p \times p$  matrix

$$\mathbf{N}_k(\mathbf{u}) = \begin{pmatrix} \nabla w_1(\mathbf{u})^T \\ \vdots \\ \nabla w_{p-1}(\mathbf{u})^T \\ \mathbf{l}_k(\mathbf{u})^T \end{pmatrix}.$$

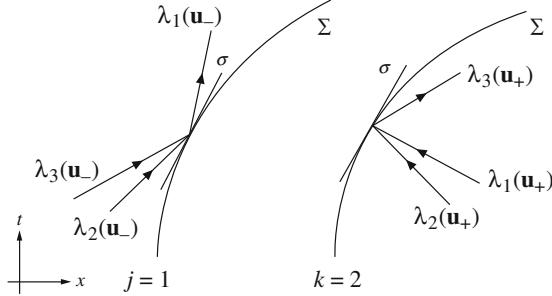
Since  $\nabla w_j(\mathbf{u})^T \mathbf{r}_k(\mathbf{u}) = 0$ ,  $1 \leq j \leq p-1$ , and  $\mathbf{l}_k(\mathbf{u})^T \mathbf{r}_k(\mathbf{u}) = 1$ , the matrix  $\mathbf{N}_k(\mathbf{u})$  is nonsingular so that in  $D$

$$\frac{\partial \mathbf{u}}{\partial t} + \lambda_k(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}.$$

Hence,  $\mathbf{u}$  is constant along each  $C_k$ -characteristic, and  $C_k$  is indeed a straight line.  $\square$

## 5.2 The Lax Entropy Conditions

Let us consider again the problem of determining all the states  $\mathbf{u}_R \in \Omega$  that can be connected to a given state  $\mathbf{u}_L \in \Omega$  by a discontinuity wave. We have already noticed in the scalar case that not any shock discontinuity was admissible. In the case of systems, we want to show how heuristic arguments based on characteristics lead us to exclude a part of the  $k$ -shock curve  $\mathcal{S}_k(\mathbf{u}_L)$ . Observe first that in the smoothness regions of a weak solution  $\mathbf{u}$  of (2.1),



**Fig. 5.1** Characteristics and discontinuity curve

the characteristic curves propagate the information from the boundary data. In particular, if  $\Sigma$  is a curve  $x = \xi(t)$  in the  $(x, t)$ -plane with  $\sigma = \xi'(t)$  and if

$$\lambda_1(\mathbf{u}) < \dots < \lambda_k(\mathbf{u}) < \sigma < \lambda_{k+1}(\mathbf{u}) < \dots < \lambda_p(\mathbf{u}) \text{ on } \Sigma$$

(see Fig. 5.1), we need to give  $(p - k)$  boundary conditions on  $\Sigma$  (resp.  $k$  boundary conditions on  $\Sigma$ ) in order to specify the solution  $\mathbf{u}$  in the region  $\{(x, t), x > \sigma t\}$  (resp. in the region  $x \leq \sigma t$ ).

Now, if a piecewise  $C^1$  weak solution  $\mathbf{u}$  of (2.1) is discontinuous across the curve  $\Sigma$  and satisfies

$$\lambda_k(\mathbf{u}_+) < \sigma < \lambda_{k+1}(\mathbf{u}_+), \quad (5.12)$$

we need to specify  $(p - k)$  conditions on the right boundary of the discontinuity. Similarly, if

$$\lambda_j(\mathbf{u}_-) < \sigma < \lambda_{j+1}(\mathbf{u}_-), \quad (5.13)$$

we have to give  $j$  conditions on the left boundary of the discontinuity, which yields in total  $p - k + j$  conditions. On the other hand, eliminating  $\alpha$  from the Rankine-Hugoniot jump relations

$$\sigma(\mathbf{u}_+ - \mathbf{u}_-) = \mathbf{f}(\mathbf{u}_+) - \mathbf{f}(\mathbf{u}_-)$$

provides  $(p - 1)$  relations between  $\mathbf{u}_+$  and  $\mathbf{u}_-$ . Hence, assuming that (5.12) holds, it seems natural to require that (5.13) holds with  $j = k - 1$ .

If  $\mathbf{u}$  satisfies  $\lambda_k(\mathbf{u}_+) = \sigma$ , the same argument requires that  $\lambda_k(\mathbf{u}_-) = \sigma$ . Thus, we introduce the following definition.

### Definition 5.1

We shall say that the discontinuity satisfies the Lax entropy conditions if there exists an index  $k \in \{1, 2, \dots, p\}$  such that we have either

(i)

$$\begin{cases} \lambda_k(\mathbf{u}_+) < \sigma < \lambda_{k+1}(\mathbf{u}_+), \\ \lambda_{k-1}(\mathbf{u}_-) < \sigma < \lambda_k(\mathbf{u}_-) \end{cases} \quad (5.14)$$

if the  $k$ th characteristic field is genuinely nonlinear; or

(ii)

$$\lambda_k(\mathbf{u}_-) = \sigma = \lambda_k(\mathbf{u}_+) \quad (5.15)$$

if the  $k$ th characteristic field is linearly degenerate.

Then, using the parametrization of Theorem 4.1, we define  $\mathcal{S}_k^a(\mathbf{u}_L)$  as the set of states  $\Psi_k(\varepsilon) \in \mathcal{S}_k(\mathbf{u}_L)$  that can be connected to  $\mathbf{u}_L$  (on the right of  $\mathbf{u}_L$ ) by a  $k$ -discontinuity wave that satisfies the Lax entropy conditions.

*Theorem 5.2*

If the  $k$ th characteristic field is genuinely nonlinear, the curve  $\mathcal{S}_k^a(\mathbf{u}_L)$  consists of the states  $\Psi_k(\varepsilon) \in \mathcal{S}_k(\mathbf{u}_L)$  that satisfy

$$\varepsilon \leq 0, |\varepsilon| \leq \varepsilon_1 \text{ small enough} \quad (5.16)$$

(for the normalization (2.53)).

If the  $k$ th characteristic field is linearly degenerate,  $\mathcal{S}_k^a(\mathbf{u}_L)$  coincides with the whole curve  $\mathcal{S}_k(\mathbf{u}_L)$ .

*Proof.* Assume that the  $k$ th characteristic field is genuinely nonlinear. Then, setting

$$\mathbf{u}(\varepsilon) = \Psi_k(\varepsilon), \quad \sigma(\varepsilon) = \sigma(\mathbf{u}_L, \Psi_k(\varepsilon)), \quad (5.17)$$

we have by (4.5) and (4.21)

$$\begin{aligned} \mathbf{u}(\varepsilon) &= \mathbf{u}_L + \varepsilon \mathbf{r}_k(\mathbf{u}_L) + \mathcal{O}(\varepsilon^2) \\ \sigma(\varepsilon) &= \lambda_k(\mathbf{u}_L) + \frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^2), \end{aligned}$$

so that by (2.53)

$$\begin{aligned} \lambda_k(\mathbf{u}(\varepsilon)) &= \lambda_k(\mathbf{u}_L) + \varepsilon D\lambda_k(\mathbf{u}_L) \cdot \mathbf{r}_k(\mathbf{u}_L) + \mathcal{O}(\varepsilon^2) \\ &= \lambda_k(\mathbf{u}_L) + \varepsilon + 0(\varepsilon^2) = \sigma(\varepsilon) + \frac{\varepsilon}{2} + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Now, we need to find the conditions that ensure

$$\begin{aligned} \lambda_k(\mathbf{u}(\varepsilon)) &< \sigma(\varepsilon) < \lambda_{k+1}(\mathbf{u}(\varepsilon)), \\ \lambda_{k-1}(\mathbf{u}_L) &< \sigma(\varepsilon) < \lambda_k(\mathbf{u}_L). \end{aligned}$$

First, we have  $\lambda_k(\mathbf{u}(\varepsilon)) < \sigma(\varepsilon)$  for  $\varepsilon$  small enough if and only if  $\varepsilon$  is  $< 0$ . Next, since  $\lambda_k(\mathbf{u}_L) < \lambda_{k+1}(\mathbf{u}_L)$  and

$$\sigma(\varepsilon) \rightarrow \lambda_k(\mathbf{u}_L), \quad \lambda_{k+1}(\mathbf{u}(\varepsilon)) \rightarrow \lambda_{k+1}(\mathbf{u}_L) \text{ as } \varepsilon \rightarrow 0,$$

we obtain  $\sigma(\varepsilon) < \lambda_{k+1}(\mathbf{u}_L)$  for  $|\varepsilon|$  small enough.

On the other hand, we have  $\sigma(\varepsilon) < \lambda_k(\mathbf{u}_L)$  for  $|\varepsilon|$  small enough if and only if  $\varepsilon$  is  $< 0$ . Finally, since  $\lambda_{k-1}(\mathbf{u}_L) < \lambda_k(\mathbf{u}_L)$ , we get  $\lambda_{k-1}(\mathbf{u}_L) < \sigma(\varepsilon)$  for  $|\varepsilon|$  small enough. This proves the first part of the theorem.

When the  $k$ th characteristic field is linearly degenerate, the result of Theorem 4.2 ensures that the whole curve  $\mathcal{S}_k(\mathbf{u}_L)$  satisfies the Lax entropy conditions.  $\square$

### 5.3 Other Entropy Conditions

#### 5.3.1 Entropy-Entropy Flux

Another way of selecting the relevant part of the curve  $\mathcal{S}_k(\mathbf{u}_L)$  is based on entropy considerations (see the Chap. I, Sect. 5). Let us recall that a convex smooth function  $U : \Omega \rightarrow \mathbb{R}$  is an entropy if there exists a smooth function  $F : \Omega \rightarrow \mathbb{R}$ , called entropy flux, such that

$$U'(\mathbf{u})\mathbf{A}(\mathbf{u}) = F'(\mathbf{u}), \quad \forall \mathbf{u} \in \Omega.$$

Then, a piecewise  $C^1$  weak solution of (2.1) is an entropy solution if, across each discontinuity line, it satisfies together with the Rankine-Hugoniot jump condition (4.1) the following entropy condition:

$$\sigma[U(\mathbf{u})] \geq [F(\mathbf{u})] \quad (5.18)$$

for all entropy pairs  $(U, F)$ .

Hence, using again the parametrization of Theorem 4.1, we want to determine the states  $\Psi_k(\varepsilon) \in \mathcal{S}_k(\mathbf{u}_L)$  that satisfy the inequality

$$\sigma(\mathbf{u}_L, \Psi_k(\varepsilon))(U(\Psi_k(\varepsilon)) - U(\mathbf{u}_L)) \geq F(\Psi_k(\varepsilon)) - F(\mathbf{u}_L). \quad (5.19)$$

*Theorem 5.3*

Let  $(U, F)$  be an entropy pair. If the  $k$ th characteristic field is genuinely nonlinear and if  $U$  is strictly convex, the inequality (5.19) holds for  $|\varepsilon|$  small enough if and only if  $\varepsilon \leq 0$ .

If the  $k$ th characteristic field is linearly degenerate, we have

$$\sigma(\mathbf{u}_L, \mathbf{u})(U(\mathbf{u}) - U(\mathbf{u}_L)) = F(\mathbf{u}) - F(\mathbf{u}_L), \quad \forall \mathbf{u} \in \mathcal{S}_k(\mathbf{u}_L). \quad (5.20)$$

*Proof.* Given an entropy pair  $(U, F)$ , we set

$$E(\varepsilon) = \sigma(\varepsilon)(U(\mathbf{u}(\varepsilon)) - U(\mathbf{u}_L)) - (F(\mathbf{u}(\varepsilon)) - F(\mathbf{u}_L)) \quad (5.21)$$

where  $\mathbf{u}(\varepsilon)$  and  $\sigma(\varepsilon)$  are defined as in (5.17). We want to determine the values of  $\varepsilon$  for which

$$E(\varepsilon) \geq 0, \quad |\varepsilon| \text{ small enough.}$$

Clearly, we have

$$E(0) = 0.$$

Next, differentiating (5.21) gives

$$E'(\varepsilon) = \sigma'(\varepsilon)(U(\mathbf{u}(\varepsilon)) - U(\mathbf{u}_L)) + \sigma(\varepsilon)DU(\mathbf{u}(\varepsilon)) \cdot \mathbf{u}'(\varepsilon) - DF(\mathbf{u}(\varepsilon)) \cdot \mathbf{u}'(\varepsilon),$$

and by the condition  $U' \mathbf{A} = F'$

$$E'(\varepsilon) = \sigma'(\varepsilon)(U(\mathbf{u}(\varepsilon)) - U(\mathbf{u}_L)) - DU(\mathbf{u}(\varepsilon)) \cdot (\mathbf{A}(\mathbf{u}(\varepsilon)) - \sigma(\varepsilon)\mathbf{u}'(\varepsilon)). \quad (5.22)$$

On the other hand, by differentiating the Rankine-Hugoniot jump condition

$$\sigma(\varepsilon)(\mathbf{u}(\varepsilon) - \mathbf{u}_L) = f(\mathbf{u}(\varepsilon)) - f(\mathbf{u}_L),$$

we obtain

$$\sigma'(\varepsilon)(\mathbf{u}(\varepsilon) - \mathbf{u}_L) = (\mathbf{A}(\mathbf{u}(\varepsilon)) - \sigma(\varepsilon)\mathbf{u}'(\varepsilon)). \quad (5.23)$$

Hence, by combining (5.22) and (5.23), we get

$$E'(\varepsilon) = \sigma'(\varepsilon)\{U(\mathbf{u}(\varepsilon)) - U(\mathbf{u}_L) - DU(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}(\varepsilon) - \mathbf{u}_L)\} \quad (5.24)$$

and therefore

$$E'(0) = 0.$$

Again, differentiating (5.24) gives

$$\begin{cases} E''(\varepsilon) = \sigma''(\varepsilon)(U(\mathbf{u}(\varepsilon)) - U(\mathbf{u}_L) - DU(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}(\varepsilon) - \mathbf{u}_L) \\ \quad - \sigma'(\varepsilon)D^2U(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}'(\varepsilon), \mathbf{u}(\varepsilon) - \mathbf{u}_L) \end{cases} \quad (5.25)$$

so that

$$E''(0) = 0.$$

Differentiating (5.25) once more, we find

$$\begin{aligned} E'''(\varepsilon) &= \sigma'''(\varepsilon)\{U(\mathbf{u}(\varepsilon)) - U(\mathbf{u}_L) - DU(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}(\varepsilon) - \mathbf{u}_L)\} \\ &\quad - 2\sigma''(\varepsilon)D^2U(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}'(\varepsilon), \mathbf{u}(\varepsilon) - \mathbf{u}_L) \\ &\quad - \sigma'(\varepsilon)\{D^3U(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}'(\varepsilon), \mathbf{u}'(\varepsilon), \mathbf{u}(\varepsilon) - \mathbf{u}_L) \\ &\quad + D^2U(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}''(\varepsilon), \mathbf{u}(\varepsilon) - \mathbf{u}_L) + D^2U(\mathbf{u}(\varepsilon)) \cdot (\mathbf{u}'(\varepsilon), \mathbf{u}'(\varepsilon))\} \end{aligned}$$

and

$$E'''(0) = -\sigma'(0)D^2U(\mathbf{u}_L)(\mathbf{u}'(0), \mathbf{u}'(0)).$$

Assume now that the  $k$ th characteristic field is genuinely nonlinear. We have by (4.5) and (4.21)

$$\mathbf{u}'(0) = \mathbf{r}_k(\mathbf{u}_L), \quad \sigma'(0) = \frac{1}{2},$$

and therefore

$$E(\varepsilon) = -\frac{\varepsilon^3}{12}D^2U(\mathbf{u}_L) \cdot (\mathbf{r}_k(\mathbf{u}_L), \mathbf{r}_k(\mathbf{u}_L)) + \mathcal{O}(\varepsilon^4). \quad (5.26)$$

If we suppose also that the entropy function  $U$  is strictly convex, we obtain that  $E(\varepsilon) \geq 0$  for  $|\varepsilon|$  small enough if and only if  $\varepsilon \leq 0$ .

Assume next that the  $k$ th characteristic field is linearly degenerate. Then by Theorem 4.2,  $\mathcal{S}_k(\mathbf{u}_L)$  is an integral curve of the vector field  $\mathbf{r}_k$  so that  $\mathbf{u}'(\varepsilon)$  is proportional to  $\mathbf{r}_k(\mathbf{u}(\varepsilon))$ . Moreover,  $\sigma(\varepsilon) = \lambda_k(\mathbf{u}_L) = \lambda_k(\mathbf{u}(\varepsilon))$ . Hence, it follows from (5.22) that

$$E'(\varepsilon) = 0.$$

Since  $E(0) = 0$ , we obtain  $E(\varepsilon) = 0$  for all  $\varepsilon$ , which proves (5.20).  $\square$

Hence, we have proven that, for sufficiently weak shocks, the Lax entropy conditions (5.14) are equivalent to the condition (5.18) associated with a strictly convex entropy  $U$ . In the general case, however, it is not yet clear whether the above entropy conditions are equivalent or not. In practice, it is often more convenient to use the Lax entropy conditions to determine the admissible part  $\mathcal{S}_k^a(\mathbf{u}_L)$  of the shock curve  $\mathcal{S}_k(\mathbf{u}_L)$ .

*Remark 5.1.* Assume that the  $k$ th characteristic field is genuinely nonlinear. Then, arguing exactly as in the proof of Theorem 5.2 or Theorem 5.3, one can check that the set of left states  $\Psi_k(\varepsilon) \in \mathcal{S}_k(\mathbf{u}_R)$  that can be connected to a given right state  $\mathbf{u}_R$  by an admissible  $k$ -shock wave consists of the states  $\Psi_k(\varepsilon)$  that satisfy

$$\varepsilon > 0, |\varepsilon| \text{ small enough}$$

(for the normalization (2.53)).  $\square$

### 5.3.2 Liu's Entropy Condition, Viscous Profile

Liu [822] has introduced another entropy condition, noted  $(E)$ , which extends Oleinik's condition (see G.R., Chapter 2, Lemma 6.1 [539]) to a system. It is written

$$\sigma(\mathbf{u}_L, \mathbf{u}_R) \leq \sigma(\mathbf{u}_L, \mathbf{u}) \text{ for any } \mathbf{u} \in \mathcal{S}_k(\mathbf{u}_L) \text{ between } \mathbf{u}_L \text{ and } \mathbf{u}_R, \quad (5.27)$$

(or  $\sigma(\mathbf{u}_L, \mathbf{u}_R) < \sigma(\mathbf{u}_L, \mathbf{u})$  for Liu's strict condition). Condition  $(E)$  implies the shock inequality

$$\lambda_k(\mathbf{u}_R) < \sigma(\mathbf{u}_L, \mathbf{u}_R) < \lambda_k(\mathbf{u}_L),$$

and, for a characteristic field that is genuinely nonlinear, condition  $(E)$  is equivalent to Lax's entropy inequalities (5.14) (Liu [821, 824]). However, if  $D\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u})$  vanishes at some state on  $\mathcal{S}_k(\mathbf{u}_L)$ , one uses condition  $(E)$  instead of Lax's entropy condition to solve the Riemann problem.

Moreover, Majda and Pego [843] have proven that, for weak shocks, Liu's strict entropy condition is equivalent to the existence of a viscous shock profile. We define a *viscous profile* as follows: it is a (smooth) *traveling wave*  $\mathbf{u}_\varepsilon$ , i.e.,

$$\mathbf{u}_\varepsilon(x, t) = \mathbf{v} \left( \frac{x - \sigma t}{\varepsilon} \right), \quad (5.28)$$

that is solution of a parabolic system

$$\frac{\partial \mathbf{u}_\varepsilon}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}_\varepsilon) = \varepsilon \frac{\partial}{\partial x} \left( \mathbf{D}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x} \right), \quad (5.29)$$

where  $\mathbf{D}$  is some "admissible" viscosity matrix (see the Chap. I, Remark 5.3), and such that the vanishing viscosity limit of  $\mathbf{u}_\varepsilon$  is an admissible shock (4.2), which reads

$$\mathbf{u}_\varepsilon(x, t) \longrightarrow \mathbf{u}(x, t) = \begin{cases} \mathbf{u}_L, & x < \sigma t, \\ \mathbf{u}_R, & x > \sigma t \end{cases} \quad (5.30)$$

as  $\varepsilon \rightarrow 0$ . The problem of the existence of a viscous profile is called the shock-structure problem and was introduced by Gelfand. Majda and Pego [843] have characterized the admissible matrices  $\mathbf{D}$  in terms of linear uniform  $\mathbf{L}^2$ -stability (or well-posedness) of the viscous system linearized about a constant state. Note that this linear stability criterion can be characterized in terms of the algebraic structure of  $\mathbf{D}$ . In turn, this structure of  $\mathbf{D}$  is implied by the existence of an entropy pair  $(U, F)$  such that  $U''\mathbf{D}$  is positive-definite (Mock's criterion) (see Remark 5.3 in the Chap. I).

Conversely, the existence of viscous profiles can also be used to derive a viscosity criterion ("chord condition") for selecting admissible shocks, for instance, in the case of the system of one-dimensional elasticity (which corresponds to a  $p$ -system that is not strictly hyperbolic since  $p''$  vanishes at one point and  $p'$  twice). In this example, the viscosity matrix is diagonal and is obtained from physical considerations (viscoelastic system), and the "admissible" shocks are thus selected as the limit of viscous profiles (see Pego [931] and the references therein). Freistühler [490] has considered the same problem for hyperbolic systems with rotational invariance (which are not strictly hyperbolic).

All these criteria for selecting reasonable weak solutions are thus linked but not always equivalent (depending on the problems), strictly hyperbolic or not, and weak or strong shocks (we refer to the above mentioned papers for details; see also Dafermos [382], Menikoff and Plohr [862], Brio [204], Azevedo and Marchesin [69], Holden et al. [623], Shearer and Scheder [1050] and Scheder et al. [1015], Warnecke [1182], Temple [1111], Schochet [1111], Schochet and Tadmor [1021], Čanic and Plohr [231, 684], Serre [1037] see also [1042]), and the subject is connected to the stability of nonlinear waves [1085].

Moreover, we have restricted ourselves to viscous shock profiles; diffusive-dispersive traveling waves, in link with nonclassical shock waves, are discussed in [753].

Before concluding this remark, let us find the conditions satisfied by a viscous profile, i.e., a solution of (5.28)–(5.30). First, we notice that in order that  $\mathbf{u}_\varepsilon$  be a solution of (5.29),  $\mathbf{v}(\xi)$  in (5.28) should satisfy the differential system

$$-\sigma \mathbf{v}' + \mathbf{A}(\mathbf{v})\mathbf{v}' = (\mathbf{D}(\mathbf{v})\mathbf{v}')', \quad (5.31)$$

which no longer depends on the viscosity coefficient  $\varepsilon$ . Also, if  $\mathbf{u}_\varepsilon$  is a viscous profile, by assumption the limit of  $\mathbf{u}_\varepsilon$  must be an admissible shock (5.30). Hence, on the one hand, the speed  $\sigma$  of the traveling wave is given by the Rankine-Hugoniot condition

$$\sigma(\mathbf{u}_R - \mathbf{u}_L) = \mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)$$

and thus depends on the conservation form but not on the viscosity matrix  $\mathbf{D}$ . On the other hand, we see that  $\mathbf{v}$  should satisfy the “boundary conditions”

$$\mathbf{v}(\xi) \rightarrow \mathbf{u}_L \text{ as } \xi \rightarrow -\infty, \quad \mathbf{v}(\xi) \rightarrow \mathbf{u}_R \text{ as } \xi \rightarrow +\infty. \quad (5.32)$$

Indeed, the family  $\mathbf{v}_\varepsilon(\xi) = \mathbf{v}(\frac{\xi}{\varepsilon})$  converges a.e. as  $\varepsilon \rightarrow 0, \varepsilon > 0$ , to  $\mathbf{v}_0(\xi)$

$$\mathbf{v}\left(\frac{\xi}{\varepsilon}\right) \longrightarrow \mathbf{v}_0(\xi) = \begin{cases} \lim_{\xi \rightarrow -\infty} \mathbf{v}(\xi), & \xi < 0, \\ \lim_{\xi \rightarrow +\infty} \mathbf{v}(\xi), & \xi > 0. \end{cases} \quad (5.33)$$

Thus, setting  $\xi = x - \sigma t$  and identifying the limits yields (5.32).

Now, integrating the system (5.31), we get

$$-\sigma \mathbf{v} + \mathbf{f}(\mathbf{v}) = \mathbf{D}(\mathbf{v})\mathbf{v}' + C, \quad (5.34)$$

and taking the limit of (5.34) as  $\xi \rightarrow \pm\infty$ , we obtain

$$C = -\sigma \mathbf{u}_L + \mathbf{f}(\mathbf{u}_L) = -\sigma \mathbf{u}_R + \mathbf{f}(\mathbf{u}_R).$$

Thus, a viscous profile, if it exists (which is the case if the shock connecting  $\mathbf{u}_R$  to  $\mathbf{u}_L$  satisfies Liu's strict entropy condition and  $\mathbf{D}$  is admissible), is a solution of the nonlinear system of ordinary differential equations

$$\mathbf{D}(\mathbf{v})\mathbf{v}' = \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}_L) - \sigma(\mathbf{v} - \mathbf{u}_L) \quad (5.35)$$

together with (5.32), i.e.,  $\mathbf{v}$  is an orbit (or trajectory) connecting the critical (or rest) points  $\mathbf{u}_L$  to  $\mathbf{u}_R$ .

*Remark 5.2.* In many practical problems, we encounter nonlinear hyperbolic systems in *nonconservative* form

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}. \quad (5.36)$$

In order to solve the Riemann problem for such a system, we need to define what we mean by a shock wave. Following the lines of the previous remark, we consider a parabolic regularization of the nonconservative system

$$\frac{\partial \mathbf{u}_\varepsilon}{\partial t} + \mathbf{A}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x} = \varepsilon \frac{\partial}{\partial x} \left( \mathbf{D}(\mathbf{u}_\varepsilon) \frac{\partial \mathbf{u}_\varepsilon}{\partial x} \right), \quad (5.37)$$

where  $\mathbf{D}$  is a given viscosity matrix. A traveling wave solution of (5.36) is a smooth function  $\mathbf{v}$ , with say  $\mathbf{v}' \in \mathbf{L}^1(\mathbb{R})$ , of the form

$$\mathbf{u}_\varepsilon(x, t) = \mathbf{v} \left( \frac{x - \sigma t}{\varepsilon} \right). \quad (5.38)$$

We check that the corresponding differential system for  $\mathbf{v}$  is still (5.31), which does not depend on  $\varepsilon$ , but it cannot be integrated in the form (5.34)(5.33),

If there exists a solution  $\mathbf{v}$  to (5.31) with the “boundary conditions” (5.32), this gives a way of defining a shock wave solution  $\mathbf{u}$  of the nonconservative system (5.36) that connects  $\mathbf{u}_L, \mathbf{u}_R$  by setting

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_L, & x < \sigma t, \\ \mathbf{u}_R, & x > \sigma t. \end{cases}$$

This solution, i.e., the triple  $(\mathbf{u}_L, \mathbf{u}_R, \sigma)$ , now depends on the diffusion (or viscosity) matrix  $\mathbf{D}$  (we shall say that it is consistent with  $\mathbf{D}$ ) and is the limit of viscous profiles when the viscosity tends to 0. Indeed, set  $\mathbf{u}_\varepsilon(x, t) = \mathbf{v} \left( \frac{x - \sigma t}{\varepsilon} \right)$ , where  $\mathbf{v}$  satisfies (5.31), (5.32). Then  $\mathbf{u}_\varepsilon$  satisfies (5.37) and, using (5.33), we have a.e. as  $\varepsilon \rightarrow 0$

$$\mathbf{u}_\varepsilon(x, t) \longrightarrow \begin{cases} \mathbf{u}_L, & x < \sigma t, \\ \mathbf{u}_R, & x > \sigma t. \end{cases} \quad (5.39)$$

We emphasize again that when the system is conservative ( $\mathbf{A}(\mathbf{u}) = \mathbf{f}'(\mathbf{u})$ ), a necessary condition for the existence of a solution  $\mathbf{v}$  to (5.31), (5.32) is the Rankine-Hugoniot condition, which links the triple  $(\mathbf{u}_L, \mathbf{u}_R, \sigma)$ , and the shock wave solutions do not depend on  $\mathbf{D}$ .

Other approaches for defining discontinuous solutions to (5.36) suppose that one can give a meaning in some way or other to the *nonconservative product*  $\mathbf{A}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x}$  when  $\mathbf{u}$  is, say, a Heaviside function. This can be done following the theory of Dal Maso et al. [387], who extend the work of Volpert; the definition of the jump  $[\mathbf{A}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x}]_\Phi$  then depends on a path  $\Phi$  connecting  $\mathbf{u}_L$  and  $\mathbf{u}_R$  in the states space (a straight line in the case of Volpert’s product); see also [759]. A different approach, using the generalized functions, consists roughly in defining the product  $HH'$ , where  $H$  is the Heaviside function, by

$HH' \sim aH'$ ; here  $a$  is some parameter that is obtained in a unique way when the equations have a conservative form. For instance, in the case of Burgers' equation written in nonconservation form

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

substituting a shock  $u(x, t) = (u_R - u_L)H(x - \sigma t) + u_L$  gives formally

$$-\sigma H' \Delta u + (\Delta u H + u_L) \Delta u H' = 0,$$

and thus

$$\Delta u H H' = (\sigma - u_L) H'.$$

Since

$$\sigma = \frac{u_L + u_R}{2}$$

for the corresponding conservative equation, this yields  $HH' = \frac{1}{2}H'$  (i.e.,  $a = \frac{1}{2}$ ). Otherwise, one must invoke physical considerations when the equations are derived from physics (see Colombeau et al. [334]).

In any case, some extra information (choice of  $\Phi$  or  $a$ ) is needed in order to define discontinuous solutions of (5.36). The above considerations show that this information can also be taken from a diffusion matrix  $\mathbf{D}$ , which leads to a definition of weak shocks as the limit of viscous profiles (see Raviart and Sainsaulieu [968], and Sainsaulieu [999]). The choice of  $\mathbf{D}$  is usually guided by physical considerations.

We will come again on nonconservative systems in Chap. VII, since they are involved in the numerical approximation of systems with geometric source terms. In that case, we will have to consider contact discontinuities, for which the extra information will appear naturally.  $\square$

## 6 Solution of the Riemann Problem

We are now able to solve the Riemann problem for the system (2.1)

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = \mathbf{0}, & x \in \mathbb{R}, \quad t > 0, \\ \mathbf{u}(x, 0) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0 \end{cases} \end{cases} \quad (6.1)$$

when the data  $\mathbf{u}_L$  and  $\mathbf{u}_R$  are sufficiently close ( $\mathbf{u}_L - \mathbf{u}_R$  is “small”). To do so, we begin by summarizing some of the results of the previous sections. Assume first that the  $k$ th characteristic field is genuinely nonlinear and normalized

by (2.53). It follows from (3.6) and (4.5) that the rarefaction and shock curves  $\mathcal{R}_k(\mathbf{u}_L)$  and  $\mathcal{S}_k(\mathbf{u}_L)$  are osculatory at  $\varepsilon = 0$ . Hence, the function  $\chi_k : (\varepsilon, \mathbf{u}_L) \mapsto \chi_k(\varepsilon; \mathbf{u}_L)$  defined for  $|\varepsilon|$  small enough by

$$\chi_k(\varepsilon; \mathbf{u}_L) = \begin{cases} \Phi_k(\varepsilon), & \varepsilon \geq 0, \\ \Psi_k(\varepsilon), & \varepsilon \leq 0 \end{cases} \quad (6.2)$$

is of class  $C^2$ . Moreover, using Theorems 3.1 and 5.2, we obtain that the set

$$\{\chi_k(\varepsilon; \mathbf{u}_L), |\varepsilon| \text{ small enough}\} \quad (6.3)$$

is exactly the set of all neighboring states  $\mathbf{u}$  that can be connected to  $\mathbf{u}_L$  (on the right of  $\mathbf{u}_L$ ) either by a  $k$ -rarefaction wave or by an admissible  $k$ -shock wave.

Next, when the  $k$ th characteristic field is linearly degenerate, setting

$$\chi_k(\varepsilon; \mathbf{u}_L) = \Psi_k(\varepsilon) \quad (6.4)$$

we obtain that (6.3) is the set of all neighboring states  $\mathbf{u}$  that can be connected to  $\mathbf{u}_L$  by a  $k$ -contact discontinuity.

Now, we can state the main result of this chapter.

*Theorem 6.1*

Assume that for all  $k \in 1, \dots, p$ , the  $k$ th characteristic field is either genuinely nonlinear or linearly degenerate. Then for all  $\mathbf{u}_L \in \Omega$ , there exists a neighborhood  $\vartheta$  of  $\mathbf{u}_L$  in  $\Omega$  with the following property: if  $\mathbf{u}_R$  belongs to  $\vartheta$ , the Riemann problem (6.1) has a weak solution that consists of at most  $(p+1)$  constant states separated by rarefaction waves, admissible shock waves, or contact discontinuities. Moreover, a weak solution of this kind is unique.

A solution of this kind is depicted in Fig. 6.1.

*Proof.* Let  $\mathbf{u}_L$  be in  $\Omega$ . We consider the mapping

$$\chi : \varepsilon = (\varepsilon_1, \dots, \varepsilon_p)^T \mapsto \chi(\varepsilon) = \chi_p(\varepsilon_p; \chi_{p-1}(\varepsilon_{p-1}; \dots; \chi_1(\varepsilon_1; u_L) \dots))$$

defined in a neighborhood of  $\mathbf{0}$  in  $\mathbb{R}^p$  with values in  $\Omega \in \mathbb{R}^p$ . In other words, the left state  $\mathbf{u}_L$  is connected on the right to  $\chi_1(\varepsilon_1; \mathbf{u}_L) = \mathbf{u}_1$  by a 1-wave, then  $\mathbf{u}_1$  to  $\mathbf{u}_2 = \chi_2(\varepsilon_2; \mathbf{u}_1)$  on the right by a 2-wave, ..., and  $\mathbf{u}_{p-1}$  to  $\mathbf{u}_p = \chi_p(\varepsilon_p; \mathbf{u}_{p-1})$  on the right by a  $p$ -wave. We want to check whether we can reach in that way any state  $\mathbf{u}_R \in \Omega$  located in a neighborhood of  $\mathbf{u}_L$  or, equivalently, solve the equation

$$\chi(\varepsilon) = \mathbf{u}_R. \quad (6.5)$$

We begin by noticing that  $\chi$  is a mapping of class  $C^2$  and

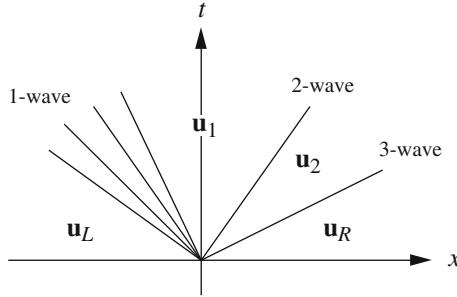
$$\chi(0) = \chi_p(0; \chi_{p-1}(0; \dots; \chi_1(0, \mathbf{u}_L) \dots)) = \mathbf{u}_L.$$

On the other hand, we have by (3.6) and (4.5)

$$\chi_k(\varepsilon_k; \mathbf{u}) = \mathbf{u} + \varepsilon_k \mathbf{r}_k(\mathbf{u}) + \mathcal{O}(\varepsilon_k^2),$$

so that

$$\begin{aligned}\chi_2(\varepsilon_2; \chi_1(\varepsilon_1; \mathbf{u}_L)) &= \chi_2(\varepsilon_2; \mathbf{u}_L + \varepsilon_1 \mathbf{r}_1(\mathbf{u}_L) + \mathcal{O}(\varepsilon_1^2)) \\ &= \mathbf{u}_L + \varepsilon_1 \mathbf{r}_1(\mathbf{u}_L) + \mathcal{O}(\varepsilon_1^2) + \varepsilon_2 \mathbf{r}_2(\mathbf{u}_L + \varepsilon_1 \mathbf{r}_1(\mathbf{u}_L) + \mathcal{O}(\varepsilon_1^2)) + \mathcal{O}(\varepsilon_2^2) \\ &= \mathbf{u}_L + \varepsilon_1 \mathbf{r}_1(\mathbf{u}_L) + \varepsilon_2 \mathbf{r}_2(\mathbf{u}_L) + \mathcal{O}(\varepsilon_1^2 + \varepsilon_2^2)\end{aligned}$$



**Fig. 6.1** Solution of the Riemann problem in  $(x, t)$ -space

and by induction

$$\chi(\varepsilon) = \mathbf{u}_L + \sum_{k=1}^p \varepsilon_k \mathbf{r}_k(\mathbf{u}_L) + \mathcal{O}(|\varepsilon|^2).$$

This means exactly that the derivative  $D\chi(0) \in \mathcal{L}(\mathbb{R}^p)$  of  $\chi$  at the origin is given by

$$D\chi(0) \cdot \boldsymbol{\eta} = \sum_{k=1}^p \eta_k \mathbf{r}_k(\mathbf{u}_L) \quad \forall \boldsymbol{\eta} = (\eta_1, \dots, \eta_p)^T \in \mathbb{R}^p.$$

Since the vectors  $\mathbf{r}_k(\mathbf{u}_L), 1 \leq k \leq p$ , are linearly independent, the linear operator  $D\chi(0)$  is invertible. By the local inversion theorem, there exists a neighborhood  $\vartheta$  of  $\mathbf{u}_L$  in  $\Omega$  such that, for all  $\mathbf{u}_R \in \vartheta$ , Eq. (6.5) has a unique solution  $\varepsilon \in \mathbb{R}^p$ . We thus obtain a solution  $\mathbf{u}$  consisting of  $(p+1)$  constant states  $\mathbf{u}_0 = \mathbf{u}_L, \mathbf{u}_1, \dots, \mathbf{u}_p = \mathbf{u}_R$  separated by  $k$ -waves,  $1 \leq k \leq p$ , that satisfy the Lax entropy conditions.  $\square$

If  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)^T$  is the solution of (6.5),  $\varepsilon_k$  is called the strength of the  $k$ th wave in the solution of the Riemann problem (6.1).

*Remark 6.1.* Consider now the case where the system is hyperbolic but not strictly hyperbolic, with a complete set of eigenvectors and eigenvalues having

constant multiplicity. More precisely, assume that at least one eigenvalue has multiplicity greater than one but that each characteristic field is still either genuinely nonlinear or linearly degenerate. Then the results of Theorem 6.1 extend since first the vectors  $\mathbf{r}_k(\mathbf{u}), 1 \leq k \leq p$ , are linearly independent and, moreover, one can prove that there is just one possibility for the corresponding wave, i.e., a contact discontinuity, because the multiple fields are linearly degenerate (see Boillat [153]). Indeed, if, say,  $\lambda_k(\mathbf{u}) = \lambda_{k+1}(\mathbf{u})$ , differentiating the identity  $(\mathbf{A}(\mathbf{u}) - \lambda(\mathbf{u}))\mathbf{r}_k(\mathbf{u}) = \mathbf{0}$  and taking the scalar product of the resulting equations with  $\mathbf{l}_k(\mathbf{u})^T$  and  $\mathbf{l}_{k+1}(\mathbf{u})^T$  yields for any vector  $\mathbf{v} \in \mathbb{R}^p$

$$\begin{aligned}\mathbf{l}_k(\mathbf{u})^T \{ \mathbf{A}'(\mathbf{u})(\mathbf{v}, \mathbf{r}_k(\mathbf{u})) - (D\lambda(\mathbf{u}) \cdot \mathbf{v})\mathbf{r}_k(\mathbf{u}) \} &= 0, \\ \mathbf{l}_{k+1}(\mathbf{u})^T \{ \mathbf{A}'(\mathbf{u})(\mathbf{v}, \mathbf{r}_k(\mathbf{u})) - (D\lambda(\mathbf{u}) \cdot \mathbf{v})\mathbf{r}_k(\mathbf{u}) \} &= 0,\end{aligned}$$

hence by (2.4)

$$\begin{cases} \mathbf{l}_k(\mathbf{u})^T \mathbf{A}'(\mathbf{u})(\mathbf{v}, \mathbf{r}_k(\mathbf{u})) = D\lambda(\mathbf{u}) \cdot \mathbf{v}, \\ \mathbf{l}_{k+1}(\mathbf{u})^T \mathbf{A}'(\mathbf{u})(\mathbf{v}, \mathbf{r}_k(\mathbf{u})) = 0. \end{cases} \quad (6.6)$$

Similarly, for  $\mathbf{r}_{k+1}(\mathbf{u})$

$$\begin{cases} \mathbf{l}_{k+1}(\mathbf{u})^T \mathbf{A}'(\mathbf{u})(\mathbf{v}, \mathbf{r}_{k+1}(\mathbf{u})) = D\lambda(\mathbf{u}) \cdot \mathbf{v}, \\ \mathbf{l}_k(\mathbf{u})^T \mathbf{A}'(\mathbf{u})(\mathbf{v}, \mathbf{r}_{k+1}(\mathbf{u})) = 0. \end{cases} \quad (6.7)$$

Taking  $\mathbf{v} = \mathbf{r}_{k+1}(\mathbf{u})$  in (6.5) (resp.  $\mathbf{r}_k(\mathbf{u})$  in (6.6)) and using the symmetry of the Hessian gives

$$D\lambda(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = D\lambda(\mathbf{u}) \cdot \mathbf{r}_{k+1}(\mathbf{u}) = 0.$$

One encounters such a situation of multiple eigenvalues in the multicomponent Euler equations (see Chap. III, Remark 3.5) and for the one-dimensional system obtained by projecting the 2D Euler system (see Chap. V, Sect. 2, Remark 2.6).  $\square$

*Remark 6.2.* Theorem 6.1 and the above remark give the solution of the Riemann problem for nearby states when the system is “convex,” i.e., when each characteristic field is either genuinely nonlinear or linearly degenerate. However, when the system is strictly hyperbolic but not convex, i.e., if  $D\lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u})$  vanishes at some state on  $\mathcal{S}_k(\mathbf{u}_L)$ , then we may have  $\sigma(\mathbf{u}_L, \mathbf{u}_1) = \lambda_k(\mathbf{u}_1)$  at some state  $\mathbf{u}_1$ . Hence, if  $\mathbf{u}_R$  lies beyond  $\mathbf{u}_1$  on  $\mathcal{S}_k(\mathbf{u}_L)$ , we cannot connect  $\mathbf{u}_L$  to  $\mathbf{u}_R$  by a unique shock satisfying Lax’s condition. Instead, following the argument of the scalar case when  $f$  is nonconvex (see G.R., Chapter 2, Section 6 [539]), we must consider composite  $k$ -waves, i.e., a succession of  $k$ -shocks (admissible in the sense of Liu; see Sect. 5.3.2), and  $k$ -rarefactions. The states  $\mathbf{u}_i$ , where we switch from a shock to a rarefaction,

satisfy  $\sigma(\mathbf{u}_i, \mathbf{u}_{i+1}) = \lambda_k(\mathbf{u}_i)$ . We refer to Liu [821] for details and Menikoff and Plohr [862] for the case of fluid dynamics.

The question of nonuniqueness (in the class described by Theorem 6.1) of the “entropy” solution for the Riemann problem appears for a system that is not strictly hyperbolic if there exists an “umbilic point,” i.e., a state  $\mathbf{u}_0$  such that two eigenvalues coincide at  $\mathbf{u}_0$ ; the set of eigenvectors may be complete at  $\mathbf{u}_0$  (hyperbolic degeneracy) or not (parabolic degeneracy) (see Shearer et al. [1049], Shearer and Schecter, Isaacson et al. [644, 645], Glimm [531], Freistühler and Pitman [492], Liu and Xin [826], Chen and Kan [288, 289]). In that case, other criteria for selecting the “relevant” solution must be considered (see Remark 5.2). Related problems concern “resonant” systems (Isaacson and Temple [646]). Similar nonuniqueness problems also occur for systems of mixed type, i.e., such that the eigenvalues of the Jacobian  $\mathbf{A}$  become complex in some region, called the elliptic region (see, for instance, Azevedo and Marchesin [69], Hsiao [632]).

The question of nonexistence of weak solutions can also arise for some nonstrictly hyperbolic systems, for which measure solutions (delta-shock waves) are needed (see LeFloch [752], and Tan et al. [1099] and the references therein).  $\square$

## 7 Examples of Systems of Two Equations

For systems of two equations ( $p = 2$ ), the resolution of the Riemann problem can often be well understood by representing the wave curves in the plane of states. It may be the plane of conservative states or, more often, a plane of chosen nonconservative variables which are easier to deal with.

### 7.1 The Case of a Linear or a Linearly Degenerate System

Let us begin by the simple example of a linear system of dimension  $p = 2$ , already seen in Sect. 1 (we use in the following lines the same notations). The wave curves are the discontinuity curves  $S_k(\mathbf{u}_L)$ ,  $k = 1, 2$ , and they are the straight lines  $\mathbf{u}_L + \mathbb{R}\mathbf{r}_k$ . The only intermediate state in the solution of the Riemann problem, say  $\bar{\mathbf{u}}$ , is obtained by intersecting the two lines  $\mathbf{u}_L + \mathbb{R}\mathbf{r}_1$  and  $\mathbf{u}_R + \mathbb{R}\mathbf{r}_2$  in the plane of states  $(u_1, u_2)^T$ .

If we choose to work in the plane of characteristic variables  $\mathbf{v} = \mathbf{T}^{-1}\mathbf{u}$ , with components which we noted  $\mathbf{v} = (\alpha_1, \alpha_2)^T$ , the wave curves are lines parallel to the axis  $\alpha_2 = \alpha_{2,L}$  and  $\alpha_1 = \alpha_{1,R}$ , and  $\bar{\mathbf{v}} = (\alpha_{1,R}, \alpha_{2,L})^T$ . This means equivalently that  $\alpha_2$  is a 1-Riemann invariant and  $\alpha_1$  is a 2-Riemann invariant

Now, consider the particular case of the linear system

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial v}{\partial x} = 0, \\ \frac{\partial v}{\partial t} + a^2 \frac{\partial}{\partial x} u = \frac{1}{\varepsilon} (f(u) - v). \end{cases} \quad (7.1)$$

to which we have added a *relaxation* source term, which is very often used as a semi-linear relaxation system for a scalar conservation law scalar conservation law (see [1040]) and for this reason it is also involved in numerical applications, because it is at the basis of the Jin-Xin relaxation model (we will come over on this example in Chap. IV, Sects. 7.4 and 8.2). If we focus on the left-hand side of (7.1), we have two eigenvalues  $\lambda_1 = -a < \lambda_2 = a$ . The characteristic variables are often noted  $w = v - au, z = v + au$ , where  $z$  (resp.  $w$ ) is a 1-Riemann invariant (resp. a 2-Riemann invariant) and  $w, z$  satisfy the decoupled diagonal system

$$\begin{cases} \frac{\partial w}{\partial t} - a \frac{\partial w}{\partial x} = 0, \\ \frac{\partial z}{\partial t} + a \frac{\partial z}{\partial x} = 0. \end{cases}$$

We have  $u = (z - w)/2a, v = (w + z)/2$ , and the intermediate state in the solution of the Riemann problem in conservative variables is then  $(\bar{u}, \bar{v})^T$  with  $\bar{u} = \frac{1}{2a}(z_L - w_R) = \frac{1}{2a}(v_L - v_R) + \frac{1}{2}(u_L + u_R), \bar{v} = \frac{1}{2}(w_R + z_L) = \frac{1}{2}(v_L + v_R) - \frac{a}{2}(u_R - u_L)$ .

*Remark 7.1.* Let us mention another system which is much studied, more from a theoretical point of view. It is a system with two linearly degenerate fields which writes

$$\begin{cases} \frac{\partial w}{\partial t} + z \frac{\partial w}{\partial x} = 0, \\ \frac{\partial z}{\partial t} + w \frac{\partial z}{\partial x} = 0, \end{cases}$$

(compare with (5.3)) where  $w \in [a, b], z \in [c, d], c < d < 0 < a < b$ . It is written above in a simple diagonal nonconservative form, but admits a conservative formulation. Setting  $u = \frac{1}{w-z}, v = \frac{z}{w-z}$ , one gets

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial v}{\partial x} = 0, \\ \frac{\partial v}{\partial t} + \frac{\partial}{\partial x} \frac{v(1+v)}{u} = 0. \end{cases}$$

In fact, it is in a sense the canonical  $2 \times 2$  strictly hyperbolic system with two linearly degenerate characteristic fields. Indeed, let us consider a  $2 \times 2$  strictly hyperbolic system,  $\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{0}$ , with Riemann invariants, say  $w_1, w_2$ , satisfying  $Dw_i \cdot \mathbf{r}_i = 0$ . Then necessarily, for  $j \neq i$ ,  $Dw_i \cdot \mathbf{r}_j \neq 0$ , and one can normalize the eigenvectors  $\mathbf{r}_i$  so that  $Dw_i \cdot \mathbf{r}_j = \delta_{ij}$  (Kronecker symbol). Then the two linear forms  $Dw_i, i = 1, 2$  correspond to the two “left eigenvectors”  $\mathbf{l}_j^T, j = 2, 1$  and multiplying the system on the left by  $Dw_i, i = 1, 3$ , one gets (for smooth solutions)  $\partial_t w_i + \lambda_j \partial_x w_i = 0, j \neq i$  (as already stated in Example 5.1). Now, if the fields are both linearly degenerate, one can take  $w_i = \lambda_i$ .

Then the Riemann problem is easily solved, following the lines of the linear case. In variables  $(w, z)$ , the intermediate state in the solution is  $(\bar{w}, \bar{z})^T$  with  $\bar{w} = w_R, \bar{z} = z_L$ . For more complete results on the Cauchy problem for measure-valued solution, we refer to [455, 938] and the references therein (in particular [1031]).

Besides, we refer to the book of C. Dafermos (Chapter 12) [384] for a thorough study of general genuinely nonlinear  $2 \times 2$  systems, exploiting the presence of a coordinate system of Riemann invariants mentioned above.  $\square$

## 7.2 The Riemann Problem for the $p$ -System

In the case of the  $p$ -system already introduced in Example 3.1, even though the two fields are genuinely nonlinear, one can also carry out the computations and solve the Riemann problem for almost any two states. Recall that the  $p$ -system is given by (2.7)

$$\begin{cases} \frac{\partial v}{\partial t} - \frac{\partial u}{\partial x} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} p(v) = 0. \end{cases} \quad (2.7)$$

Note that, compared with (7.1), the traditional notations we use for the  $p$ -system exchange the role of  $u$  which is now the second component and  $-v$ , and if one linearizes the pressure  $p(v)$  at a state say  $v_0$ , since we assume  $p'(v_0) < 0$ , setting  $p'(v_0) = -a^2$  we check that system (7.1) (the homogeneous part) is indeed the system obtained by linearization of the  $p$ -system.

If  $p'(v) < 0$ , the two eigenvalues of  $\mathbf{A} = \mathbf{f}'$  are

$$\lambda_1(v) = -\sqrt{-p'(v)} < 0 < \lambda_2(v) = +\sqrt{-p'(v)},$$

and the corresponding eigenvectors  $\mathbf{r}_k, k = 1, 2$  are given by (2.9). Assuming, moreover,  $p'' > 0$ , the two characteristic fields are genuinely nonlinear. Observe that  $\lambda_k$  and  $\mathbf{r}_k$  are functions of  $v$  only.

Let us construct first the curve  $\mathcal{R}_1(\mathbf{w}_L)$  (noted  $\mathcal{R}_1^L$  in the illustrations) of states that can be connected to a given state  $\mathbf{w}_L = (v_L, u_L)$  by a 1-rarefaction

wave. It is the integral curve of  $\mathbf{r}_1$  issued from  $\mathbf{w}_L = (v_L, u_L)$ . By (2.9), we have

$$\begin{cases} v'(\xi) = 1, u'(\xi) = \sqrt{-p'(v(\xi))}, \\ v(\lambda_1(v_L)) = v_L, \quad u(\lambda_1(v_L)) = u_L. \end{cases}$$

Since  $v'(\xi) \neq 0$ , we can take  $v$  as the parameter and we get

$$u'(v) = \sqrt{-p'(v)},$$

and by integrating we obtain

$$u(v) = u_L + \int_{v_L}^v \sqrt{-p'(y)} dy.$$

Due to the assumption  $p'' > 0$ , the condition  $\lambda_1(v) \geq \lambda_1(v_L)$  gives

$$v \geq v_L.$$

The 1-rarefaction curve is thus given in the  $(v, u)$  plane by

$$u(v) = u_L + \int_{v_L}^v \sqrt{-p'(y)} dy, \quad v \geq v_L. \quad (7.2a)$$

It satisfies

$$u'(v) = \sqrt{-p'(v)} > 0, \quad u''(v) = -\frac{p''(v)}{2} \sqrt{-p'(v)} > 0,$$

If  $\mathbf{w}_R = (v_R, u_R)$  is any point on  $\mathcal{R}_1(\mathbf{w}_L)$ , the corresponding solution of the Riemann problem is depicted in Fig. 7.1 (remember that  $\lambda_1 < 0$ ).

In the same way, the 2-rarefaction curve  $\mathcal{R}_2(\mathbf{w}_L)$  (noted  $\mathcal{R}_2^L$  in the illustrations) is given in the  $(v, u)$  plane by

$$u(v) = u_L + \int_v^{v_L} \sqrt{-p'(y)} dy, \quad v \leq v_L. \quad (7.2)$$

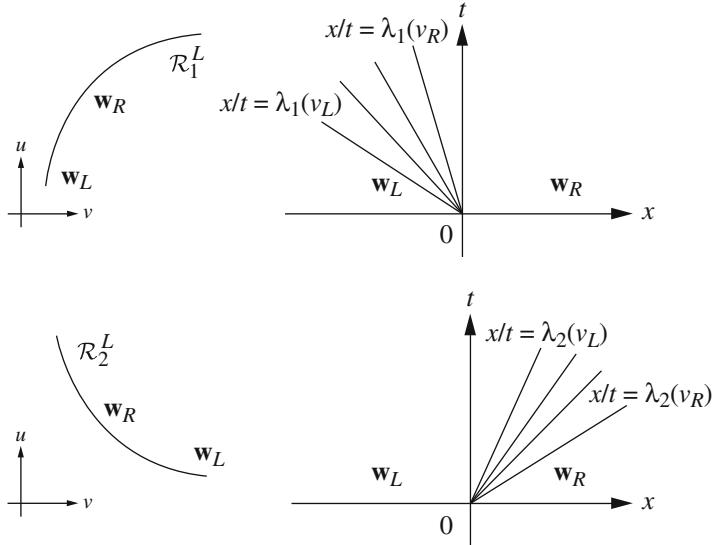
It satisfies

$$u'(v) = -\sqrt{-p'(v)} < 0, \quad u''(v) = \frac{p''(v)}{2} \sqrt{-p'(v)} > 0,$$

and is illustrated in Fig. 7.1 together with the solution of the corresponding Riemann problem associated with  $\mathbf{w}_R = (v_R, u_R) \in \mathcal{R}_2(\mathbf{w}_L)$ .

Let us turn now to the description of the shock curves. The Rankine-Hugoniot jump condition is given in the present case by

$$\begin{cases} \sigma(v - v_L) = -(u - u_L), \\ \sigma(u - u_L) = p(v) - p(v_L). \end{cases} \quad (7.3)$$



**Fig. 7.1** 1- and 2-rarefaction curves

Eliminating  $\sigma$  from these two equations gives

$$(v - v_L)(p(v) - p(v_L)) = -(u - u_L)^2.$$

Since the function  $p$  is decreasing, we get

$$u - u_L = \pm \sqrt{(p(v) - p(v_L))(v_L - v)}. \quad (7.4)$$

Theorem 4.1 enables us to distinguish between 1- and 2-shock curves. Due to (4.5), for a 1-shock, we have

$$\frac{u - u_L}{v - v_L} = +\sqrt{-p'(v_L)} + \mathcal{O}(\varepsilon).$$

Thus

$$u - u_L \text{ is } \begin{cases} \geq 0 & \text{for } v \geq v_L, \\ \leq 0 & \text{for } v \leq v_L, \end{cases}$$

and, from (7.4), we deduce that the 1-shock curve is given by

$$u - u_L = \begin{cases} -\sqrt{(p(v) - p(v_L))(v_L - v)}, & v \leq v_L, \\ +\sqrt{(p(v) - p(v_L))(v_L - v)}, & v \geq v_L. \end{cases} \quad (7.5a)$$

Notice that by (7.3), we have

$$\sigma_1 = -\frac{u - u_L}{v - v_L} < 0.$$

Similarly, for a 2-shock, we have by (4.5)

$$\frac{u - u_L}{v - v_L} = -\sqrt{-p'(v_L)} + \mathcal{O}(\varepsilon).$$

It follows that the 2-shock curve is given by

$$u - u_L = \begin{cases} +\sqrt{(p(v) - p(v_L))(v_L - v)}, & v \leq v_L, \\ -\sqrt{(p(v) - p(v_L))(v_L - v)}, & v \geq v_L \end{cases} \quad (7.5)$$

with

$$\sigma_2 = -\frac{u - u_L}{v - v_L} > 0.$$

Let us now select the admissible shocks, i.e., those that satisfy the Lax entropy conditions (5.14). For the 1-shock (5.14) yields

$$\begin{cases} \sigma_1 < \lambda_1(v_L), \\ \lambda_1(v) < \sigma_1 < \lambda_2(v). \end{cases}$$

Since  $\lambda_1 < 0$  and  $\lambda_2 > 0$ , we have only to impose

$$\lambda_1(v) < \sigma_1 < \lambda_1(v_L).$$

First, since  $p'$  is monotonically increasing,  $\lambda_1(v) < \lambda_1(v_L)$  implies

$$v < v_L.$$

Consider now the following part of the 1-shock curve:

$$u = u_L - \sqrt{(p(v) - p(v_L))(v_L - v)}, \quad v \leq v_L. \quad (7.6)$$

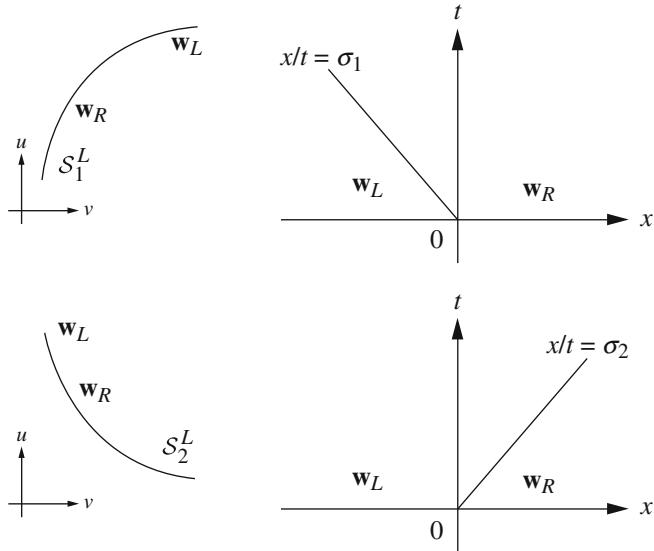
By (7.3), the corresponding shock speed

$$\sigma_1 = -\frac{u - u_L}{v - v_L} = -\sqrt{(p(v) - p(v_L))/(v_L - v)}$$

readily satisfies the Lax condition due to the convexity of  $p$ .

Thus,  $\mathcal{S}_1^a(\mathbf{w}_L)$  (also noted  $\mathcal{S}_1^L$ ) is indeed given by (7.6). We have

$$u'(v) = \frac{p'(v)(v - v_L) + p(v) - p(v_L)}{2\sqrt{(p(v) - p(v_L))(v_L - v)}} > 0 \text{ for } v \leq v_L$$



**Fig. 7.2** 1- and 2-shock curves

(see Fig. 7.2). With a similar analysis, we find that an admissible 2-shock satisfies

$$\begin{aligned}\lambda_2(v) &< \sigma_2, \\ \lambda_1(v_L) &< \sigma_2 < \lambda_2(v_L),\end{aligned}$$

or equivalently

$$\lambda_2(v) < \sigma_2 < \lambda_2(v_L),$$

which implies  $v \geq v_L$ . The curve  $\mathcal{S}_2^a(\mathbf{w}_L)$  (also noted  $\mathcal{S}_2^L$ ) is thus

$$u = u_L - \sqrt{(p(v) - p(v_L))(v_L - v)}, \quad v \geq v_L. \quad (7.7)$$

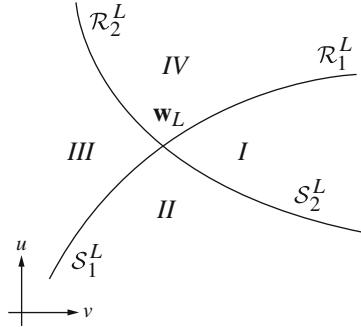
For a given right state  $\mathbf{w}_R = (v_R, u_R)$ , the speed of the shock connecting  $\mathbf{w}_L = (v_L, u_L)$  and  $\mathbf{w}_R = (v_R, u_R)$  is

$$\sigma_2 = -\frac{u_R - u_L}{v_R - v_L} = \sqrt{(p(v_L) - p(v_R))/(v_R - v_L)}.$$

*Remark 7.2.* We shall now see that the Lax entropy condition (5.14) is in the present case equivalent to the entropy condition (5.18) associated with the entropy pair  $(U, F)$  introduced in Chap. I, Example 3.1

$$U(v, u) = \frac{u^2}{2} - P(v), \quad F(v, u) = p(v)u,$$

where  $P$  is a primitive of  $p$ .



**Fig. 7.3** Regions delimited by the rarefaction and shock curves

Indeed, with this choice of  $(U, F)$ , (5.18) yields

$$\sigma \left\{ \frac{u^2 - u_L^2}{2} - \int_{v_L}^v p(y) dy \right\} \geq up(v) - u_L p(v_L).$$

Using (7.3), we can eliminate  $\sigma$  and get

$$\frac{1}{2}(u + u_L)(p(v) - p(v_L)) + \left( \frac{u - u_L}{v - v_L} \right) \int_{v_L}^v p(y) dy \geq up(v) - u_L p(v_L),$$

or equivalently

$$(u - u_L) \left\{ \frac{1}{(v - v_L)} \int_{v_L}^v p(y) dy - \frac{1}{2}(p(v) + p(v_L)) \right\} \geq 0.$$

Since, by the convexity of  $p$ , the term in brackets is always  $\leq 0$ , we obtain the condition

$$u \leq u_L.$$

Compared with (7.6), (7.7), we find the same admissible shock curves.  $\square$

Let us draw the curves  $S_i^a(\mathbf{w}_L)$  and  $\mathcal{R}_i(\mathbf{w}_L)$ ,  $i = 1, 2$  together in the  $(v, u)$  plane. In conformity with the general case (see (6.2)), the curves  $S_i^a$  and  $\mathcal{R}_i$ ,  $i = 1, 2$  are osculatory so that we obtain Fig. 7.3.

The curves above divide the  $(v, u)$ -plane into four regions (for a given left state  $(v_L, u_L,)$ ) labeled  $I$  to  $IV$ .

By Theorem 6.1, we can solve locally the Riemann problem for any right state  $\mathbf{w}_R$  sufficiently close to  $\mathbf{w}_L = (v_L, u_L)$ . The solution consists of at most three constant states (including  $(v_L, u_L)$  and  $(v_R, u_R)$ ) separated by a  $k$ -rarefaction or a  $k$ -admissible shock wave  $k = 1, 2$ . In fact, this result has a global character and holds for right states that are not necessarily in a neighborhood of  $(v_L, u_L)$ .

*Theorem 7.1*

- (i) If  $\mathbf{w}_R = (v_R, u_R)$  belongs to the regions I or III (see Fig. 7.3) the Riemann problem (6.1) for the  $p$ -system always admits a solution.
- (ii) If  $\mathbf{w}_R = (v_R, u_R) \in II$ , the Riemann problem always has a solution provided that  $p$  is defined on the whole real line or  $p$  is defined on  $]0, +\infty[$  with  $\lim_{v \rightarrow 0+} p(v) = +\infty$ .
- (iii) If  $\mathbf{w}_R = (v_R, u_R) \in IV$ , the Riemann problem may have no solution.

*Proof.* Let us consider the four different cases:

*Case 1:*  $(v_R, u_R) = \mathbf{w}_R \in I$  (see Fig. 7.4).

Let us prove that there exists a state  $\bar{\mathbf{w}}_0 = (\bar{v}_0, \bar{u}_0)$  on  $\mathcal{R}_1(w_L)$  such that  $\mathbf{w}_R$  can be connected to  $\bar{\mathbf{w}}_0$  by a 2-shock, i.e.,  $\mathbf{w}_R \in \mathcal{S}_2^a(\bar{\mathbf{w}}_0)$ .

We begin by characterizing the states  $\mathbf{w} = (v, u)$  that can be connected to  $\mathbf{w}_R$  by a 1-rarefaction and a 2-shock by means of an intermediate state  $\bar{\mathbf{w}} = (\bar{v}, \bar{u}) \in \mathcal{R}_1(\mathbf{w}_L)$ . First, we have by (7.1)

$$\bar{u} = u_L + \int_{v_L}^{\bar{v}} \sqrt{-p'(y)} dy, \quad \bar{v} > v_L.$$

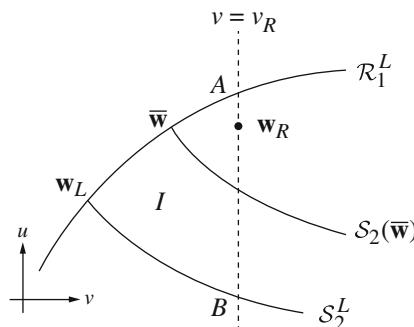
Next, by (7.7),  $\mathcal{S}_2^a(\bar{\mathbf{w}})$  is the set of states  $\mathbf{w} = (v, u)$  that satisfy

$$u = \bar{u} - \sqrt{(p(v) - p(\bar{v}))(\bar{v} - v)}, \quad v > \bar{v}$$

so that  $\mathbf{w}$  and  $\bar{\mathbf{w}}$  are linked by the relation

$$u = u_L + \int_{v_L}^{\bar{v}} \sqrt{-p'(y)} dy - \sqrt{(p(v) - p(\bar{v}))(\bar{v} - v)}, \quad v > \bar{v} > v_L.$$

Now  $\mathcal{S}_2^a(\bar{\mathbf{w}})$  intersects the vertical line  $v = v_R$  at a point  $(v_R, G(\bar{v})) = \mathbf{w}_R$ , where  $G$  is defined by



**Fig. 7.4** Region I

$$G(v) = u_L + \int_{v_L}^v \sqrt{-p'(y)} dy - \sqrt{(p(v) - p(v_R))(v_R - v)}, \quad v_R > v > v_L.$$

In order to prove that there exists a unique  $\bar{v}_0$  such that  $G(\bar{v}_0) = u_R$ , hence such that  $\bar{\mathbf{w}}_R = \mathbf{w}_R$ , we first check that  $G$  is monotonically increasing. Indeed

$$G'(v) = \sqrt{-p'(v)} - \frac{\{p(v_R) - p(v) - p'(v)(v - v_R)\}}{2\sqrt{(p(v) - p(v_R))(v_R - v)}} > 0.$$

Then clearly the line  $v = v_R$  intersects  $\mathcal{R}_1(w_L)$  at a point  $A = (v_R, u_A)$  and  $\mathcal{S}_2^a(\mathbf{w}_L)$  at a point  $B = (v_R, u_B)$  with

$$u_A = G(v_R) < u_R < u_B = G(v_L).$$

Hence, there exists a unique  $\bar{v}_0 \in ]v_L, v_R[$  such that  $G(\bar{v}_0) = u_R$ . Setting

$$\bar{u}_0 = u_L + \int_{v_L}^{\bar{v}_0} \sqrt{-p'(y)} dy,$$

we get the point  $\mathbf{w}_0 = (\bar{v}_0, \bar{u}_0)$  that we were looking for.  $\square$

In short,  $\mathbf{w}_L$ , can be connected to  $\mathbf{w}_R$  by a 1-rarefaction wave followed by a 2-shock propagating with speed  $\sigma_2 = \sqrt{\frac{(p(v_R) - p(\bar{v}_0))}{(\bar{v}_0 - v_R)}}$  (see Fig. 7.5).

*Remark 7.3.* Note that it is impossible in the preceding case to connect  $\mathbf{w}_L$  and  $\mathbf{w}_R$  by a 2-shock wave and a 1-rarefaction in this order (as Fig. 7.4 might suggest), since the condition  $\lambda_1 < 0 < \lambda_2$  cannot be violated.  $\square$

*Case 2:  $\mathbf{w}_R \in III$ .*

By a similar argument, we can prove that  $\mathbf{w}_L$  can be connected to  $\mathbf{w}_R$  by a 1-shock followed by a 2-rarefaction wave (see Fig. 7.6). The 1-shock propagates with speed  $\sigma_1 = \sqrt{\frac{(p(\bar{v}_0) - p(v_L))}{(v_L - \bar{v}_0)}}$ .

*Case 3:  $\mathbf{w}_R \in II$ .*

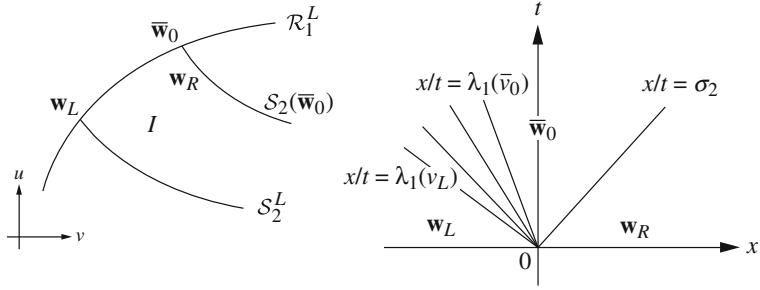
Let us check that the horizontal line  $u = u_R$  meets  $\mathcal{S}_1^a(\mathbf{w}_L)$  and  $\mathcal{S}_2^a(\mathbf{w}_L)$ , respectively, at two points  $A, B$  that are uniquely determined. To prove the existence and uniqueness of  $B$ , we have to solve by (7.7)

$$u_L - \sqrt{(p(v) - p(v_L))(v_L - v)} = u_R, \quad v \geq v_L. \quad (7.8)$$

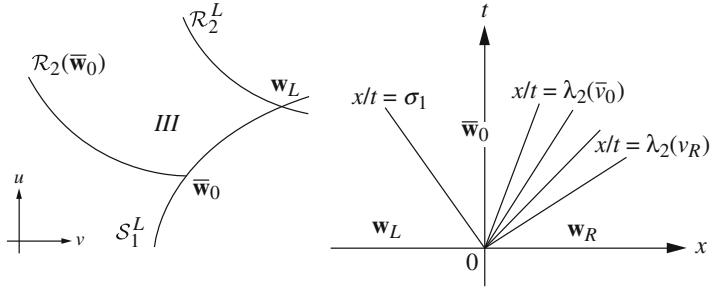
But the function

$$v \rightarrow u_L - \sqrt{(p(v) - p(v_L))(v_L - v)}$$

is easily seen to decrease for  $v \in [v_L, +\infty[$  from  $u_L$  to  $-\infty$ . Hence, since  $u_R < u_L$ , there exists a unique  $v = v_B$  solution of (7.8).



**Fig. 7.5** Solution of the Riemann problem: 1-rarefaction and 2-shock



**Fig. 7.6** Solution of the Riemann problem: 1-shock and 2-rarefaction

In the same way, the point where the line  $u = u_R$  meets  $\mathcal{S}_1^a(\mathbf{w}_L)$  satisfies by (7.6)

$$u_L - \sqrt{(p(v) - p(v_L))(v_L - v)} = u_R, \quad v \leq v_L. \quad (7.9)$$

The mapping  $v \mapsto (p(v) - p(v_L))(v_L - v)$  is decreasing for  $v \leq v_L$ . Moreover, on the one hand, if  $p$  is defined on  $\mathbb{R}$ , we have

$$\lim_{v \rightarrow -\infty} (p(v) - p(v_L))(v_L - v) = +\infty,$$

and, on the other hand, if  $p$  is only defined on  $]0, +\infty[$  we have by assumption

$$\lim_{v \rightarrow 0^+} (p(v) - p(v_L))(v_L - v) = +\infty.$$

Thus, in both cases, we can find a unique  $v = v_A < v_L$  solution of (7.9).

Then, we can conclude as in case 1 that there exists a unique state  $\bar{\mathbf{w}}_0 \in \mathcal{S}_a^1(\mathbf{w}_L)$  such that  $\mathbf{w}_R$  belongs to  $\mathcal{S}_2^a(\bar{\mathbf{w}}_0)$  (see Fig. 7.7). Hence, we can connect  $\mathbf{w}_L$  to  $\mathbf{w}_R$  by a 1-shock followed by a 2-shock.

*Case 4:  $\mathbf{w}_R \in IV$ .*

In this case, the Riemann problem may have no solution. Indeed, assume

$$u_\infty = \int_{v_L}^{+\infty} \sqrt{-p'(y)} dy < +\infty. \quad (7.10)$$

Then  $\mathcal{R}_1(u_L)$  has a horizontal asymptote, namely, the line  $u = u_L + u_\infty$ . Now take, for instance, a state  $\mathbf{w}_R$  in IV with

$$v_R = v_L \text{ and } u_R > u_L + 2u_\infty$$

(see Fig. 7.8).

Then, for any  $\bar{\mathbf{w}} \in \mathcal{R}_1(\mathbf{w}_L)$ ,

$$\bar{u} = u_L + \int_{v_L}^{\bar{u}} \sqrt{-p'(y)} dy < u_L + u_\infty, \quad \bar{v} \geq v_L.$$

The 2-rarefaction curve passing by  $\bar{\mathbf{w}}$  is given by

$$u = \bar{u} + \int_v^{\bar{v}} \sqrt{-p'(y)} dy, \quad v \leq \bar{v}$$

or

$$u = u_L + \int_{v_L}^{\bar{v}} \sqrt{-p'(y)} dy + \int_v^{\bar{v}} \sqrt{-p'(y)} dy.$$

It meets the vertical line  $v = v_L$  at a point  $(v_L, u)$  with

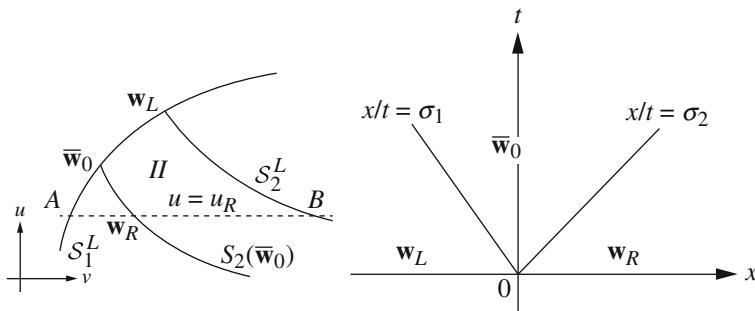
$$u = u_L + 2 \int_{v_L}^{\bar{v}} \sqrt{-p'(y)} dy \leq u_L + 2u_\infty,$$

which proves that the point  $\mathbf{w}_R$  cannot be reached by a 2-rarefaction wave (nor by a 2-shock of course!).

When the Riemann problem can be solved, which occurs at least for  $\mathbf{w}_R$  close to  $\mathbf{w}_L$ , one can find  $\bar{\mathbf{w}}_0 \in \mathcal{R}_1(\mathbf{w}_L)$  such that  $\mathbf{w}_R \in \mathcal{R}_2(\bar{\mathbf{w}}_0)$  and the solution consists of two rarefaction waves (Fig. 7.9).

*Example 7.1. The vacuum problem.*

To illustrate the situation in case 4, take  $p(v) = Av^{-\gamma}$  with  $\gamma > 1$ , so that (7.10) holds. Also, we notice that as  $v \rightarrow +\infty$



**Fig. 7.7** Solution of the Riemann problem: 1-shock and 2-shock

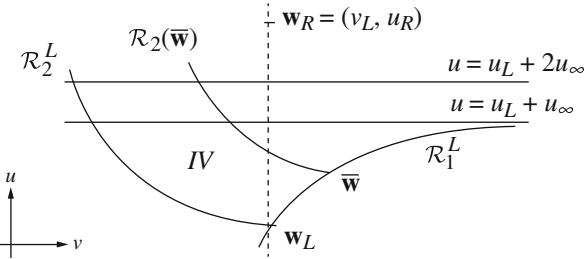
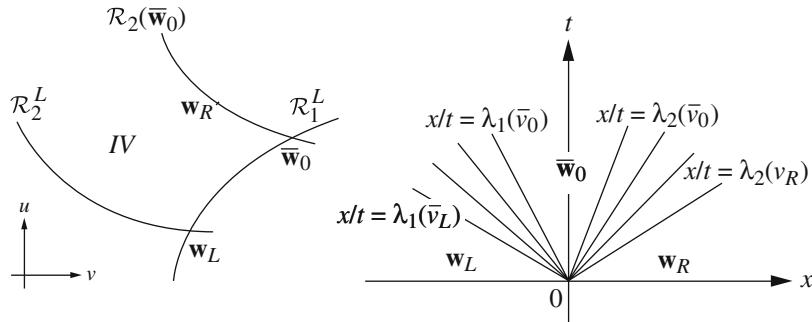
Fig. 7.8 Region  $IV$ 

Fig. 7.9 Solution of the Riemann problem: 1-and 2-rarefaction

$$\lambda_i(v) = \pm \sqrt{-p'(v)} \rightarrow 0, \quad i = 1, 2.$$

To understand what happens, take the limit situation  $\mathbf{w}_R = (v_L, u_L + 2u_\infty)$ . Then  $\mathbf{w}_R$  belongs to  $\mathcal{R}_2(\bar{\mathbf{w}}_\infty)$ , where  $\bar{\mathbf{w}}_\infty = (+\infty, u_L + u_\infty)$  and the solution is given by two complete rarefaction waves (as depicted in Fig. 7.10), which means that the fan of the 1-(resp. the 2-)rarefaction wave is bordered on the right (resp. the left) by the line  $x = 0$ . On this line, we have  $v = +\infty$ , which implies the formation of a vacuum since  $\rho = \frac{1}{v} = 0$  (in this case, [1213] allows the specific volume to be a Radon measure).  $\square$

*Remark 7.4.* It is possible to extend the analysis to a nonconvex equation of state (we refer to Wendroff [1185]) provided composite waves are considered to solve the Riemann problem. We can also consider the case of a van der Waals-type equation of state that takes into account phase transition in a mixture of fluids, complemented with Maxwell law. Assume, for instance, that the two phases are ideal gases with an isothermal law; the pressure law of the mixture writes

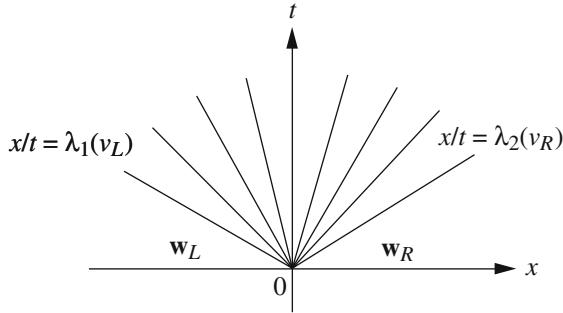


Fig. 7.10 Vacuum

$$p(v) = \begin{cases} a_2/v, & \text{if } 0 < v < v_2^* \\ a_2/v_2^*, & \text{if } v_2^* < v < v_1^* \\ a_1/v, & \text{if } v > v_1^*, \end{cases}$$

where  $a_1 > a_2$  are two positive constants and  $a_2/v_2^* = a_1/v_1^*$ . The pressure law is constant in the mixture zone, and we are left with the well-known system of pressureless gases (see [161]). However, the zone  $(v_2^*, v_1^*)$  where  $p'$  vanishes is bounded, and thus it is possible to exhibit a solution of the Riemann problem involving multiple waves but without measure (we refer to [541] for details).  $\square$

### 7.3 The Riemann Problem for the Barotropic Euler System

We can similarly solve the Riemann problem for the barotropic Euler system in Eulerian coordinates

$$\begin{cases} \frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \end{cases} \quad (7.11)$$

where the barotropic pressure law  $p = p(\rho)$  satisfies:  $p(\rho) > 0$ ,  $p'(\rho) > 0$ ,  $p''(\rho) + \frac{2}{\rho}p'(\rho) > 0$ .

### 7.3.1 The Barotropic Euler System

Let us check that the barotropic assumption in the Euler system (2.35) leads to (7.11) with an energy equation which is redundant for smooth solutions. Indeed, one can easily prove that (for smooth solutions) the energy conservation results from the conservation of mass and momentum equations. The momentum equation leads to

$$\partial_t u + u \partial_x u + \frac{1}{\rho} \partial_x p = 0,$$

as in (2.39), while the mass conservation equation multiplied by  $\varepsilon'(\rho)$  gives

$$\partial_t \varepsilon(\rho) + u \partial_x \varepsilon(\rho) + \rho^2 \varepsilon'(\rho) \partial_x u = 0.$$

Thus, setting  $\varepsilon'(\rho) = p(\rho)/\rho^2$  in order to define the internal energy  $\varepsilon$ , we get easily the energy conservation equation with the same definition  $e = \frac{u^2}{2} + \varepsilon$ .

Note that with this definition of  $\varepsilon$ , the second law of thermodynamics (2.27) gives  $ds = 0$ , and the flow is isentropic.

Then many results obtained for the full Euler system may be simplified to get the formulas for the barotropic case, for instance, the speed of sound is  $c^2 = p'(\rho)$ , the formulas for the eigenvectors may be derived from (2.43), and the 1- and 2- Riemann invariants (cf. Example 3.3) are respectively  $u + \ell$  and  $u - \ell$  with  $\ell'(\rho) = c/\rho$ .

### 7.3.2 Correspondence with the $p$ -System

As we have already seen, the barotropic Euler system in Lagrangian coordinates coincides with the  $p$ -system, provided we use the notation  $v$  instead of  $\tau$  for  $\frac{1}{\rho}$  and restrict to  $\rho > 0$ . The assumptions on  $p$  are strictly equivalent to the assumptions made on the function  $p(v)$  (for the  $p$ -system, they ensure that the two fields are genuinely nonlinear). Indeed, setting now  $\tilde{p}(v) = p(\rho)$ , where  $v = \frac{1}{\rho}$ , we have  $\tilde{p}'(v) = -p'(\rho)/v^2$ , and  $\tilde{p}''(v) = (p''(\rho) + 2vp'(\rho))/v^4$ .

Thus, the general result of Theorem 2.2 concerning the change of frame gives the definition of the energy (which is a mathematical entropy). From the formulas recalled in Remark 7.2:  $U = u^2/2 - P(v)$ ,  $F = pu$ , we get the corresponding entropy-entropy flux pair for system (7.11):  $\mathcal{U} = \rho U = \rho e$ ,  $\mathcal{F} = \rho u U + F = (\rho e + p)u$ .

Then the structure of the Riemann problem is the same as for the  $p$ -system. Note that the waves may now propagate to the left or to the right, according to the sign of the eigenvalues  $u \pm c$ , where  $c^2 = p'(\rho)$ . The flow is subsonic if  $|u|/c < 1$  (supersonic if  $|u|/c > 1$ ).

The Riemann invariants are the same, as we have already proved (see Remark 3.2), given by  $u \pm \ell$ , with  $\ell'(\rho) = c/\rho$ . Thus, if we draw the rarefaction curves in the  $(\tau, u)$  plane instead of the  $(\rho, u)$ -plane, they coincide. The shock curves also coincide in the  $(\tau, u)$ -plane. We can prove it in the general framework of fluid systems, using the notations (2.16) of Sect. 2.1 above.

First, writing the Rankine-Hugoniot conditions in Eulerian coordinates for a shock propagating with speed  $\sigma^{\mathcal{E}}$ ,  $[\mathbf{F}(\mathbf{U})] = \sigma^{\mathcal{E}}[\mathbf{U}]$ , gives

$$\begin{cases} [\rho u] = \sigma^{\mathcal{E}}[\rho], \\ [\rho u \Phi + \mathbf{g}] = \sigma^{\mathcal{E}}[\rho \Phi] \end{cases}$$

If we eliminate  $\sigma^{\mathcal{E}}$ , we get  $[\rho u][\rho \Phi] = [\rho][\rho u \Phi + \mathbf{g}]$  from which we deduce easily the relation  $[\Phi][u] = -[\frac{1}{\rho}][\mathbf{g}]$ . This relation says that the discontinuity curves also coincide in the  $(\tau, u)$  plane. Indeed, writing the Rankine-Hugoniot conditions now in Lagrangian coordinates,  $[\mathbf{G}(\mathbf{V})] = \sigma^{\mathcal{L}}[\mathbf{V}]$ , gives

$$\begin{cases} [u] = -\sigma^{\mathcal{L}}[\tau], \\ [\mathbf{g}] = \sigma^{\mathcal{L}}[\Phi] \end{cases}$$

which by eliminating  $\sigma^{\mathcal{L}}$ , leads to the same relation. The entropy condition gives naturally the same admissible parts of the discontinuity curves.

For an isothermal flow,  $p(\rho) = c^2\rho$  is linear, and the computations are particularly simple. The rarefaction curves  $\mathcal{R}_i(\mathbf{u}_L)$ ,  $i = 1, 2$  are given by

$$u - u_L = \begin{cases} c \log(\frac{\tau}{\tau_L}), & \tau \geq \tau_L, \quad 1\text{-rarefaction} \\ -c \log(\frac{\tau}{\tau_L}), & \tau \leq \tau_L, \quad 2\text{-rarefaction} \end{cases}$$

and the shock curves  $\mathcal{S}_i(\mathbf{u}_L)$ ,  $i = 1, 2$  are given by

$$u - u_L = \begin{cases} -c \frac{\tau_L - \tau}{\sqrt{\tau\tau_L}}, & \tau \leq \tau_L, \quad 1\text{-shock} \\ -c \frac{\tau - \tau_L}{\sqrt{\tau\tau_L}}, & \tau \geq \tau_L, \quad 2\text{-shock} \end{cases}$$

with shocks propagating at the respective velocities:  $\sigma_1^{\mathcal{E}} = u_L - c\sqrt{\tau_L/\tau} = u - c\sqrt{\tau/\tau_L}$  for the 1-shock, which can also be written  $\sigma_1^{\mathcal{E}} = u_L - c\tau_L\sqrt{1/\tau_L\tau} = u - c\tau\sqrt{1/\tau\tau_L}$ , and  $\sigma_2^{\mathcal{E}} = u_L + c\sqrt{\tau_L/\tau} = u + c\sqrt{\tau/\tau_L} = u + c$  for the 2-shock. The corresponding Lagrangian speeds are  $\sigma_i^{\mathcal{L}} = \pm c\sqrt{1/\tau\tau_L}$ .

We have directly analog formulas expressed in terms of  $\rho$ .

*Remark 7.5.* As already observed in the preceding remark, the solution of the Riemann problem in the nonconvex case (when the function  $\tilde{p}$  is not necessarily convex) is studied in [1184, 1185] (see also [1047]). The Lax entropy

condition has to be replaced by the Liu entropy condition [822]. The solution of the Riemann problem then involves composite waves. Let us also mention [541] where a simple model for phase transition involves a nonconvex pressure law exhibiting a zone where the pressure is constant, thus leading to a resonance phenomena. The Riemann problem may however be solved without measure, contrary to the case of zero pressure gas dynamics, where Dirac measures have to be introduced in order to solve the Riemann problem (see [161, 1052]).  $\square$

### 7.3.3 The Saint-Venant System

Let us now consider the case of the Saint-Venant system (see Example 3.1, system (3.2) in the Chap. I),

$$\begin{cases} \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0, \\ \frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}(hu^2 + \frac{g}{2}h^2) = -ghZ'(x), \end{cases} \quad (7.12)$$

which corresponds to a barotropic Euler system with a quadratic pressure law  $p(h) = \frac{g}{2}h^2$ . Then, the analog of  $\varepsilon$  is simply  $\frac{g}{2}h$  (see Sect. 7.3.1), and the energy is given by  $E(h, hu) = \frac{1}{2}(hu^2 + gh^2)$ . From this correspondence, and multiplying on the left system (7.12) by  $E'(h, hu)$ , we get that smooth solutions of (7.12) satisfy

$$\frac{\partial}{\partial t}\left(\frac{1}{2}(hu^2 + gh^2)\right) + \frac{\partial}{\partial x}\left(u\left(\frac{1}{2}hu^2 + gh^2\right)\right) = -ghuZ'(x),$$

replaced by an inequality for admissible discontinuous solution. Now, this nonconservative equation leads to a conservative one, related to the modified energy, now involving  $Z$ ,

$$E(h, hu, Z) = \frac{1}{2}(hu^2 + gh^2) + gZh.$$

Indeed, smooth solutions of (7.12) are easily shown to satisfy the energy conservation equation (exhibited in Example 3.1, Chap. I)

$$\partial_t E + \operatorname{div}\left(u\left(E + \frac{g}{2}h^2\right)\right) = 0.$$

The solution of the Riemann problem in the case where the topography is constant (flat bottom  $Z' = 0$ ) is obtained following the above lines for the barotropic Euler system. The analog of the sound speed is  $\sqrt{gh}$ , and the eigenvalues are thus  $u \pm \sqrt{gh}$ . The Froude number is defined as  $|u|/\sqrt{gh}$ ;

when  $Fr > 1$  (resp.  $Fr < 1$ ), the flow is called supercritical or torrential (resp. subcritical or fluvial) and critical when  $Fr = 1$ .

The Riemann invariants (analog of  $u \pm \ell$ ) are  $w_{\pm} = u \pm 2\sqrt{gh}$ .

In the general case ( $Z$  nonconstant), the solution of the Riemann problem for (7.12) is no longer self similar and has to take into account the topography. In Chap. VII (Sect. 2.4), introducing the topography  $Z$  as a new dependent variable satisfying  $\partial_t Z = 0$ , we will consider the Riemann problem for the system in  $(h, hu, Z)$ , which means for piecewise constant, discontinuous topography. This approach is used in the design of numerical schemes, and the subject will be developed more generally for systems with a geometric source term.

## Notes

We refer naturally to the important paper of Lax [745], to the book of Smoller [1066], and particularly to the very complete one of Serre [1038] (or its translated version [1039]); see also the illuminating introductory notes of Tartar (Instituto de Analisi del CNR, Pavia [1108]); the papers of DiPerna [422, 423], Dafermos [381], and Liu [824]; the books of LeVeque [774, 777], Chang and Hsiao [278], and Li Ta-tsien [794].

In Sect. 6, Theorem 6.1 ensures the existence of an entropy solution to the Riemann problem for nearby states; note that the solution may not exist for general Riemann data (Kranzer and Keyfitz [708], Serre [1035] and the references therein, and Keyfitz and Kranzer [692]). The existence for more general small BV Cauchy data (near a constant state) is proven by Glimm [530] and Schochet [1019] for small BV perturbation of a solvable Riemann problem (see the references therein for other results and also Temple and Young [1113]) via the convergence of the random-choice method. The existence of solutions in the general case is mostly an open problem. For smooth local solutions, see Kato [682], and for the use of convergent finite difference approximations, see Nishida and Smoller [902]. In some cases, global existence results are obtained by the vanishing viscosity method together with compensated compactness (see Tartar [1106]), for instance, for isentropic gas dynamics by DiPerna [424], for the class of B. Temple systems [1112] by LeVeque and Temple and Serre [1034], and for “rich” systems by Serre [1036]; see also Chen [283], Heibig [598, 599], Liu [823], Rascle [966], Peng [937], Benzoni-Gavage and Serre [111], and Rubino [993, 994]. An alternative technique is wave front tracking (Bressan and Colombo [200], and the references therein). We also mention a result by Freistühler [491] concerning a class of (nonstrictly) hyperbolic systems and related results to which he refers and the papers of Isaacson and Temple [646] concerning “resonant” systems. For the question of uniqueness, see LeFloch and Xin [760], which contains a survey of the known results, and the references therein.

In the isentropic case, the kinetic formulation provides important results [811], and we refer to B. Perthame's textbook [944].

We have considered “convex” systems, i.e., strictly hyperbolic systems whose characteristic fields are either genuinely nonlinear or linearly degenerate. The existence of a solution to the Riemann problem for more general (nonconvex) systems is given in Liu [821]; see also Menikoff and Plohr [862] and the interesting paper of Wendroff [1184] concerning the  $p$ -system when  $p$  is not convex. As in the scalar case, we must admit composite waves corresponding to a characteristic field that is non-genuinely nonlinear (see Remarks 6.2 and 7.4); otherwise, the conclusions of Theorem 6.1 are valid.

For the extension of these results to nonstrictly hyperbolic systems, the proof of Theorem 6.1 obviously fails since the rank of  $\mathbf{M}(\mathbf{u}_0)$  is  $< p - 1$  at an “umbilic point” (coinciding eigenvalues; see Remark 6.1), which poses a bifurcation problem (see Shearer et al. [1049], and Chen and Kan [288]); see also [1223]). Besides the references at the end of Chapter 17 of Smoller's book [1066], see Keyfitz and Kranzer [691], Tveito and Winther [1140], the papers of Keyfitz [686, 687] in the above mentioned AMS proceedings Keyfitz and Kranzer Eds. [691], and those of Saint-Etienne et al. [232] and also [688, 690], those of Brio [204], Freistühler [489], and Glimm [531] in the Proceedings of the 2nd International conference on nonlinear hyperbolic problems (Ballmann and Jeltsch Eds. [75]), those of Azevedo and Marchesin [69], Hsiao [633], Liu and Xin [826], LeFloch et al. [1050], and Schaeffer and Shearer [1013], in *Nonlinear evolution equations that change type*, Keyfitz and Shearer [694], and Lindquist [808]. All those proceedings contain other interesting related papers; see also Smoller [1065].

We did not consider nonclassical shock waves, for which we refer to P.G. LeFloch's textbook [753].

For what concerns “convex” hyperbolic systems in nonconservative form, which are introduced in Remark 5.2, the existence of weak entropy BV solutions is proven in LeFloch and Liu [755], for small BV data, by means of the convergence of Glimm's scheme; see also [1223].

Though it is very important in the applications, we did not mention the problem of invariant regions for system (2.1) (a subset  $S$  of the set of states is invariant if  $(\mathbf{u}_0(x) \in S, \forall x \Rightarrow \mathbf{u}(x, t) \in S, \forall (x, t))$ ; their characterization can be found in Chueh et al. [307], Hoff [620], and Serre [1032].

We did not consider systems with source terms. However, the case of reacting flows will be studied in the next chapter. Some other examples can be found in LeVeque and Yee [786], LeVeque and Wang [785], E [454], Fan and Hale [468], Chen et al. [291], Isaacson and Temple [647], Kevorkian et al. [685], and Chen and Glimm [286], Klingenberg and Lu [702], and Schroll et al. [1023] in Glimm et al. [532]. We will come over the subject in Chap. VII, devoted to some aspects of systems with source terms and their numerical treatment.

*Note Added in the Second Edition*

Some interesting subjects would have deserved more development, such as measure-valued solutions [504, 1082, 1083], and more recently [485], duality solutions [171], delta-shock waves [388, 1099], renormalized entropy solutions [109], after [43, 110], when the flux is less regular; also in link with linearization, and references therein, singular shocks [692], Dafermos regularization [380] [805, 1014, 1016, 1143]. Then let us mention interesting results concerning the Riemann problem for the isentropic system of gas dynamics in two space dimensions in [302] (system of three equations, with initial data piecewise constant on each side of  $x_2 = 0$ : existence and uniqueness of self-similar solutions, functions of  $x_2/t$ ; nonuniqueness of more general solutions). In the applications modeled by  $2 \times 2$  systems, we can add traffic models, such as the Aw-Rascle model [68], and more references can be found in Chap. VII. As already written the above list is nonexhaustive.



# III

## Gas Dynamics and Reacting Flows

### 1 Preliminaries

#### 1.1 Properties of the Physical Entropy

Let us consider a fluid in a local thermodynamical equilibrium. Then we know from thermodynamics that the thermodynamical state of the fluid is completely determined by any two thermodynamic variables. Most often, we shall note by the same letter the corresponding mathematical functions, though they differ. For instance, we shall use

$$p = p(\tau, s) = p(\rho, \varepsilon), \quad \varepsilon = \varepsilon(\tau, s) = \varepsilon(\tau, p), \quad T = T(\tau, s) = T(\tau, p), \dots$$

Let us choose in particular the specific volume  $\tau$  and the specific (physical) entropy  $s$  and concentrate on the function  $(\tau, s) \mapsto \varepsilon(\tau, s)$ . By the second law of thermodynamics, we have

$$d\varepsilon = T ds - p d\tau \tag{1.1}$$

so that

$$\frac{\partial \varepsilon}{\partial \tau} = -p < 0, \quad \frac{\partial \varepsilon}{\partial s} = T > 0.$$

Moreover, always assuming that the fluid is in a local thermodynamic equilibrium, we have the following important assumption:

$$\text{the function } (\tau, s) \mapsto \varepsilon(\tau, s) \quad \text{is strictly convex.} \tag{1.2}$$

This means that the Hessian matrix of this function,

$$\begin{pmatrix} \frac{\partial^2 \varepsilon}{\partial \tau^2} & \frac{\partial^2 \varepsilon}{\partial \tau \partial s} \\ \frac{\partial^2 \varepsilon}{\partial \tau \partial s} & \frac{\partial^2 \varepsilon}{\partial s^2} \end{pmatrix},$$

is a (symmetric) positive-definite matrix, or equivalently that we have

$$\frac{\partial^2 \varepsilon}{\partial s^2} > 0, \quad \frac{\partial^2 \varepsilon}{\partial \tau^2} \frac{\partial^2 \varepsilon}{\partial s^2} - \left( \frac{\partial^2 \varepsilon}{\partial \tau \partial s} \right)^2 > 0. \quad (1.3)$$

As a consequence, we obtain

$$\frac{\partial p}{\partial \tau} = -\frac{\partial^2 \varepsilon}{\partial \tau^2} < 0,$$

which enables us to define the sound speed  $c$  by

$$c = \tau \sqrt{-\partial p / \partial \tau}. \quad (1.4)$$

Now, since  $\frac{\partial \varepsilon}{\partial s} > 0$ , one can invert the function  $s \mapsto \varepsilon(\tau, s)$  and introduce the function  $(\tau, s) \mapsto s(\tau, \varepsilon)$ . The purpose of this section is to prove the following result.

### Theorem 1.1

*The three following assertions are equivalent:*

- (i) *the function  $(\tau, s) \mapsto \varepsilon(\tau, s)$  is strictly convex;*
- (ii) *the function  $(\tau, \mathbf{u} = (u_1, \dots, u_d), e) \mapsto -s(\tau, e - \frac{|\mathbf{u}|^2}{2})$  is strictly convex;*
- (iii) *the function  $(\rho, \mathbf{q} = (q_1, \dots, q_d), E) \mapsto -\rho s(\frac{1}{\rho}, \frac{E}{\rho} - \frac{|\mathbf{q}|^2}{2\rho^2})$  is strictly convex.*

The proof of this result needs a few simple steps, all of which rely on the differentiation of a compound function. Let us recall that if  $f$  and  $g$  are two  $C^2$  functions defined on an open subset  $U$  (respectively  $V$ ) of a Banach space  $E$  (respectively  $F$ ) with  $f : U \mapsto V$  and  $g : V \mapsto G$  ( $G$  a Banach space), we have the following formula expressing the second total derivative of  $g \circ f$ : for all  $a \in U, (x_1, x_2) \in E \times E$

$$\begin{cases} D^2(g \circ f)(a)(x_1, x_2) \\ = Dg(f(a)) \cdot (D^2f(a) \cdot (x_1, x_2)) \\ + D^2g(f(a)) \cdot (Df(a) \cdot x_1, Df(a) \cdot x_2). \end{cases} \quad (1.5)$$

### Lemma 1.1

*The two following assertions are equivalent:*

- (i) *the function  $(\tau, s) \mapsto \varepsilon(\tau, s)$  is strictly convex;*
- (ii) *the function  $(\tau, \varepsilon) \mapsto -s(\tau, \varepsilon)$  is strictly convex.*

*Proof.* We apply formula (1.5) to the functions  $f(\tau, \varepsilon) = (\tau, s(\tau, \varepsilon))$  and  $g(\tau, s) = \varepsilon(\tau, s)$ . Hence  $g \circ f(\tau, s) = \varepsilon(\tau, s(\tau, \varepsilon)) = \varepsilon$ , and in that case

$$D^2(\varepsilon \circ f)(\tau, s) = 0. \quad (1.6)$$

Let us compute the corresponding right-hand side of (1.5) with  $x_1 = x_2 = (\tau_1, \varepsilon_1)$ . We have

$$\begin{cases} Df(\tau, \varepsilon)(\tau_1, \varepsilon_1) = (\tau_1, Ds(\tau, \varepsilon) \cdot (\tau_1, \varepsilon_1)), \\ D^2 f(\tau, \varepsilon)(\tau_1, \varepsilon_1)^{(2)} = (0, D^2 s(\tau, \varepsilon) \cdot (\tau_1, \varepsilon_1)^{(2)}), \end{cases} \quad (1.7)$$

and also for all  $(\tau_1, s_1)$

$$D\varepsilon(\tau, s) \cdot (\tau_1, s_1) = \frac{\partial \varepsilon}{\partial \tau}(\tau, s)\tau_1 + \frac{\partial \varepsilon}{\partial s}(\tau, s)s_1,$$

which together with (1.7) yields

$$\begin{cases} D\varepsilon(\tau, s(\tau, \varepsilon)) \cdot (D^2 f(\tau, \varepsilon)(\tau_1, \varepsilon_1)^{(2)}) \\ = \frac{\partial \varepsilon}{\partial s}(\tau, s(\tau, \varepsilon)) D^2 s(\tau, s) \cdot (\tau_1, \varepsilon_1)^{(2)} \end{cases} \quad (1.8)$$

and

$$\begin{cases} D^2 \varepsilon(\tau, s(\tau, \varepsilon)) \cdot (Df(\tau, \varepsilon)(\tau_1, \varepsilon_1)^{(2)}) \\ = D^2 \varepsilon(\tau, s(\tau, \varepsilon)) \cdot (\tau_1, Ds(\tau, \varepsilon) \cdot (\tau_1, \varepsilon_1))^ {(2)}. \end{cases} \quad (1.9)$$

Combining (1.5)–(1.9), we get

$$D^2 s(\tau, \varepsilon) \cdot (\tau_1, \varepsilon_1)^{(2)} = -\left(\frac{\partial \varepsilon}{\partial s}\right)^{-1} D^2 \varepsilon(\tau, s(\tau, \varepsilon)) \cdot (\tau_1 \cdot Ds(\tau, \varepsilon) \cdot (\tau_1, \varepsilon_1))^ {(2)}.$$

Since  $\frac{\partial \varepsilon}{\partial s} > 0$ , this proves that (i)  $\implies$  (ii). The converse follows by exchanging the roles of the functions  $\varepsilon$  and  $s$ .  $\square$

The next step consists in proving the following equivalence property.

*Lemma 1.2*

*The two following assertions are equivalent:*

- (i) *the function  $(\tau, \varepsilon) \mapsto -s(\tau, \varepsilon)$  is strictly convex;*
- (ii) *the function  $(\tau, \mathbf{u}, e) \mapsto -s(\tau, e - \frac{|\mathbf{u}|^2}{2})$  is strictly convex.*

*Proof.* Let us now choose  $f(\tau, \mathbf{u}, e) = (\tau, e - \frac{|\mathbf{u}|^2}{2})$  and  $g(\tau, \varepsilon) = s(\tau, \varepsilon)$ . Then, an easy computation shows that

$$\begin{aligned} Df(\tau, \mathbf{u}, e) \cdot (\tau_1, \mathbf{u}_1, e_1) &= (\tau_1, e_1 - \mathbf{u} \cdot \mathbf{u}_1), \\ D^2 f(\tau, \mathbf{u}, e) \cdot (\tau_1, \mathbf{u}_1, e_1)^{(2)} &= (0, -|\mathbf{u}_1|^2), \end{aligned}$$

and using (1.5) we get for all triple  $(\tau_1, \mathbf{u}_1, e_1)$

$$\begin{cases} D^2(s \circ f)(\tau, \mathbf{u}, e) \cdot (\tau_1, \mathbf{u}_1, e_1)^{(2)} \\ = -\frac{\partial s}{\partial \varepsilon}\left(\tau, e - \frac{|\mathbf{u}|^2}{2}\right)|\mathbf{u}_1|^2 + D^2 s\left(\tau, e - \frac{|\mathbf{u}|^2}{2}\right) \cdot (\tau_1, e_1 - \mathbf{u} \cdot \mathbf{u}_1)^{(2)}. \end{cases} \quad (1.10)$$

Since  $\frac{\partial s}{\partial \varepsilon} > 0$ , this proves that (i)  $\implies$  (ii).

Conversely, if we choose  $\mathbf{u}_1 = \mathbf{0}$  in (1.10), we obtain for all pair  $(\tau_1, e_1)$

$$D^2 s\left(\tau, e - \frac{|\mathbf{u}|^2}{2}\right) \cdot (\tau_1, e_1)^{(2)} = D^2(s \circ f)(\tau, \mathbf{u}, e) \cdot (\tau_1, \mathbf{0}, e_1)^{(2)},$$

which proves in turn that (ii)  $\implies$  (i).  $\square$

Now, let  $C$  be a convex cone of  $\mathbb{R}^p$ , and let  $\pi : \mathbb{R}_+^* \times C \mapsto \mathbb{R}$  be a  $C^2$  function. We define the function  $\omega : \mathbb{R}_+^* \times C \mapsto \mathbb{R}$  by

$$\omega(\mu, \mathbf{v}) = \mu \pi\left(\frac{1}{\mu}, \frac{\mathbf{v}}{\mu}\right). \quad (1.11)$$

*Lemma 1.3*

The function  $\omega$  is strictly convex (resp. convex) if and only if the function  $\pi$  is strictly convex (resp. convex).

*Proof.* Let us set

$$\hat{\pi}(\mu, \mathbf{v}) = \pi\left(\frac{1}{\mu}, \frac{\mathbf{v}}{\mu}\right)$$

so that

$$\omega(\mu, \mathbf{v}) = \mu \hat{\pi}(\mu, \mathbf{v}).$$

Then, simply by differentiating the bilinear form  $(\mu, \mathbf{v}) \mapsto \mu \mathbf{v}$ , we get

$$D^2 \omega(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)} = \mu D^2 \hat{\pi}(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)} + 2\mu_1 D\hat{\pi}(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1). \quad (1.12)$$

Let us next define on  $\mathbb{R}_+^* \times C$

$$r(\mu, \mathbf{v}) = \left(\frac{1}{\mu}, \frac{\mathbf{v}}{\mu}\right). \quad (1.13)$$

Since  $\hat{\pi}(\mu, \mathbf{v}) = \pi \circ r(\mu, \mathbf{v})$ , we have, again using (1.5),

$$\begin{cases} D\hat{\pi}(\mu, \mathbf{v})(\mu_1, \mathbf{v}_1) = D\pi\left(\frac{1}{\mu}, \frac{\mathbf{v}}{\mu}\right) \cdot (Dr(\mu, \mathbf{v})(\mu_1, \mathbf{v}_1)), \\ D^2 \hat{\pi}(\mu, \mathbf{v})(\mu_1, \mathbf{v}_1)^{(2)} = D\pi\left(\frac{1}{\mu}, \frac{\mathbf{v}}{\mu}\right) \cdot (D^2 r(\mu, \mathbf{v})(\mu_1, \mathbf{v}_1)^{(2)}) \\ \quad + D^2 \pi\left(\frac{1}{\mu}, \frac{\mathbf{v}}{\mu}\right) \cdot (Dr(\mu, \mathbf{v})(\mu_1, \mathbf{v}_1))^{(2)}. \end{cases} \quad (1.14)$$

Now, we notice that differentiating two times the relation

$$\mu r(\mu, \mathbf{v}) = (1, \mathbf{v})$$

yields the identity

$$(0, 0) = 2\mu_1 \ Dr(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1) + \mu D^2 r(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)}. \quad (1.15)$$

Substituting (1.14) in (1.12) and using (1.15), we get

$$\begin{aligned} D^2 \omega(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)} &= \mu D^2 \pi \left( \frac{1}{\mu}, \frac{\mathbf{v}}{\mu} \right) \cdot (Dr(\mu, \mathbf{v})(\mu_1, \mathbf{v}_1))^{(2)} \\ &\quad + D\pi \left( \frac{1}{\mu}, \frac{\mathbf{v}}{\mu} \right) \cdot \{2\mu_1 Dr(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1) + \mu D^2 r(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)}\} \\ &= \mu D^2 \pi \left( \frac{1}{\mu}, \frac{\mathbf{v}}{\mu} \right) \cdot Dr(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)}, \end{aligned}$$

and hence

$$D^2 \omega(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)} = \mu D^2 \pi \left( \frac{1}{\mu}, \frac{\mathbf{v}}{\mu} \right) \cdot \left( -\frac{\mu_1}{\mu^2}, -\left( \frac{\mu_1}{\mu^2} \right) \mathbf{v} + \left( \frac{1}{\mu} \right) \mathbf{v}_1 \right)^{(2)}.$$

This proves that  $\omega$  is convex (resp. strictly convex) if  $\pi$  is convex (resp. strictly convex). The reverse property is clear since

$$\pi(\lambda, \mathbf{u}) = \lambda \omega \left( \frac{1}{\lambda}, \frac{\mathbf{u}}{\lambda} \right)$$

enables us to exchange  $\pi$  and  $\omega$ . □

Next, let  $C$  be a convex subset of  $\mathbb{R}_+ \times \mathbb{R}^p$ , and let  $\omega : C \mapsto \mathbb{R}$  be  $C^2$ . We define a convex subset  $\overline{C}$  of  $\mathbb{R}^n \times \mathbb{R}^p$  by

$$\overline{C} : \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^p, (|\mathbf{x}|, \mathbf{y}) \in C\},$$

and the function  $\overline{\omega} : \overline{C} \mapsto \mathbb{R}$  by

$$\overline{\omega}(\mathbf{u}, \mathbf{v}) = \omega(|\mathbf{u}|, \mathbf{v}).$$

*Lemma 1.4*

*Assume that the function  $\omega$  is monotonically increasing with respect to the first variable. Then, the function  $\overline{\omega}$  is strictly convex (resp. convex) if and only if the function  $\omega$  is strictly convex (resp. convex).*

*Proof.* Setting  $n(\mathbf{x}) = |\mathbf{x}|$ , and using once more (1.5), we have

$$\begin{aligned} D^2 \overline{\omega}(\mathbf{u}, \mathbf{v}) \cdot (\mathbf{u}_1, \mathbf{v}_1)^{(2)} &= \frac{\partial \omega}{\partial \mu} (|\mathbf{u}|, \mathbf{v}) D^2 n(\mathbf{u}) \cdot (\mathbf{u}_1)^{(2)} + D^2 \omega(|\mathbf{u}|, \mathbf{v}) \cdot (Dn(\mathbf{u}) \mathbf{u}_1, \mathbf{v}_1)^{(2)} \end{aligned}$$

with

$$Dn(\mathbf{u}) \cdot \mathbf{u}_1 = \frac{\mathbf{u}^T \mathbf{u}_1}{|\mathbf{u}|}$$

so that

$$D^2\bar{\omega}(\mathbf{u}, \mathbf{v}) \cdot (\mathbf{u}_1, \mathbf{v}_1)^{(2)} = \frac{\partial \omega}{\partial \mu}(|\mathbf{u}|, \mathbf{v}) D^2n(\mathbf{u}) \cdot (\mathbf{u}_1)^{(2)} + D^2\omega(|\mathbf{u}|, \mathbf{v}) \cdot \left( \frac{\mathbf{u}^T \mathbf{u}^1}{|\mathbf{u}|}, \mathbf{v}_1 \right)^{(2)}.$$

By assumption,  $\frac{\partial \omega}{\partial \mu} > 0$ ; hence, if  $\omega$  is convex or strictly convex so is  $\bar{\omega}$ .

Conversely, we may write

$$\omega(\mu, \mathbf{v}) = \bar{\omega}(\mu \mathbf{e}, \mathbf{v}),$$

where  $\mathbf{e}$  is any fixed unit vector of  $\mathbb{R}^n$ . Then

$$D^2\omega(\mu, \mathbf{v}) \cdot (\mu_1, \mathbf{v}_1)^{(2)} = D^2\bar{\omega}(\mu \mathbf{e}, v) \cdot (\mu_1 \mathbf{e}, \mathbf{v}_1)^{(2)},$$

which proves the reverse property.  $\square$

*Proof of Theorem 1.1.* In the case  $d = 1$ , thanks to Lemmas 1.1 to 1.3, (1.2) implies that the function

$$(\rho, q, E) \mapsto -\rho s\left(\frac{1}{\rho}, \frac{E}{\rho} - \frac{q^2}{2\rho}\right)$$

is strictly convex. In fact, taking  $\pi(\tau, (u, e)) = -s(\tau, e - \frac{u^2}{2})$  in (1.11) gives

$$\omega(\rho, (q, E)) = -\rho s\left(\frac{1}{\rho}, \frac{E}{\rho} - \frac{q^2}{2\rho}\right) = -\rho s\left(\tau, e - \frac{u^2}{2}\right) = -\rho s(\tau, \varepsilon).$$

In the case  $d > 1$ , we may apply Lemma 1.4 to the function  $\omega$  defined by

$$(\mu, \mathbf{v} = (\tau, e)) \mapsto \omega(\mu, \mathbf{v}) = -s\left(\tau, e - \frac{\mu^2}{2}\right).$$

Since

$$\frac{\partial \omega}{\partial \mu}(\mu, \mathbf{v}) = +\mu \frac{\partial s}{\partial \varepsilon}\left(\tau, e - \frac{\mu^2}{2}\right) > 0,$$

Lemmas 1.1 to 1.3 imply the theorem.

Next, applying Lemma 1.4 with

$$\Phi = (\mathbf{u}, e), \quad \eta(\tau, \Phi) = -s(\tau, e - \frac{1}{2}|\mathbf{u}|^2)$$

proves the equivalence of the assertions (ii) and (iii) of Theorem 1.1.  $\square$

We now turn to the general situation of a system of conservation laws which can be written either in Eulerian coordinates (cf. Chap. II, Eq. (2.16)) or in Lagrangian coordinates (cf. Chap. II, Eq. (2.17)). As a consequence of Lemma 1.3, we obtain

*Lemma 1.5*

A function  $(\tau, \Phi) \mapsto \eta(\tau, \Phi)$  is strictly convex (resp. convex) if and only if the function  $(\rho, \rho\Phi) \mapsto \rho\eta\left(\frac{1}{\rho}, \frac{\rho\Phi}{\rho}\right)$  is strictly convex (resp. convex).

*Proof.* Setting

$$\pi(\tau, \Phi) = \eta(\tau, \Phi),$$

we apply Lemma 1.3 with

$$\mu = \rho = \frac{1}{\tau}, \quad \mathbf{v} = \rho\Phi$$

so that

$$\omega(\rho, \rho\Phi) = \rho\eta\left(\frac{1}{\rho}, \frac{\rho\Phi}{\rho}\right)$$

□

Next, applying Lemma 1.4 with

$$\Phi = (\mathbf{u}, e), \quad \eta(\tau, \Phi) = -s(\tau, e - \frac{1}{2}|\mathbf{u}|^2)$$

proves the equivalence of the assertions (ii) and (iii) and the validity of Theorem 1.1 for general fluid systems.

As another application of the above results, we consider the ideal MHD system written in Lagrangian coordinates or in Eulerian coordinates (cf. Chap. I, Examples 2.2 and 2.4, Chap. II, Example 2.5). Setting

$$\mathbf{u} = (u, v, w)^T, \quad \mathbf{B}_\perp = (B_y, B_z)^T,$$

we can state

*Theorem 1.2*

*The function*

$$(\tau, \mathbf{u}, \tau\mathbf{B}_\perp, e^*) \mapsto -s\left(\tau, e^* - \frac{1}{2}|\mathbf{u}|^2 - \frac{\tau|\mathbf{B}|^2}{2\mu}\right)$$

*is a strictly convex entropy for the ideal MHD system written in Lagrangian coordinates. Equivalently, the function*

$$(\rho, \rho\mathbf{u}, \mathbf{B}_\perp, \rho e^*) \mapsto -\rho s\left(\frac{1}{\rho}, e^* - \frac{1}{2}|\mathbf{u}|^2 - \frac{|\mathbf{B}|^2}{2\mu\rho}\right)$$

*is a strictly convex entropy for the ideal MHD system in Eulerian coordinates.*

*Proof.* We first check that the function

$$(\tau, \mathbf{u}, \tau\mathbf{B}_\perp, e^*) \mapsto -s\left(\tau, e^* - \frac{1}{2}|\mathbf{u}|^2 - \frac{\tau|\mathbf{B}|^2}{2\mu}\right)$$

is strictly convex. We set

$$\begin{aligned} f(\tau, \mathbf{u}, \tau \mathbf{B}_\perp, e^*) &= \left( \tau, e^* - \frac{1}{2} |\mathbf{u}|^2 - \frac{\tau |\mathbf{B}|^2}{2\mu} \right) = \\ &= \left( \tau, e^* - \frac{1}{2} |\mathbf{u}|^2 - \frac{B_x^2}{2\mu} \tau - \frac{1}{2\mu\tau} |\tau \mathbf{B}_\perp|^2 \right). \end{aligned}$$

We have

$$\begin{aligned} Df(\tau, \mathbf{u}, \tau \mathbf{B}_\perp, e^*) \cdot (\tau_1, \mathbf{u}_1, (\tau \mathbf{B}_\perp)_1, e_1^*) &= \\ &= \left( \tau_1, e_1^* - \mathbf{u} \cdot \mathbf{u}_1 - \frac{1}{\mu\tau} (\tau \mathbf{B}_\perp) \cdot (\tau \mathbf{B}_\perp)_1 + \frac{1}{2\mu} \left( \frac{|\tau \mathbf{B}_\perp|^2}{\tau^2} - B_x^2 \right) \tau_1 \right) \end{aligned}$$

and then

$$\begin{aligned} D^2 f(\tau, \mathbf{u}, \tau \mathbf{B}_\perp, e^*) \cdot (\tau_1, \mathbf{u}_1, (\tau \mathbf{B}_\perp)_1, e_1^*)^{(2)} &= \\ &= \left( 0, -|\mathbf{u}_1|^2 - \frac{1}{\mu\tau} |(\tau \mathbf{B}_\perp) \frac{\tau_1}{\tau} - (\tau \mathbf{B}_\perp)_1|^2 \right). \end{aligned}$$

Then, using (1.5), we obtain

$$\begin{cases} D^2(s \circ f)(\tau, \mathbf{u}, \tau \mathbf{B}_\perp, e^*) \cdot (\tau_1, \mathbf{u}_1, (\tau \mathbf{B}_\perp)_1, e_1^*)^{(2)} = \\ = -\frac{\partial s}{\partial \varepsilon}(\tau, \varepsilon) \left( |\mathbf{u}_1|^2 + \frac{1}{\mu\tau} |(\tau \mathbf{B}_\perp) \frac{\tau_1}{\tau} - (\tau \mathbf{B}_\perp)_1|^2 \right) + \\ + D^2 s(\tau, \varepsilon) \cdot \left( \tau_1, e_1^* - \mathbf{u} \cdot \mathbf{u}_1 - \frac{1}{\mu\tau} (\tau \mathbf{B}_\perp) \cdot (\tau \mathbf{B}_\perp)_1 + \right. \\ \left. + \frac{1}{2\mu} \left( \frac{|\tau \mathbf{B}_\perp|^2}{\tau^2} - B_x^2 \right) \tau_1 \right)^{(2)}, \end{cases}$$

where

$$\varepsilon = e^* - \frac{1}{2} |\mathbf{u}|^2 - \frac{\tau |\mathbf{B}|^2}{2\mu}.$$

Since  $\frac{\partial s}{\partial \varepsilon}$  is  $> 0$ , the right-hand side is  $\leq 0$  which proves that  $-s$  is a convex function of the conservative variables  $(\tau, \mathbf{u}, \tau \mathbf{B}_\perp, e^*)$ . Let us next check that this function is strictly convex. Indeed, assume that the right-hand side vanishes. Then we have on the one hand

$$|\mathbf{u}_1|^2 + \frac{1}{\mu\tau} |(\tau \mathbf{B}_\perp) \frac{\tau_1}{\tau} - (\tau \mathbf{B}_\perp)_1|^2 = 0$$

and on the other hand

$$\begin{cases} \tau_1 = 0, \\ e_1^* - \mathbf{u} \cdot \mathbf{u}_1 - \frac{1}{\mu\tau} (\tau \mathbf{B}_\perp) \cdot (\tau \mathbf{B}_\perp)_1 + \frac{1}{2\mu} \left( \frac{|\tau \mathbf{B}_\perp|^2}{\tau^2} - B_x^2 \right) \tau_1 = 0. \end{cases}$$

This implies

$$\mathbf{u}_1 = \mathbf{0}, \quad (\tau \mathbf{B}_\perp)_1 = \mathbf{0}, \quad e_1^* = 0$$

which yields the strict convexity property.  $\square$

## 1.2 Ideal Gases

Before concluding this introductory section, let us consider more closely the case of an *ideal gas*, or perfect gas, i.e., such that the equation of state satisfies Gay-Lussac and Boyle's law

$$p = \rho RT = \frac{RT}{\tau}, \quad (1.16)$$

where  $T$  is the temperature and  $R$  is the specific gas constant;  $R = \eta \mathcal{R}$ , where  $\eta$  is the mole-mass fraction and  $\mathcal{R}$  is the universal gas constant. From (1.1) and (1.16), one deduces

$$R \frac{\partial \varepsilon}{\partial s} + \tau \frac{\partial \varepsilon}{\partial \tau} = 0,$$

which implies that  $\varepsilon$  is a function of  $\tau \exp(-\frac{s}{R})$ , that is,

$$\varepsilon = \varphi\left(\tau \exp\left(-\frac{s}{R}\right)\right)$$

for some function  $\varphi$ . Computing  $T = \frac{\partial \varepsilon}{\partial s}$  and taking hypothesis (1.3) into account, one can prove that  $\varepsilon$  is then a function of  $T$  only (see Courant and Friedrichs [371], Chapter I, Section 4]),  $\varepsilon = \varepsilon(T)$ . One defines the function  $\gamma(T)$  by

$$\frac{d\varepsilon}{dT} = \frac{R}{\gamma(T) - 1}, \quad (1.17)$$

and  $\frac{d\varepsilon}{dT}$  is the “specific heat at constant volume”  $C_v$ . Hence

$$p = \rho RT, \quad \text{where } R \frac{dT}{d\varepsilon} = \gamma(T) - 1$$

(which does not mean that  $p = (\gamma(T) - 1)\rho\varepsilon$ , unless  $\gamma$  is constant, see (1.19) below). Since  $R > 0$  and  $C_v > 0$ , we have  $\gamma(T) > 1$ .

In the same way, introducing the specific enthalpy

$$h = \varepsilon + p\tau,$$

which satisfies

$$dh = T ds + \tau dp,$$

we obtain that for an ideal gas

$$h = \varepsilon + RT,$$

$h$  is also a function of  $T$  only, and

$$\frac{dh}{dT} = \frac{R\gamma(T)}{(\gamma(T) - 1)},$$

$\frac{dh}{dT}$  is the “specific heat at constant pressure”  $C_p$ . Hence

$$\gamma(T) = \frac{C_p}{C_v}$$

is the ratio of two specific heats. Note also that from (1.4) and (1.16)

$$c^2 = -\tau^2 \frac{\partial p}{\partial \tau}(\tau, s) = RT - R\tau \frac{\partial T}{\partial \tau}.$$

Now, since from (1.1), (1.16), and (1.17)

$$\tau R \frac{\partial T}{\partial \tau} = \tau \frac{dT}{d\varepsilon} \frac{\partial \varepsilon}{\partial \tau} = -p\tau(\gamma(T) - 1) = -RT(\gamma(T) - 1),$$

we get

$$c^2 = \gamma(T)RT = \frac{\gamma(T)p}{\rho}. \quad (1.18)$$

One sometimes says that the gas is thermally perfect (but calorically imperfect).

Let us now focus on the case of a *polytropic* ideal gas (thermally and calorically perfect) for which  $\varepsilon$  is proportional to  $T$ . Thus  $\gamma$ , the adiabatic exponent,  $C_v = \frac{R}{(\gamma-1)}$ , and  $C_p = \frac{R\gamma}{(\gamma-1)}$  are constant, and

$$\varepsilon = C_v T = \frac{RT}{\gamma-1}, \quad h = C_p T = \frac{\gamma RT}{\gamma-1}, \quad (1.19)$$

which yields the traditional  $\gamma$ -law

$$p = (\gamma - 1)\rho\varepsilon. \quad (1.20)$$

Note that for usual cases  $1 < \gamma \leq \frac{5}{3}$ . We shall see in Chap. IV, Sect. 7 (Remarks 7.2 and (7.18)) that  $\gamma = \frac{5}{3}$  corresponds to a monatomic gas in dimension  $d = 3$  and  $\gamma = \frac{7}{5} = 1.4$  for diatomic molecules. Now, we also obtain from (1.1) (see Chap. I, Example 5.3)

$$ds = C_v \left( \frac{d\varepsilon}{\varepsilon} - (\gamma - 1) \frac{d\rho}{\rho} \right),$$

which gives

$$s - s_0 = C_v \operatorname{Log} \left( \frac{\varepsilon}{\rho^{(\gamma-1)}} \right)$$

or

$$\varepsilon = \rho^{(\gamma-1)} \exp \left( \frac{(s - s_0)}{C_v} \right).$$

Thus

$$p = (\gamma - 1) \exp\left(\frac{(s - s_0)}{C_v}\right) \rho^\gamma = A(s) \rho^\gamma.$$

Note also that

$$c^2 = \gamma RT = (\gamma - 1)h,$$

and the function  $\ell$  such that  $u \pm \ell$  are the 1- and 3-Riemann invariants is given by

$$\ell = \int \left(\frac{c}{\rho}\right) d\rho = \frac{2c}{\gamma - 1}.$$

*Example 1.1.* Air under normal conditions ( $p$  and  $T$  moderate enough) can be considered as a perfect gas with  $\gamma = \frac{7}{5}$  (approximately a mixture of two diatomic molecular species: 20% of O<sub>2</sub>, 80% N<sub>2</sub>). When the temperature increases, the vibrational motion of oxygen and nitrogen molecules in air becomes important, and specific heats vary with temperature so that  $\gamma$  is no longer constant but depends on temperature: the air is thermally perfect. However, at even higher temperatures, the molecules of oxygen and nitrogen begin to dissociate, which is symbolized by the reactions



The air becomes chemically reacting, and the specific heats are functions of both  $T$  and  $p$ . These facts are improperly termed *real gas effects*.  $\square$

*Remark 1.1.* A gas should rather be defined as *real* when intermolecular forces become important (at very cold temperatures and high pressures). The perfect gas equation of state must be replaced by more accurate relations such as the van der Waals equation

$$\left(p + \frac{a}{\tau^2}\right)(\tau - b) = RT,$$

where  $a$  and  $b$  are constants depending on the gas.  $\square$

*Remark 1.2.* A mixture of reacting gases such as air at high temperatures is said to be in chemical equilibrium if the forward and reverse reactions  $AB \rightleftharpoons A + B$  are balanced so that the species are present in fixed amount. The chemical composition is then uniquely determined by  $p$  and  $T$ . Most problems can be treated assuming a mixture of perfect gases, for which Dalton's law states that the total pressure is the sum of the partial pressures

$$p = \sum_\alpha p_\alpha,$$

where the summation is taken over all species, together with  $p_\alpha = \rho \eta_\alpha \mathcal{R}T$  ( $\eta_\alpha$  is the mole-mass fraction of species  $\alpha$ ).  $\square$

Let us now consider the more general case of a gas with an incomplete equation of state expressed in the form

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

which will prove convenient in Chap. IV. We set  $\tilde{\varepsilon} = \rho \varepsilon$  and

$$\kappa = \frac{\partial p}{\partial \tilde{\varepsilon}}(\rho, \tilde{\varepsilon}), \quad \chi = \frac{\partial p}{\partial \rho}(\rho, \tilde{\varepsilon}), \quad (1.20)$$

and define

$$\gamma = \frac{\rho c^2}{p},$$

where  $c$  is the speed of sound. Then, for an ideal gas, this definition of  $\gamma$  and the previous one,  $\gamma(T) = \frac{C_p}{C_v}$ , coincide (which is not the case for a more general gas). Moreover, computing  $\kappa$  and  $\chi$

$$\kappa = R\rho \left( \frac{dT}{d\varepsilon} \right) \frac{\partial \varepsilon}{\partial \tilde{\varepsilon}} \Big|_{\rho=\text{cte}} = RT'(\varepsilon) = \gamma(T) - 1$$

and

$$\chi = RT + \rho R \left( \frac{dT}{d\varepsilon} \right) \frac{\partial \varepsilon}{\partial \rho} \Big|_{\tilde{\varepsilon}=\text{cte}} = R \left( T - \varepsilon \frac{dT}{d\varepsilon} \right) = RT - (\gamma(T) - 1)\varepsilon,$$

we see that they depend only on  $T$  (or  $\varepsilon$ ). Finally, we have

$$\chi\rho + \kappa\tilde{\varepsilon} = (RT - (\gamma(T) - 1)\varepsilon)\rho + (\gamma(T) - 1)\tilde{\varepsilon} = \rho RT.$$

Therefore, for an ideal gas, the function  $p = p(\rho, \tilde{\varepsilon})$  satisfies the identity

$$p = \rho \frac{\partial p}{\partial \rho}(\rho, \tilde{\varepsilon}) + \tilde{\varepsilon} \frac{\partial p}{\partial \tilde{\varepsilon}}(\rho, \tilde{\varepsilon}). \quad (1.21)$$

Conversely, given a function  $p = p(\rho, \varepsilon)$  such that (1.21) holds, we can show that it corresponds to an ideal gas. Indeed, assuming (1.21), an easy computation gives

$$\frac{\partial(p(\rho, \varepsilon)/\rho)}{\partial \rho} = 0.$$

Hence  $\frac{p}{\rho}$  is a function of  $\varepsilon$  only that reads

$$\frac{p}{\rho} = T(\varepsilon),$$

where we have incorporated the constant  $R$  in the function  $T$ . Moreover, since

$$\frac{\partial(p(\rho, \varepsilon)/\rho)}{\partial \varepsilon} = \frac{\partial p}{\partial \tilde{\varepsilon}}(\rho, \tilde{\varepsilon})$$

and  $\frac{\partial p}{\partial \tilde{\varepsilon}}(\rho, \tilde{\varepsilon}) = \kappa$  by definition (1.20) of  $\kappa$ , we find that  $\kappa$  is a function of  $\varepsilon$  only and that the function  $T$  satisfies

$$T'(\varepsilon) = \kappa.$$

*Example 1.2.* Consider a *stiffened* equation of Grüneisen type, which we write

$$p = (\gamma - 1)\rho\varepsilon + c_{\text{ref}}^2(\rho - \rho_{\text{ref}}), \quad \gamma > 1. \quad (1.22)$$

Such an equation is obtained by linearization from a Grüneisen equation of state for a metal (see Menikoff and Plohr [862]). We have  $\kappa = \gamma - 1$  and  $\chi = c_{\text{ref}}^2$ . We shall see (Chap. IV, Lemma 4.3) that

$$c^2 = \kappa h + \chi,$$

which gives

$$c^2 = \gamma(p + p_\infty)\tau,$$

where we have set  $p_\infty = \frac{c_{\text{ref}}^2 \rho_{\text{ref}}}{\gamma}$  for convenience.  $\square$

We shall again use these notions in Chap. IV, Sects. 4–6 for the application of the usual schemes to gas dynamics, and we shall also return to the subject in Chap. V, Sect. 2.

## 2 Entropy Satisfying Shock Conditions

Let us consider again the gas dynamics equations in slab symmetry written in conservation form and in Eulerian coordinates,

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = 0. \end{cases} \quad (2.1)$$

We supplement equations (2.1) with an equation of state, which we can take to be of the form

$$\varepsilon = \varepsilon(\tau, p). \quad (2.2)$$

In this section, using the ideas of Chap. II, Sect. 5, we want to characterize the admissible shock discontinuities for the physical (or entropy) weak solutions of (2.1). In this particular case of fundamental importance, we will be able to give a global description of the shock curves, in contrast with the general situation of Chap. II where only a local description of shock curves was obtained. We start from the Rankine-Hugoniot jump conditions for the system (2.1):

$$\begin{cases} \sigma[\rho] = [\rho u], \\ \sigma[\rho u] = [\rho u^2 + p], \\ \sigma[\rho e] = [(\rho e + p)u], \end{cases} \quad (2.3)$$

where  $\sigma$  is the speed of the discontinuity.

In addition, since by Theorem 1.1  $-\rho s$  is a strictly convex function of the conservative variables  $\rho, q = \rho u, E = \rho e$ , we obtain that  $-\rho s$  is a strictly convex (mathematical) entropy associated with the entropy flux  $-\rho su$  (the fact that  $-\rho su$  is indeed the associated entropy flux is not difficult to check). Therefore, the entropy considerations of Chap. I, Sect. 5 and Chap. II, Sect. 5 lead us to require that any physical weak solution of (2.1), (2.2) satisfies the entropy inequality (in the sense of distributions)

$$\frac{\partial}{\partial t}(\rho s) + \frac{\partial}{\partial x}(\rho su) \geq 0. \quad (2.4)$$

This means that any jump discontinuity must satisfy the entropy jump condition

$$\sigma[\rho s] \leq [\rho su]. \quad (2.5)$$

Let us denote by (0) and (1) the two states that are connected by the discontinuity. Later on, we shall give a more precise meaning to the indices 0 and 1, but for the moment, this is just a way of discriminating between the two states under consideration. We set

$$v_i = u_i - \sigma, \quad i = 0, 1, \quad (2.6)$$

so that  $v_i$  is the flow velocity relative to the discontinuity.

*Lemma 2.1*

*The Rankine-Hugoniot jump relations (2.3) can be equivalently written in the form*

$$\begin{cases} \rho_0 v_0 = \rho_1 v_1, \\ \rho_0 v_0^2 + p_0 = \rho_1 v_1^2 + p_1, \\ \left( \rho_0 \left( \varepsilon_0 + \frac{v_0^2}{2} \right) + p_0 \right) v_0 = \left( \rho_1 \left( \varepsilon_1 + \frac{v_1^2}{2} \right) + p_1 \right) v_1, \end{cases} \quad (2.7)$$

with  $\sigma = u_i - v_i$ .

*Proof.* The first equation (2.7) coincides with the first equation (2.3). Next, we have

$$\rho_0 v_0^2 + p_0 = \rho_0(u_0 - \sigma)^2 + p_0 = \rho_0 u_0^2 + p_0 - \sigma \rho_0 u_0 - \sigma \rho_0 v_0.$$

Hence, the second equation (2.3) together with the first equation (2.7) implies the second equation (2.7). Finally, we write

$$\begin{aligned} \left( \rho_0 \left( \varepsilon_0 + \frac{v_0^2}{2} \right) + p_0 \right) v_0 &= \left( \rho_0 \left( \varepsilon_0 + \frac{(u_0 - \sigma)^2}{2} \right) + p_0 \right) (u_0 - \sigma) \\ &= \{(\rho_0 e_0 + p_0) u_0 - \sigma \rho_0 e_0\} - \sigma (\rho_0 v_0^2 + p_0) - \sigma^2 \rho_0 \frac{v_0}{2}, \end{aligned}$$

so that the third equation (2.3) together with the first two equations (2.7) implies the third equation (2.7).  $\square$

Next, we set

$$M = \rho_0 v_0 = \rho_1 v_1, \quad (2.8)$$

so that  $M$  is the mass flux through the curve of discontinuity. We have to distinguish between the two cases  $M = 0$  and  $M \neq 0$ . In the case  $M = 0$ , we have necessarily  $v_0 = v_1 = 0$ , and therefore

$$\begin{cases} u_0 = u_1 = \sigma, \\ p_0 = p_1. \end{cases} \quad (2.9)$$

Since  $\rho_0 \neq \rho_1$  (otherwise the discontinuity would not exist), we obtain a physical contact discontinuity (or slip line) that indeed corresponds to a mathematical 2-contact discontinuity with

$$\sigma = \lambda_2 = u.$$

When  $M \neq 0$ , we obtain a shock discontinuity that may be a 1-shock or a 3-shock.

*Lemma 2.2*

For  $M \neq 0$ , the Rankine-Hugoniot jump conditions can be equivalently written

$$M = \frac{u_1 - u_0}{\tau_1 - \tau_0}, \quad (2.10)$$

$$M^2 = -\frac{p_1 - p_0}{\tau_1 - \tau_0}, \quad (2.11)$$

$$\varepsilon_1 - \varepsilon_0 + \frac{1}{2}(p_1 + p_0)(\tau_1 - \tau_0) = 0. \quad (2.12)$$

*Proof.* First using (2.8), we have

$$M\tau_i = v_i, \quad i = 0, 1 \quad (2.13)$$

and by (2.6)

$$\sigma = u_0 - M\tau_0 = u_1 - M\tau_1,$$

which gives (2.10). Next, again using (2.8), the second equation (2.7) yields

$$Mv_0 + p_0 = Mv_1 + p_1,$$

so that by (2.13)

$$\left( \rho_i \left( \varepsilon_i + \frac{v_i^2}{2} \right) + p_i \right) v_i = M \left( \varepsilon_i + \frac{v_i^2}{2} \right) + p_i v_i = M \left( \varepsilon_i + \frac{v_i^2}{2} + p_i \tau_i \right).$$

Hence, the third equation (2.7) becomes

$$\varepsilon_0 + \frac{v_0^2}{2} + p_0 \tau_0 = \varepsilon_1 + \frac{v_1^2}{2} + p_1 \tau_1,$$

or

$$\varepsilon_1 - \varepsilon_0 + \frac{(v_1^2 - v_0^2)}{2} + p_1 \tau_1 - p_0 \tau_0 = 0.$$

Since by (2.13) and (2.11)

$$v_1^2 - v_0^2 = M^2 (\tau_1^2 - \tau_0^2) = -(p_1 - p_0)(\tau_1 + \tau_0),$$

we obtain

$$\varepsilon_1 - \varepsilon_0 - \frac{(p_1 - p_0)(\tau_1 + \tau_0)}{2} + p_1 \tau_1 - p_0 \tau_0 = 0,$$

and (2.12) follows.

Conversely, it is an easy matter to check that the equations (2.10)–(2.12) imply the Rankine-Hugoniot jump conditions (2.7) if we set  $\sigma = u_i - M\tau_i$ ,  $i = 0, 1$ .  $\square$

*Remark 2.1.* We could have equivalently derived the relations (2.10)–(2.12) starting from the gas dynamics equations written in Lagrangian coordinates,

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial}{\partial m} (pu) = 0. \end{cases} \quad (2.14)$$

Recall that in (2.14)  $m$  stands for a mass variable. Indeed, the Rankine-Hugoniot jump conditions associated with the system (2.14) are given by

$$\begin{cases} \sigma_L[\tau] = -[u], \\ \sigma_L[u] = [p], \\ \sigma_L[e] = [pu], \end{cases} \quad \text{or} \quad \begin{cases} \sigma_L(\tau_1 - \tau_0) = -(u_1 - u_0), \\ \sigma_L(u_1 - u_0) = p_1 - p_0, \\ \sigma_L(e_1 - e_0) = p_1 u_1 - p_0 u, \end{cases} \quad (2.15)$$

where  $\sigma_L$  is the speed of propagation of the discontinuity with respect to the mass variable.

Comparing the first equation (2.15) with (2.10) gives

$$\sigma_L = -M. \quad (2.16)$$

In the case of a contact discontinuity, we have  $\sigma_L = 0$ , which implies

$$u_1 = u_0 \quad \text{and} \quad p_1 = p_0.$$

On the other hand, when  $M \neq 0$ , the first two equations (2.15) together with (2.16) imply trivially (2.11). Moreover, the third equation (2.15) can be written

$$M\left(\varepsilon_1 - \varepsilon_0 + \frac{(u_1^2 - u_0^2)}{2}\right) + p_1 u_1 - p_0 u_0 = 0,$$

or

$$\varepsilon_1 - \varepsilon_0 + \frac{(u_1 + u_0)}{2} \left\{ (u_1 - u_0) + \frac{(p_1 - p_0)}{M} \right\} + \frac{(p_1 + p_0)(u_1 - u_0)}{2M} = 0,$$

which together with (2.10) and (2.11) gives (2.12).  $\square$

Several comments on equations (2.10)–(2.12) are now in order. First, combining (2.10) and (2.11) yields

$$M = -\frac{p_1 - p_0}{u_1 - u_0}. \quad (2.17)$$

Next, we observe that due to (2.11) the pressure  $p$  and the specific volume  $\tau$  vary in opposite ways along a shock curve. Finally, we notice that the equation (2.12), called the *Hugoniot equation*, is of purely thermodynamic nature since by (2.2) it involves the thermodynamic variables  $\tau$  and  $p$ . In other words, (2.12) is the projection onto the  $(\tau, p)$ -plane of the shock relations. Let us then introduce the *Hugoniot function*  $\mathcal{H} = \mathcal{H}(\tau, p)$  with *center*  $(\tau_0, p_0)$  defined by

$$\mathcal{H}(\tau, p) = \varepsilon(\tau, p) - \varepsilon(\tau_0, p_0) + \frac{1}{2}(p + p_0)(\tau - \tau_0), \quad (2.18)$$

so that the Hugoniot equation (2.12) becomes

$$\mathcal{H}(\tau_1, p_1) = 0.$$

The graph  $\mathcal{H}$  of the Hugoniot function with center  $(\tau_0, p_0)$  in the  $(\tau, p)$ -plane is called the *Hugoniot curve* or the *shock adiabatic with center*  $(\tau_0, p_0)$ . This is the set of all states  $(\tau, p)$  that can be connected to the state  $(\tau_0, p_0)$  by a shock (that is not necessarily admissible).

*Example 2.1.* Consider a polytropic ideal gas for which

$$\varepsilon = \frac{p\tau}{\gamma - 1}, \quad \gamma > 1.$$

One can write in this case

$$\begin{aligned}\mathcal{H}(\tau, p) &= \frac{p\tau - p_0\tau_0}{\gamma - 1} + \frac{1}{2}(p + p_0)(\tau - \tau_0) \\ &= \frac{1}{2} \left\{ \left( \tau \frac{(\gamma + 1)}{\gamma - 1} - \tau_0 \right) p + \left( \tau - \tau_0 \frac{(\gamma + 1)}{(\gamma - 1)} \right) p_0 \right\}.\end{aligned}$$

Setting

$$\mu^2 = \frac{\gamma - 1}{\gamma + 1}, \quad (2.19)$$

we find

$$2\mu^2\mathcal{H}(\tau, p) = (\tau - \mu^2\tau_0)p - (\tau_0 - \mu^2\tau)p_0. \quad (2.20)$$

In this case, the Hugoniot curve is a rectangular hyperbola. Observe that for  $\tau > \mu^{-2}\tau_0$  the pressure becomes negative so that the corresponding part of the Hugoniot curve has no physical meaning. Hence, along the Hugoniot curve, the values of  $\tau$  may vary between two limits  $\tau_{\min} = \mu^2\tau_0$  and  $\tau_{\max} = \mu^{-2}\tau_0$ , while the pressure  $p$  varies between 0 and  $+\infty$  (see Fig. 2.1).  $\square$

*Example 2.2.* If we take more generally a stiffened equation of state of Grüneisen type (Example 1.2), which we write

$$\varepsilon = \frac{1}{\gamma - 1}p\tau + \frac{\gamma}{\gamma - 1}p_\infty\tau - \varepsilon_\infty,$$

where  $\varepsilon_\infty = \frac{c_{\text{ref}}^2}{\gamma - 1}$ , we find that the Hugoniot curve is again a hyperbola

$$2\mu^2\mathcal{H}(\tau, p) = (\tau - \mu^2\tau_0)(p + p_\infty) - (\tau_0 - \mu^2\tau)(p_0 + p_\infty).$$

We still have  $p$  varying between 0 and  $+\infty$ ,  $\tau_{\min} = \mu^2\tau_0$ , but

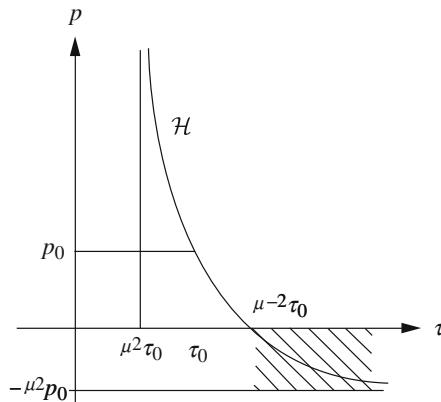


Fig. 2.1 Hugoniot curve  $\mathcal{H}$  for a polytropic ideal gas

$$\tau_{\max} = \tau_0 \frac{p_\infty(1 + \mu^2) + p_0}{p_\infty(1 + \mu^2) + \mu^2 p_0} \leq \mu^{-2} \tau_0$$

so that the upper limit of  $\tau$  is smaller.  $\square$

Let us next study the properties of the Hugoniot curves. For the sake of brevity, we denote by  $A_0 = (\tau_0, p_0)$  the center of the Hugoniot curve, and we set

$$s_0 = s(\tau_0, p_0).$$

*Lemma 2.3*

Assume that the function  $(\tau, s) \mapsto p(\tau, s)$  satisfies the conditions

$$\frac{\partial p}{\partial \tau}(\tau, s) < 0, \quad \frac{\partial p}{\partial s}(\tau, s) > 0. \quad (2.21)$$

Then, we have at point  $A_0 = (\tau_0, p_0)$

$$\frac{\partial \mathcal{H}}{\partial \tau}(\tau_0, p_0) > 0, \quad \frac{\partial \mathcal{H}}{\partial p}(\tau_0, p_0) > 0. \quad (2.22)$$

*Proof.* Differentiating (2.18), we have

$$d\mathcal{H} = d\varepsilon + \frac{1}{2}(p + p_0)d\tau + \frac{1}{2}(\tau - \tau_0)dp$$

and by (1.1)

$$d\mathcal{H} = T ds - \frac{1}{2}(p - p_0)d\tau + \frac{1}{2}(\tau - \tau_0)dp. \quad (2.23)$$

Thus

$$\begin{cases} \frac{\partial \mathcal{H}}{\partial \tau}(\tau, p_0) = T \frac{\partial s}{\partial \tau}(\tau, p_0), \\ \frac{\partial \mathcal{H}}{\partial p}(\tau_0, p) = T \frac{\partial s}{\partial p}(\tau_0, p). \end{cases} \quad (2.24)$$

Now it remains to compute the first partial derivatives of the function  $s(\tau, p)$ . From the identity

$$s = s(\tau, p(\tau, s)),$$

we deduce

$$\begin{cases} \frac{\partial s}{\partial \tau}(\tau, p) + \frac{\partial s}{\partial p}(\tau, p) \frac{\partial p}{\partial \tau}(\tau, s) = 0, \\ \frac{\partial s}{\partial p}(\tau, p) \frac{\partial p}{\partial s}(\tau, s) = 1. \end{cases}$$

Therefore, thanks to assumptions (2.21), it yields

$$\begin{cases} \frac{\partial s}{\partial \tau}(\tau, p) = -\left(\frac{\partial p}{\partial s}(\tau, s)\right)^{-1} \frac{\partial p}{\partial \tau}(\tau, s) > 0, \\ \frac{\partial s}{\partial p}(\tau, p) = \left(\frac{\partial p}{\partial s}(\tau, s)\right)^{-1} > 0, \end{cases} \quad (2.25)$$

which we substitute into (2.24) to get the desired result.  $\square$

*Remark 2.2.* As we have already noted, the first inequality (2.21) follows from the strict convexity of the function  $\varepsilon(\tau, s)$  since (1.1) gives

$$\frac{\partial p}{\partial \tau}(\tau, s) = -\frac{\partial^2 \varepsilon}{\partial \tau^2}(\tau, s).$$

On the other hand, the second condition (2.21) holds in many physical cases and is indeed satisfied in the case of a polytropic ideal gas.  $\square$

Let us denote by a prime the derivative along the Hugoniot curve.

*Lemma 2.4*

Along the Hugoniot curve with center  $A_0$ , we have at point  $A_0$

$$s' = 0, \quad s'' = 0, \quad s''' = -\frac{1}{2} \frac{\partial^2 p}{\partial \tau^2}(\tau_0, s_0)(\tau')^3. \quad (2.26)$$

*Proof.* From (2.23) and the relation  $\mathcal{H}(\tau, p) = 0$ , we get

$$Ts' = (p - p_0) \frac{\tau'}{2} - (\tau - \tau_0) \frac{p'}{2}, \quad (2.27)$$

so that

$$s' = 0 \text{ at } A_0.$$

Next, differentiating (2.27), we write

$$(Ts')' = (p - p_0) \frac{\tau''}{2} - (\tau - \tau_0) \frac{p''}{2},$$

which yields

$$(Ts')' = 0 \text{ at } A_0.$$

But

$$(Ts')' = Ts'' + T's' = Ts'' \text{ at } A_0$$

and

$$s'' = 0 \text{ at } A_0.$$

Finally, differentiating once more (2.27) gives

$$(Ts'')'' = \frac{(p'\tau'' - \tau' p'')}{2} + \frac{(p - p_0)\tau'''}{2} - (\tau - \tau_0)p''',$$

so that

$$(Ts'')'' = Ts''' = \frac{(p'\tau'' - \tau' p'')}{2} \quad \text{at } A_0.$$

On the other hand, using the equation of state in the form  $p = p(\tau, s)$ , we obtain

$$dp = \frac{\partial p}{\partial \tau} d\tau + \frac{\partial p}{\partial s} ds$$

and

$$d^2 p = \frac{\partial^2 p}{\partial \tau^2} (d\tau)^2 + 2 \left( \frac{\partial^2 p}{\partial \tau \partial s} \right) d\tau \, ds + \frac{\partial^2 p}{\partial s^2} (ds)^2 + \frac{\partial p}{\partial \tau} d^2 \tau + \frac{\partial p}{\partial s} d^2 s.$$

Hence, along the Hugoniot curve  $\mathcal{H} = 0$ , we find

$$\begin{cases} p' = \frac{\partial p}{\partial \tau} \tau' & \text{at } A_0, \\ p'' = \frac{\partial^2 p}{\partial \tau^2} (\tau')^2 + \frac{\partial p}{\partial \tau} \tau'' & \text{at } A_0, \end{cases} \quad (2.28)$$

and

$$Ts''' = -\frac{1}{2} \frac{\partial^2 p}{\partial \tau^2} (\tau')^3 \quad \text{at } A_0,$$

which proves the result.  $\square$

As a first consequence of Lemma 2.4, we obtain that the Hugoniot curve with center  $A_0$  and the isentropic curve  $s = s_0$  passing through the point  $A_0$  are osculatory at  $A_0$ . In fact, this property follows from the general theory of Chap. II. When the  $k$ th characteristic field is genuinely nonlinear, we know from Chap. II, Sect. 4 that the  $k$ -shock curve  $\mathcal{S}_k(u_L)$  and the  $k$ -rarefaction curve  $\mathcal{R}_k(u_L)$  are osculatory at the point  $u_L$ . This is exactly the situation here: the 1- and 3-shock curves are projected in the  $(\tau, p)$ -plane onto the Hugoniot curve, while the 1- and 3-rarefaction curves are projected onto the isentropic curve.

Let us point out another consequence of Lemmas 2.3 and 2.4. Since  $\frac{\partial \mathcal{H}}{\partial p} > 0$  at  $A_0$ , we may parametrize the Hugoniot curve  $\mathcal{H} = 0$  in the form  $p = p(\tau)$  in a neighborhood of  $A_0$ . Moreover, if in addition to (2.21) we assume

$$\frac{\partial^2 p}{\partial \tau^2}(\tau, s) > 0, \quad (2.29)$$

it follows from (2.26) that along the Hugoniot curve the physical entropy  $s$  is a decreasing function of  $\tau$  in a neighborhood of  $A_0$ . More generally, we can state the following result.

**Theorem 2.1**

Assume that the function  $p = p(\tau, s)$  satisfies the conditions (2.21) and (2.29). Then the entropy  $s$  is strictly monotone all along the Hugoniot curve with center  $A_0$  and has a unique critical point at  $A_0$ .

*Proof.* By a critical point, we mean a point of the Hugoniot curve for which  $s' = 0$ . Let us prove that, along the Hugoniot curve, we have  $s' \neq 0$  except at  $A_0$ . We begin by considering the straight lines  $\Delta$  passing through  $A_0$  (i.e., the Rayleigh lines) and parametrized by

$$\begin{cases} \tau = \tau_0 + a\alpha, \\ p = p_0 + b\alpha. \end{cases}$$

Observe that along such a line  $\Delta$ , we have

$$(p_0 - p)d\tau + (\tau - \tau_0)dp = 0.$$

Hence, using (2.23), we obtain

$$\frac{d\mathcal{H}}{d\alpha} = T \frac{ds}{d\alpha} \quad \text{along } \Delta. \quad (2.30)$$

Let us check that

$$\text{along } \Delta, s \text{ has at most one critical point, that is a maximum.} \quad (2.31)$$

We first consider a straight line distinct from  $\tau = \tau_0$ . Along such a line, we have

$$\frac{dp}{d\alpha} = \frac{\partial p}{\partial \tau} \frac{d\tau}{d\alpha} + \frac{\partial p}{\partial s} \frac{ds}{d\alpha},$$

and since

$$\frac{d^2\tau}{d\alpha^2} = \frac{d^2p}{d\alpha^2} = 0,$$

we find

$$0 = \frac{d^2p}{d\alpha^2} = \frac{\partial^2p}{\partial\tau^2} \left( \frac{d\tau}{d\alpha} \right)^2 + 2 \left( \frac{\partial^2p}{\partial\tau\partial s} \right) \frac{d\tau}{d\alpha} \frac{ds}{d\alpha} + \frac{\partial^2p}{\partial s^2} \left( \frac{ds}{d\alpha} \right)^2 + \frac{\partial p}{\partial s} \frac{ds}{d\alpha}.$$

Now, we obtain at a critical point of  $s$  along  $\Delta$

$$\frac{\partial^2p}{\partial\tau^2} \left( \frac{d\tau}{d\alpha} \right)^2 + \frac{\partial p}{\partial s} \frac{d^2s}{d\alpha^2} = 0.$$

Since  $\frac{d\tau}{d\alpha} \neq 0$ , it follows from (2.21) and (2.29) that at such a critical point, we have

$$\frac{d^2s}{d\alpha^2} < 0.$$

In other words, a critical point of  $s$  along  $\Delta$  is necessarily a local maximum. This implies the property (2.31).

It remains to consider the case of the line  $\tau = \tau_0$ . Since  $\frac{\partial p}{\partial s} > 0$ , the function  $s \mapsto p(\tau_0, s)$  is strictly increasing and the same is true of the reciprocal function  $p \mapsto s(\tau_0, p)$  so that  $s$  has no critical point along  $\tau = \tau_0$ . This proves our assertion (2.31).

Next, combining (2.30) and (2.31), we obtain that  $\mathcal{H}$  has at most one critical point along  $\Delta$ .

Then, let  $A_1 = (\tau_1, p_1) \neq A_0$  be another critical point of  $s$  along the Hugoniot curve with center  $A_0$ . Since  $\mathcal{H}' = 0$  along this Hugoniot curve, we use (2.27) to obtain

$$(p_0 - p_1)\tau' + (\tau_1 - \tau_0)p' = 0 \text{ at } A_1.$$

This means that the straight line  $A_0 A_1$  is tangent to the Hugoniot curve at the point  $A_1$ . Therefore, along  $A_0 A_1$ ,  $\mathcal{H}$  is critical at  $A_1$ . On the other hand,  $\mathcal{H}$  vanishes at the points  $A_0$  and  $A_1$  and must be critical along  $A_0 A_1$  at some intermediate point  $A_2$  distinct from  $A_0$  and  $A_1$ . Hence, along  $A_0 A_1$ ,  $\mathcal{H}$  has two distinct critical points  $A_1$  and  $A_2$ , which violates the above property.

Thus, along the Hugoniot curve with center  $A_0$ , the entropy  $s$  is critical at the point  $A_0$  alone.  $\square$

As a consequence of Lemma 2.4 and Theorem 2.1, we obtain the following corollary.

*Corollary 2.1*

*Assume, moreover, that the Hugoniot curve may be parametrized by  $\tau$ . Then the entropy  $s$  is a strictly decreasing function of  $\tau$  along the Hugoniot curve.*

This is the case when  $\frac{\partial \mathcal{H}}{\partial p} > 0$  and is indeed satisfied for a polytropic ideal gas (Example 2.1).

It remains to characterize the admissible shocks. We first notice that the sign of  $M$  enables us to recognize a 1-shock from a 3-shock. Indeed, since by (2.21), (2.24), and (2.25)

$$\frac{\partial \mathcal{H}}{\partial \tau}(\tau, p_0) = -T \left( \frac{\partial p}{\partial s} \right)^{-1} \frac{\partial p}{\partial \tau}(\tau, p_0) > 0,$$

the function  $\tau \mapsto \mathcal{H}(\tau, p_0)$  is strictly increasing, so that

$$M^2 = -\frac{p - p_0}{\tau - \tau_0}$$

cannot vanish when the point  $(\tau, p)$  varies on the Hugoniot curve with center  $A_0$ . Hence  $M$  keeps a constant sign on a shock curve. In order to determine the sign of  $M$ , we use the parametrization of a shock curve derived in Chap. II, Example 4.2. We have (by (4.35) and (4.36))

$$\tau = \tau_0 - \varepsilon \tau_0 + \mathcal{O}(\varepsilon^2)$$

and

$$u = u_0 - \varepsilon c_0 + \mathcal{O}(\varepsilon^2) \quad \text{for a 1-shock,}$$

$$u = u_0 + \varepsilon c_0 + \mathcal{O}(\varepsilon^2) \quad \text{for a 3-shock.}$$

Hence

$$M = \frac{u - u_0}{\tau - \tau_0} = \begin{cases} c_0 \rho_0 + \mathcal{O}(\varepsilon) & \text{for a 1-shock,} \\ -c_0 \rho_0 + \mathcal{O}(\varepsilon) & \text{for a 3-shock,} \end{cases}$$

so that

$$\begin{cases} M > 0 & \text{for a 1-shock,} \\ M < 0 & \text{for a 3-shock.} \end{cases} \quad (2.32)$$

Let us now introduce some notations. Since  $M \neq 0$ , it follows from (2.8) that the gas crosses a shock. Then the side of the shock front through which the gas enters is called the *front side* (or the side ahead of the shock front), while the other side is called the *back side* (or the side behind). This means that the gas crosses the shock front from the front toward the back side, and one can say that the shock faces the front side. Hereafter, we shall use the following convention: the state (0) will refer to the state of the gas at the *front side*, and the state (1) will refer to the state of the gas at the *back side*. In other words, we are in one of the two situations below (see Fig. 2.2). Thus

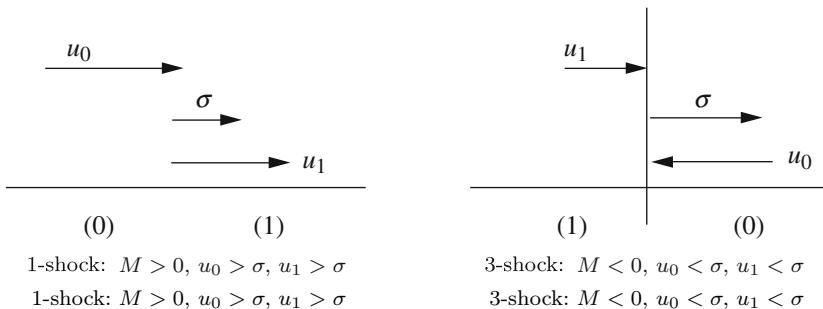
for a 1-shock the state (0) is the left state

while

for a 3-shock the state (1) is the left state.

The left (resp. right) state is usually noted with the subscript – or  $L$  (resp. + or  $R$ ).

*Remark 2.3.* The direction of propagation of the shock is given by the sign of  $\sigma$  and must be distinguished from the direction toward which the shock front faces, which depends on the sign of the relative velocities  $u_i - \sigma$ . Consider the situation illustrated in Fig. 2.2 (left): the fluid particles are crossing the



**Fig. 2.2** Relative velocities for a 1- and a 3-shock

shock front from the left to the right since  $u_0 > \sigma$ . Hence, the left state is indeed the front state. This argument is of course independent of whether the front advances or recedes, and the conclusion holds with a negative  $\sigma$ . Note also that we might have taken  $u_0 < u_1$ . We shall see, however, below that this is excluded for an admissible shock.  $\square$

Now, a shock discontinuity is said to be *admissible* if it satisfies the entropy jump condition (2.5). Moreover, a shock is called a *compressive shock* if it satisfies  $p_1 > p_0$ , i.e., if the pressure increases as the gas crosses the shock front.

*Lemma 2.5*

Assume the hypotheses of Corollary 2.1 of Theorem 2.1. Then, a shock is admissible if and only if it is a compressive shock.

*Proof.* Using (2.8), the entropy jump condition (2.5) can be written

$$M[s] = [\rho s(u - \sigma)] \geq 0$$

or, equivalently,

$$s_1 \geq s_0. \quad (2.33)$$

Indeed, when  $M$  is  $> 0$ , we have (see Fig. 2.2)

$$M[s] = M(s_+ - s_-) = M(s_1 - s_0) \geq 0.$$

On the other hand, when  $M$  is  $< 0$ , we obtain

$$M[s] = M(s_0 - s_1) \geq 0,$$

so that the entropy increases across a shock and our assertion (2.33) follows.

Now, assuming the hypotheses of Corollary 2.1, we know that along the Hugoniot curve with center  $A_0$ ,  $s$  is a strictly decreasing function of  $\tau$  and therefore a strictly increasing function of  $p$ , so that

$$s_1 \geq s_0 \iff p_1 \geq p_0$$

and the shock is indeed compressive.  $\square$

*Remark 2.4.* Note that the entropy jump condition  $M[s] \geq 0$  is obtained in a more direct way if we use the Lagrangian variables since  $-M$  is the shock velocity with respect to the mass variable (see Remark 2.1).  $\square$

As a simple consequence of Lemma 2.5, we obtain the following result.

**Theorem 2.2**

Assume the hypotheses of Corollary 2.1 of Theorem 2.1. A shock is admissible if and only if one of the four following equivalent properties holds:

$$\begin{cases} (i) & \rho_1 \geq \rho_0, \\ (ii) & p_1 \geq p_0, \\ (iii) & s_1 \geq s_0, \\ (iv) & u_- \geq u_+. \end{cases} \quad (2.34)$$

*Proof.* Since along the Hugoniot curve with center  $A_0$  we have

$$\rho_1 \geq \rho_0 \iff p_1 \geq p_0 \iff s_1 \geq s_0,$$

the first three characterizations of an admissible shock are an obvious re-statement of Lemma 2.5. They are also equivalent to  $\tau_1 \leq \tau_0$ . Let us check the fourth characterization (2.34). Using

$$\rho_0(u_0 - \sigma) = \rho_1(u_1 - \sigma) = M,$$

the condition  $\rho_1 \geq \rho_0$  gives for  $M > 0$

$$u_0 - \sigma \geq u_1 - \sigma$$

i.e., since the state (0) is the left state (see Fig. 2.2)

$$u_0 = u_- \geq u_+ = u_1,$$

and for  $M < 0$

$$u_0 - \sigma \leq u_1 - \sigma$$

i.e., (see Fig. 2.2 again)

$$u_1 = u_- \geq u_+ = u_0.$$

Hence, for an admissible shock, the density, pressure, and entropy increase when the gas crosses the shock front.  $\square$

**Theorem 2.3**

Assume the hypotheses of the Corollary 2.1 of Theorem 2.1. Then, a shock is admissible if and only if

$$c_0 \leq |u_0 - \sigma|, \quad |u_1 - \sigma| \leq c_1, \quad (2.35)$$

i.e., if and only if the gas velocity relative to the shock front is supersonic at the front side and subsonic at the back side.

*Proof.* Assume that  $A_1 = (\tau_1, p_1)$  belongs to the Hugoniot curve with center  $A_0$ , and consider the straight line  $\Delta$  joining the states  $A_0$  and  $A_1$ ,

parametrized by  $\alpha$  as in the proof of Theorem 2.1, with  $\alpha_1$  (corresponding to  $A_1$ )  $> 0$ , for instance. Since  $\mathcal{H}(\tau_0, p_0) = \mathcal{H}(\tau_1, p_1) = 0$ , along  $\Delta$  we have  $\frac{d\mathcal{H}}{d\alpha} = 0$  at some intermediate point  $A_2$  of  $A_0 A_1$  distinct from  $A_0$  and  $A_1$ . Then, it follows from (2.30) and (2.31) that along the line  $\Delta$ ,  $s$  has exactly one critical point  $A_2$ , which is a maximum. Hence, we obtain

$$\frac{ds}{d\alpha} > 0 \text{ at } A_0, \quad \frac{ds}{d\alpha} < 0 \text{ at } A_1.$$

Since

$$\begin{aligned} \frac{ds}{d\alpha} &= \frac{\partial s}{\partial \tau} \frac{d\tau}{d\alpha} + \frac{\partial s}{\partial p} \frac{dp}{d\alpha} \\ &= \frac{(\tau_1 - \tau_0)}{\alpha_1} \frac{\partial s}{\partial \tau} + \frac{(p_1 - p_0)}{\alpha_1} \frac{\partial s}{\partial p}, \end{aligned}$$

(2.25) and (1.4) yield

$$\frac{ds}{d\alpha} = \frac{1}{\alpha_1} \{ \rho^2 c^2 (\tau_1 - \tau_0) + p_1 - p_0 \} \frac{\partial s}{\partial p}.$$

We find thus at  $A_0$

$$\rho_0^2 c_0^2 (\tau_1 - \tau_0) + p_1 - p_0 > 0$$

and at  $A_1$

$$\rho_1^2 c_1^2 (\tau_1 - \tau_0) + p_1 - p_0 < 0.$$

Since  $\tau_1 < \tau_0$ , this gives

$$\rho_0^2 c_0^2 < -\frac{p_1 - p_0}{\tau_1 - \tau_0} < \rho_1^2 c_1^2, \quad (2.36)$$

and by (2.11) and (2.8)

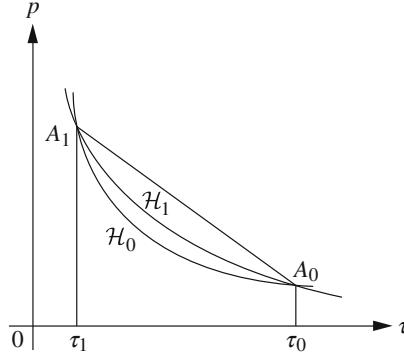
$$\rho_0^2 c_0^2 < \rho_0^2 v_0^2 = M^2 = \rho_1^2 v_1^2 < \rho_1^2 c_1^2.$$

This implies

$$c_0^2 < v_0^2, \quad v_1^2 < c_1^2,$$

which is equivalent to (2.35).  $\square$

*Remark 2.5.* For two given states, a front state  $A_0$  and a back state  $A_1$ , that can be connected by an admissible shock, the Hugoniot curves  $\mathcal{H}_0$  with center  $A_0$  and  $\mathcal{H}_1$  with center  $A_1$  intersect at points  $A_0$  and  $A_1$ , but they do not coincide (see Fig. 2.3). If the left state is  $A_0$ , the shock connecting  $A_0$  to  $A_1$  is a 1-shock (see Fig. 2.2), and a 3-shock if the left state is  $A_1$ .  $\square$



**Fig. 2.3**  $\mathcal{H}_0$  with center  $A_0$ ,  $\mathcal{H}_1$  with center  $A_1$

*Remark 2.6.* Let us prove that the inequalities (2.36) imply that a Rayleigh line through  $A_0$  intersects the Hugoniot curve at (at most) one other point.

First, we note that, due to assumption (2.28), we have by (2.29) (if the curve is parametrized by  $\tau$ )

$$p''(\tau_0) = \frac{\partial^2 p}{\partial \tau^2}(\tau_0, s_0) > 0, \quad (2.37)$$

and the curve is convex in a neighborhood of the center  $A_0$ .

Now, by (2.36) and (1.4), we have

$$-\frac{\partial p}{\partial \tau}(\tau_0, s_0) = \rho_0^2 c_0^2 < -\frac{p_1 - p_0}{\tau_1 - \tau_0} = M^2 < \rho_1^2 c_1^2 = -\frac{\partial p}{\partial \tau}(\tau_1, s_1),$$

which we write

$$\frac{\partial p}{\partial \tau}(\tau_1, s_1) < \frac{p_1 - p_0}{\tau_1 - \tau_0} < \frac{\partial p}{\partial \tau}(\tau_0, s_0). \quad (2.38)$$

Consider the Hugoniot curve  $\mathcal{H}$  with center  $A$ , and  $B$ ,  $C$  two points on  $\mathcal{H}$  with  $\tau_B < \tau_A$  and  $\tau_C > \tau_A$  (see Fig. 2.4). We first consider the left side of  $\mathcal{H}(\tau < \tau_A)$ . Since the state  $B$  is such that  $\tau_B < \tau_A$ , the shock connecting  $A$  and  $B$  is admissible if  $A$  is the front side,  $A = B_0 = (\tau_0, p_0)$ , and  $B$  the back side  $B = B_1 = (\tau_1, p_1)$ , so that  $\mathcal{H}$  is indeed the Hugoniot curve  $\mathcal{H}_0$  with center  $B_0 = A$ . When parametrized by  $\tau$ , it is given by the function  $\tau \mapsto p(\tau, s(\tau)) = p(\tau)$ . Therefore, by (2.29), the slope of the tangent is

$$p'(\tau) = \frac{\partial p}{\partial \tau} + \frac{\partial p}{\partial s} s'(\tau),$$

and

$$p'(\tau_0) = \frac{\partial p}{\partial \tau}(\tau_0, s_0).$$

Now, since  $s$  is a strictly decreasing function of  $\tau$ ,  $s' < 0$ , we have by (2.21)

$$p'(\tau_1) = \frac{\partial p}{\partial \tau}(\tau_1, s_1) + \frac{\partial p}{\partial s}(\tau_1, s_1)s'(\tau_1) < \frac{\partial p}{\partial \tau}(\tau_1, s_1). \quad (2.39)$$

Thus, in this case (2.38) implies

$$p'(\tau_1) < \frac{p_1 - p_0}{\tau_1 - \tau_0} < p'(\tau_0), \quad (2.40)$$

which proves that the line  $\Delta = A_0 A_1$  crosses the left part of the Hugoniot curve  $\{\tau < \tau_A\}$  of  $\mathcal{H}_0$  at the point  $A_1$  only.

Similarly, let us now consider the state  $C$  on  $\mathcal{H}$  such that  $\tau_C > \tau_A$ . The shock connecting  $A$  and  $C$  is admissible if  $C$  is the front side  $C = C_0 = (\tau_0, p_0)$ , and  $A = C_1 = (\tau_1, p_1)$ . So  $\mathcal{H}$  is the Hugoniot curve with center  $C_1$ , which we denote by  $\mathcal{H}_1$  and parametrize by  $\tau$ . To distinguish from the first case where the center corresponded to the index 0, we shall denote by a dot the differentiation along  $\mathcal{H}_1$ . We then have by (2.29)

$$\dot{p}(\tau_1) = \frac{\partial p}{\partial \tau}(\tau_1, s_1). \quad (2.41)$$

By using (2.27) applied to  $\mathcal{H}_1$ , we now observe that since  $s$  is a strictly decreasing function of  $\tau$ ,

$$T\dot{s}(\tau_0) = \frac{(p_0 - p_1)}{2} - \frac{(\tau_0 - \tau_1)}{2}\dot{p}(\tau_0) < 0.$$

Thus, for  $\tau_0 > \tau_1$

$$\frac{p_0 - p_1}{\tau_0 - \tau_1} < \dot{p}(\tau_0).$$

Together with (2.41) and (2.37), this yields

$$\dot{p}(\tau_1) < \frac{p_0 - p_1}{\tau_0 - \tau_1} < \dot{p}(\tau_0), \quad (2.42)$$

which proves the analogous property for the right part of the Hugoniot curve  $\mathcal{H}$ . Properties (2.37), (2.40), and (2.42) give the desired result.

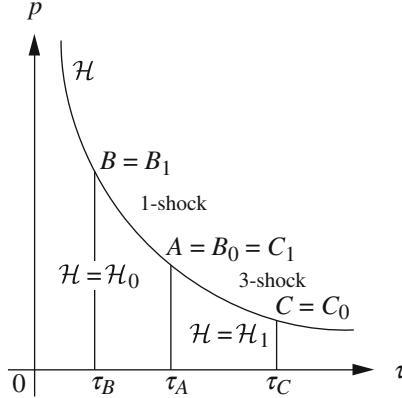
In Fig. 2.3, we had represented for two given states  $A_0$  (resp.  $A_1$ ) the part of the Hugoniot curve  $\mathcal{H}_0$  with center  $A_0$  such that  $\tau \leq \tau_0$  (resp.  $\mathcal{H}_1$  with center  $A_1$  such that  $\tau \geq \tau_1$ ). Note that the slope of the tangent to  $\mathcal{H}_0$  (resp.  $\mathcal{H}_1$ ) at point  $A_1$  is  $p'(\tau_1)$  (resp.  $\dot{p}(\tau_1)$ ) and by (2.39) and (2.41)

$$p'(\tau_1) < \dot{p}(\tau_1).$$

Similarly,

$$\dot{p}(\tau_0) = \frac{p_0 - p_1}{\tau_0 - \tau_1} - \frac{2T\dot{s}(\tau_0)}{\tau_0 - \tau_1} < p'(\tau_0),$$

so that  $\mathcal{H}_1$  lies above  $\mathcal{H}_0$ . □



**Fig. 2.4** Hugoniot curve  $\mathcal{H}$  with center  $A$

In the previous chapter, we introduced two different criteria for characterizing the admissible shocks, namely, the Lax entropy conditions (Chap. II, (5.14)) and the entropy condition based on a strictly convex entropy  $U$  (Chap. II, (5.18)). We have seen that these two criteria coincide for sufficiently weak shocks (see Chap. II, Theorem 5.3). Here, in the case of the gas dynamics equations, we want to check that the Lax entropy conditions coincide with the entropy condition (2.5) (or equivalently (2.33)) for all shocks.

*Theorem 2.4*

*Assume the hypotheses of the Corollary of Theorem 2.1. Then, for the gas dynamics equations, the Lax entropy inequalities are equivalent to the increase of the entropy across a shock.*

*Proof.* Consider a 1-shock. In Eulerian coordinates, the Lax entropy conditions give in this case

$$u_+ - c_+ < \sigma < u_+, \quad \sigma < u_- - c_-,$$

or equivalently in terms of the states (0) and (1) (see Fig. 2.2)

$$u_1 - c_1 < \sigma < u_1, \quad \sigma < u_0 - c_0.$$

These inequalities can be equivalently written

$$0 < u_1 - \sigma < c_1, \quad c_0 < u_0 - \sigma.$$

Since for a 1-shock we have  $u_i - \sigma > 0, i = 0, 1$ , Theorem 2.3 implies that the above inequalities characterize an admissible 1-shock. The case of a 3-shock is entirely similar. Note that since  $(\lambda - u)^2 = -\frac{1}{\rho^2} \frac{\partial p}{\partial \tau}$  and  $\rho_i(\sigma - u_i)^2 = -\frac{(p_1 - p_0)}{(\tau_1 - \tau_0)}$ , inequalities (2.38) also express the Lax entropy criteria.  $\square$

*Remark 2.7.* Again, one could have proven Theorem 2.4 by working in Lagrangian coordinates. Indeed, consider a 3-shock; the Lax entropy conditions give here

$$\frac{c_+}{\tau_+} < -M = \sigma_L, \quad 0 < -M = \sigma_L < \frac{c_-}{\tau_-},$$

or equivalently by (2.13)

$$c_0 < -(u_0 - \sigma), \quad 0 < -(u_1 - \sigma) < c_1.$$

Using once more Theorem 2.3, this characterizes a 3-shock.  $\square$

*Remark 2.8.* Smoller et al. [1067] have proven that for a polytropic ideal gas with  $1 < \gamma < \frac{5}{3}$ , rarefaction shocks (of moderate strength), i.e., shocks that violate the Lax entropy condition (with  $|p_+ - p_-|$  not too large), are unstable in the class of smooth solutions. This means that there exists a sequence of  $C^2$  solutions (defined uniformly on  $\mathbb{R} \times [0, T]$  for some  $T > 0$ ) that converges in every  $\mathbf{L}^p (p \geq 1)$  to the given discontinuous data at  $t = 0$  but does not converge to the given (rarefaction) shock for any  $t \in ]0, T]$ .

Thus, stability with respect to smoothing appears as another criterion for selecting “admissible” weak shocks (see Chap. II, Sect. 5.3.2).  $\square$

### 3 Solution of the Riemann Problem

In this section, we want to solve the Riemann problem for the gas dynamics equations either in Eulerian coordinates or in Lagrangian coordinates. In fact, it will be convenient in all the following to characterize the state of the gas by the three dependent variables  $(\rho, u, p)$  (primitive variables). We look for an entropy solution of the system of equations (2.1), (2.2) satisfying the initial condition

$$(\rho, u, p)(x, 0) = \begin{cases} (\rho_L, u_L, p_L), & x < 0, \\ (\rho_R, u_R, p_R), & x > 0. \end{cases} \quad (3.1)$$

Now, it follows from the general theory of Chap. II, Sect. 6 that (at least for sufficiently close left and right states) we can find an entropy solution of the Riemann problem of the following form: the left state  $(\rho_L, u_L, p_L)$  is connected to the right state  $(\rho_R, u_R, p_R)$  by a 1-shock or a 1-rarefaction wave, a 2-contact discontinuity, and a 3-shock or a 3-rarefaction wave. The 2-contact discontinuity separates two constant states  $(\rho_I, u^*, p^*)$  and  $(\rho_{II}, u^*, p^*)$  so that  $u$  and  $p$  are continuous across the contact discontinuity. In Fig. 3.1, the 1-wave is a rarefaction and the 3-wave a shock. Then, it is adequate to work in the  $(u, p)$ -plane in order to determine the types of the 1- and 3-waves (shock or rarefaction) and the unknown variables  $u^*$ ,  $p^*$  at the contact discontinuity. It will be an easy matter afterward to compute  $\rho_I$  and  $\rho_{II}$ .

Let us first determine the projections of the shock curves onto the  $(u, p)$ -plane. We consider two states **a** and **b** connected by a shock wave. Recall that by (2.11) and (2.17)

$$M = -\frac{p_a - p_b}{u_a - u_b},$$

$$M^2 = -\frac{p_a - p_b}{\tau_a - \tau_b},$$

and by (2.12)

$$\mathcal{H}_a(\tau_b, p_b) = 0,$$

where

$$\mathcal{H}_a(\tau, p) = \varepsilon(\tau, p) - \varepsilon(\tau_a, p_a) + \frac{1}{2}(p + p_a)(\tau - \tau_a).$$

Assume that the Hugoniot curve  $\mathcal{H}_a(\tau, p) = 0$  may be parametrized by  $p \in [0, +\infty[$ , i.e., may be represented by an equation of the form

$$\tau = h_a(p), \quad \tau_a = h_a(p_a) \tag{3.2}$$

with the hypotheses

$$\lim_{p \mapsto 0} h_a(p) = \tau_{\max}, \quad \lim_{p \mapsto +\infty} h_a(p) = \tau_{\min},$$

and

$$h'_a(p) < 0, \quad \lim_{p \mapsto +\infty} \sqrt{p} h'_a(p) = 0. \tag{3.3}$$

*Remark 3.1.* Let us assume that the Hugoniot curve may be parametrized by  $\tau$  and that  $p(\tau) \mapsto +\infty$  when  $\tau \mapsto \tau_{\min}$ . Let us note that properties (3.3) hold if we assume that the Hugoniot curve is convex (decreasing). Indeed,  $h_a$  is also convex decreasing,  $h_a(p) \mapsto \tau_{\min}$  as  $p \mapsto +\infty$ , and due to the convexity of  $h_a$ , we have for any  $p_0$ ,

$$0 \leq -h'_a(p) \leq \frac{h_a(p_0) - h_a(p)}{p - p_0},$$

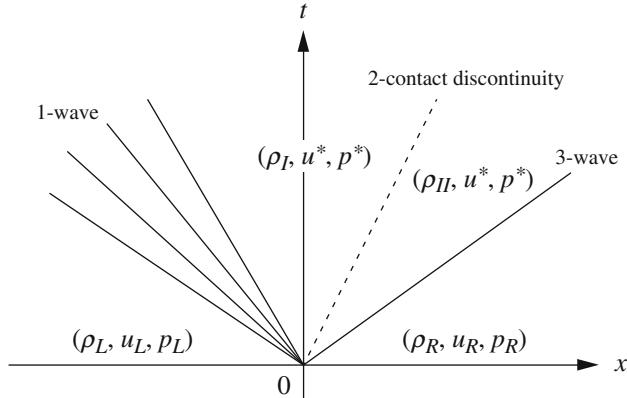
which implies the second inequality in (3.3).  $\square$

Note also that we have  $\tau_b = h_a(p_b)$ . Next, setting

$$M_a(p) = \begin{cases} \sqrt{(p - p_a)/(\tau_a - h_a(p))} & \text{for a 1-shock,} \\ -\sqrt{(p - p_a)/(\tau_a - h_a(p))} & \text{for a 3-shock,} \end{cases}$$

we have

$$u_b = u_a - \frac{p_b - p_a}{M_a(p_b)}.$$



**Fig. 3.1** Solution of the Riemann problem in the  $(x, t)$ -plane: case of a 1-rarefaction and 3-shock

Hence, defining the function

$$\Phi_a(p) = \frac{p - p_a}{|M_a(p)|} = (p - p_a) \sqrt{(\tau_a - h_a(p))/(p - p_a)},$$

or equivalently

$$\Phi_a(p) = \begin{cases} \sqrt{(p - p_a)(\tau_a - h_a(p))}, & p \geq p_a, \\ -\sqrt{(p - p_a)(\tau_a - h_a(p))}, & p \leq p_a, \end{cases} \quad (3.4)$$

we obtain

$$u_b = \begin{cases} u_a - \Phi_a(p_b) & \text{for a 1-shock,} \\ u_a + \Phi_a(p_b) & \text{for a 3-shock.} \end{cases} \quad (3.5)$$

Let us make out a list of the properties of the function  $\Phi_a$ . First, obviously

$$\Phi_a(p_a) = 0.$$

Then, since

$$\lim_{p \rightarrow 0} h_a(p) = \tau_{\max}, \quad \lim_{p \rightarrow +\infty} h_a(p) = \tau_{\min},$$

we have on the one hand

$$\Phi_a(0) = -\sqrt{p_a/(\tau_{\max} - \tau_a)}, \quad (3.6a)$$

$$\lim_{p \rightarrow +\infty} \Phi_a(p) = +\infty \quad (3.6b)$$

and, on the other hand,

$$\Phi'_a(p) = \frac{\tau_a - h_a(p) - (p - p_a)h'_a(p)}{2\Phi_a(p)}$$

so that by (3.3)

$$\Phi'_a(p) > 0, \quad (3.6c)$$

$$\lim_{p \rightarrow +\infty} \Phi'_a(p) = 0. \quad (3.6d)$$

*Example 3.1.* Consider again the polytropic ideal gas introduced in Example 2.1, for which

$$\mathcal{H}_a(\tau, p) = \frac{(\tau - \mu^2 \tau_a)p - (\tau_a - \mu^2 \tau)p_a}{2\mu^2},$$

with  $\mu^2 = \frac{(\gamma-1)}{(\gamma+1)}$  given by (2.19) and  $\tau_{\min} = \mu^2 \tau_a$ ,  $\tau_{\max} = \mu^{-2} \tau_a$ . We have

$$h_a(p) = \frac{\tau_a(\mu^2 p + p_a)}{(p + \mu^2 p_a)},$$

which implies

$$\Phi_a(p) = (p - p_a) \sqrt{(1 - \mu^2)\tau_a / (p + \mu^2 p_a)}.$$

Clearly, the function  $h_a$  satisfies the assumptions (3.3), and the properties (3.6) hold. Note also that

$$\Phi_a(0) = -\sqrt{p_a \tau_a (1 - \mu^2) / \mu^2} = -\sqrt{2p_a \tau_a / (\gamma - 1)},$$

and since  $c^2 = \gamma p \tau$

$$\Phi_a(0) = -c_a \sqrt{2/\gamma(\gamma - 1)}.$$

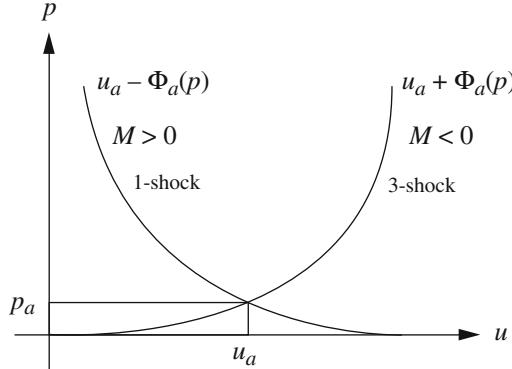
These results extend easily to a stiffened equation of state of Grüneisen type (Example 2.2).  $\square$

Let us consider the *shock curves*. The function

$$u = u_a \pm \Phi_a(p)$$

represents the states in the  $(u, p)$ -plane that can be connected to the state **a** on the right or on the left by a 1-shock or a 3-shock (see Fig. 3.2). Now, let us select the *right* states that can be connected to the state **a** by an *admissible shock*. We use the characterization (2.34)(iv) of an admissible shock  $u_- > u_+$ . Since in that case  $u_a = u_-$ , we obtain diagram (i) of Fig. 3.3.

Similarly, in order to determine the *left* states to which the state **a** can be connected by an admissible shock, we notice that  $u_a = u_+$ , so that we get diagram (ii) of Fig. 3.3.



**Fig. 3.2** Right and left states that can be connected to **a** by a shock

Let us next determine the projections of the *rarefaction curves* onto the  $(u, p)$ -plane. Defining a function  $\ell = \ell(\rho, s)$  up to an additive constant by

$$\frac{\partial \ell}{\partial \rho}(\rho, s) = \frac{c(\rho, s)}{\rho},$$

we recall that we can choose  $\{u - \ell, s\}$  as a pair of 1-Riemann invariants and  $\{u + \ell, s\}$  as a pair of 3-Riemann invariants (see Chap. II, Examples 3.2 and 3.3). In the following, it will be more convenient to use  $p, s$  as the thermodynamic independent variables. And we may equivalently define  $\ell = \ell(p, s)$  by

$$\frac{\partial \ell}{\partial p}(p, s) = \frac{1}{(\rho c)}(p, s) \quad (3.7)$$

since

$$\frac{c}{\rho} = \frac{\partial \ell}{\partial \rho}(p, s) = \frac{\partial \ell}{\partial p}(p, s) \frac{\partial p}{\partial \rho}(\rho, s) = \frac{\partial \ell}{\partial p}(p, s) c^2.$$

Now, let us consider two states **a** and **b** connected by a 1-rarefaction wave. By Theorem 3.2 of Chap. II, the 1-Riemann invariants are constant through a 1-rarefaction wave, so we have

$$\begin{cases} s_a = s_b, \\ u_a + \ell(p_a, s_a) = u_b + \ell(p_b, s_b). \end{cases} \quad (3.8)$$

Hence, setting

$$\Psi_a(p) = \ell(p, s_a) - \ell(p_a, s_a) \quad (3.9)$$

gives

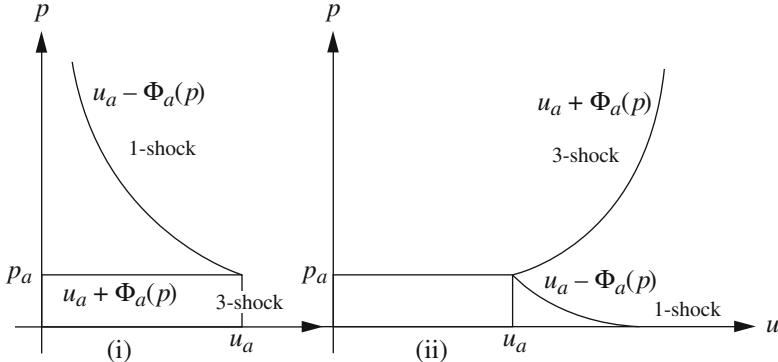
$$u_b = u_a - \Psi_a(p_b).$$

Similarly, if the states **a** and **b** are connected by a 3-rarefaction wave, we have

$$\begin{cases} s_a = s_b, \\ u_a - \ell(p_a, s_a) = u_b - \ell(p_b, s_b), \end{cases}$$

so that

$$u_b = u_a + \Psi_a(p_b).$$



**Fig. 3.3** (i) Right and (ii) left states that can be connected to  $\mathbf{a}$  by a shock

Thus, we obtain

$$u_b = \begin{cases} u_a - \Psi_a(p_b) & \text{for a 1-rarefaction,} \\ u_a + \Psi_a(p_b) & \text{for a 3-rarefaction.} \end{cases} \quad (3.10)$$

Let us study the properties of the function  $\Psi_a$ . Due to (2.30), (3.9) can be written

$$\Psi_a(p) = \int_{p_a}^p \left( \frac{1}{\rho c} \right) dp.$$

Since (see (2.30) of Chap. II)

$$\rho c = \sqrt{-\partial p / \partial \tau(\tau, s)}$$

and  $s$  is constant, we have

$$dp = \frac{\partial p}{\partial \tau} d\tau = -\rho^2 c^2 d\tau,$$

which yields

$$\Psi_a(p) = - \int_{\tau_a}^{\tau} \sqrt{-\partial p / \partial \tau} d\tau. \quad (3.11)$$

We note that

$$\Psi_a(p_b) = -\Psi_b(p_a), \quad (3.11a)$$

$$\Psi_a(0) = - \int_{\tau_a}^{+\infty} \sqrt{-\partial p / \partial \tau} d\tau < 0, \quad (3.11b)$$

and

$$\Psi'_a(p) = \frac{1}{\rho c} > 0. \quad (3.11c)$$

Now, we assume that the function  $p = p(\tau, s)$  satisfies

$$\left| \frac{\partial p}{\partial \tau}(\tau, s) \right| \geq \frac{c}{\tau^\alpha}, \quad \alpha \geq 2. \quad (3.12)$$

This implies for  $p > p_a$

$$\Psi_a(p) = \int_\tau^{\tau_a} \sqrt{-\partial p / \partial \tau} d\tau \geq c \int_\tau^{\tau_a} \left( \frac{1}{\tau^{\alpha/2}} \right) d\tau,$$

hence

$$\lim_{p \rightarrow +\infty} \Psi_a(p) = +\infty \quad (3.11d)$$

and also

$$\lim_{p \rightarrow +\infty} \Psi'_a(p) = 0. \quad (3.11e)$$

*Example 3.2.* Let us go back to the case of a polytropic ideal gas (Example 2.1) for which

$$p = p(\rho, s) = A(s)\rho^\gamma, \quad \gamma > 1.$$

We have

$$c^2 = \frac{\partial p}{\partial \rho} = \gamma A(s)\rho^{\gamma-1},$$

so that

$$\ell = \int \frac{c}{\rho} d\rho = \sqrt{\gamma A(s)} \int \rho^{(\gamma-3)/2} d\rho = \frac{2}{(\gamma-1)} \sqrt{\gamma A(s)} \rho^{(\gamma-1)/2}.$$

This yields

$$\ell = \frac{2}{\gamma-1} \sqrt{\gamma p \tau} = \frac{2c}{\gamma-1}$$

and

$$\Psi_a(p) = \frac{2\sqrt{\gamma}}{(\gamma-1)} (\sqrt{p\tau} - \sqrt{p_a\tau_a}).$$

It remains to express  $\tau$  in terms of  $p$ . Since  $s_a = s_b$ , we can write

$$\frac{p}{p_a} = \left( \frac{\rho}{\rho_a} \right)^\gamma$$

or

$$\tau = \tau_a \left( \frac{p_a}{p} \right)^{1/\gamma}.$$

Hence, we find

$$\Psi_a(p) = \frac{2\sqrt{\gamma}}{(\gamma-1)} p_a^{1/2\gamma} \tau_a^{1/2} \{ p^{(\gamma-1)/2\gamma} - p_a^{(\gamma-1)/2\gamma} \}.$$

Note that (3.12) holds with  $\alpha = \gamma + 1$ , which implies the properties (3.11). Moreover,

$$\Psi_a(0) = -2\sqrt{\gamma p_a \tau_a}/(\gamma-1) = -\frac{2c_a}{\gamma-1}.$$

The calculations are simple in this model case.  $\square$

The rarefaction curves defined by

$$u = u_a \pm \Psi_a(p)$$

represent the states in the  $(u, p)$ -plane that can be connected to the state **a** (on the right or on the left) by a 1-rarefaction or a 3-rarefaction wave (see Fig. 3.4).

It remains to select the states that can be connected to the state **a** on the right (resp. on the left) by a rarefaction wave. We first state the following lemma.

*Lemma 3.1*

*If a left state (with velocity  $u_-$ ) is connected to a right state (with velocity  $u_+$ ) by a rarefaction wave, we have*

$$u_+ > u_-. \quad (3.13)$$

*Proof.* Consider first a 1-rarefaction wave connecting a state **u** to the state **a**; we have by (3.8)

$$u + \ell = u_a + \ell_a, \quad s = s_a.$$

Hence, throughout the wave, we have

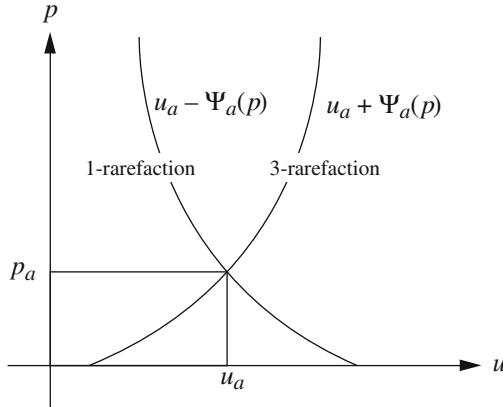
$$du = -d\ell \quad \text{and} \quad ds = 0.$$

Moreover,

$$\frac{d(u-c)}{du} = \frac{(d\ell + dc)}{d\ell} = \frac{\left( (c/\rho)d\rho + dc \right)}{\left( (c/\rho)d\rho \right)} = \frac{d(\rho c)}{cd\rho}.$$

But since

$$\rho c = \sqrt{-\partial p / \partial \tau},$$



**Fig. 3.4** Right and left states that can be connected to  $\mathbf{a}$  by a rarefaction wave

we have

$$d(\rho c) = -\frac{\partial^2 p / \partial \tau^2}{2\sqrt{-\partial p / \partial \tau}} = \left(\frac{1}{2}\right)\rho^3 c \left(\frac{\partial^2 p}{\partial \tau^2}\right) d\rho$$

and

$$\frac{d(\rho c)}{cd\rho} > 0.$$

Hence, throughout the 1-rarefaction wave, we obtain

$$\frac{d(u - c)}{du} > 0,$$

so that  $u - c$  increases with  $u$ . But for a 1-rarefaction wave,  $\lambda_1$  increases from the left to the right and (with shorthand notations for  $\lambda_1(\mathbf{U})$ )

$$u_+ - c_+ = \lambda_1(u_+) > \lambda_1(u_-) = u_- - c_-,$$

which implies (3.13).

Now, for a 3-rarefaction wave, we have

$$u - \ell = u_a - \ell_a, \quad s = s_a,$$

so that  $du = d\ell$  throughout the wave and

$$\frac{d(u + c)}{du} = \frac{(d\ell + dc)}{d\ell} = \frac{d(\rho c)}{cd\rho} > 0.$$

Hence  $u + c$  increases with  $u$ . Since

$$\lambda_3(u_+) = u_+ + c_+ > u_- + c_- = \lambda_3(u_-),$$

we obtain again (3.13). □

*Remark 3.2.* The result of Lemma 3.1 is natural if we recall the characterization (2.34) (iv) of an admissible shock. Moreover, since in a rarefaction wave

$$dp = c^2 d\rho, \quad d\ell = \frac{c}{\rho} d\rho,$$

we have

$$\begin{aligned} \rho_+ < \rho_-, \quad p_+ < p_- &\quad \text{for a 1-rarefaction wave,} \\ \rho_+ > \rho_-, \quad p_+ > p_- &\quad \text{for a 3-rarefaction wave,} \end{aligned}$$

which gives the analog of (2.34).  $\square$

We use the criterion (3.13), which characterizes a rarefaction wave, for obtaining the diagrams of Fig. 3.5.

We are now in a position to solve the Riemann problem.

*Theorem 3.1*

Assume that the function  $p(\tau, s)$  satisfies the conditions

$$\frac{\partial p}{\partial \tau} \leq -\frac{c}{\tau^\alpha}, \quad \alpha > 2, \quad \frac{\partial p}{\partial s} > 0, \quad \frac{\partial^2 p}{\partial \tau^2} > 0, \quad (3.14)$$

and that the Hugoniot curve may be parametrized as in (3.2) and (3.3). Then, the Riemann problem for the gas dynamics equations in Eulerian coordinates has a unique solution (in the class of admissible shock, contact discontinuities, and rarefaction waves separating constant states) if and only if the initial states satisfy the condition

$$u_R - u_L < -(\Psi_R(0) + \Psi_L(0)). \quad (3.15)$$

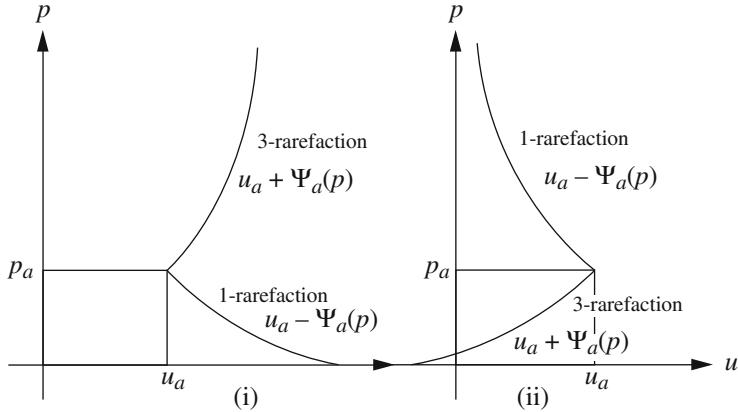
*Proof.* Let us denote by  $L$  (resp.  $R$ ) the given left (resp. right) states. Consider the set of right states that can be connected to  $L$  by a 1-wave (shock or rarefaction). It can be defined by

$$u = \begin{cases} u_L - \Phi_L(p), & p \geq p_L \quad (\mathcal{S}_1^L), \\ u_L - \Psi_L(p), & p \leq p_L \quad (\mathcal{R}_1^L). \end{cases} \quad (3.16)$$

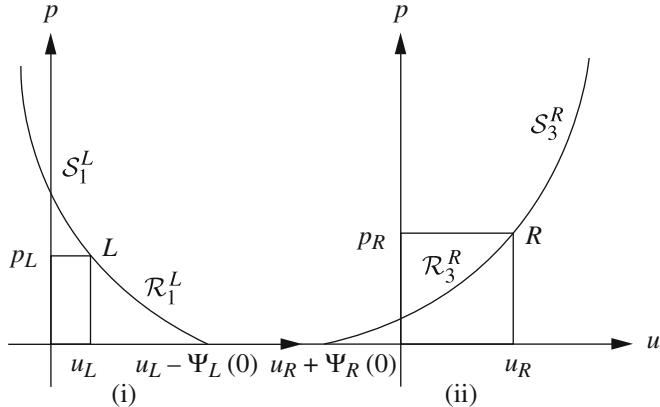
We have already observed that this curve is of class  $C^2$  and is strictly decreasing from  $u_L - \Psi_L(0)$  to  $-\infty$  when  $p$  increases from 0 to  $+\infty$  (Fig. 3.6 (i)).

Similarly, the set of left states to which  $R$  can be connected by a 3-wave (shock or rarefaction) is defined by

$$u = \begin{cases} u_R + \Phi_R(p), & p \geq p_R \quad (\mathcal{S}_3^R), \\ u_R + \Psi_R(p), & p \leq p_R \quad (\mathcal{R}_3^R). \end{cases}$$



**Fig. 3.5** (i) Right and (ii) left states that can be connected to  $\mathbf{a}$  by a rarefaction wave



**Fig. 3.6** Right states that can be connected to  $L$  by a 1-wave. Left states to which  $R$  can be connected by a 3-wave

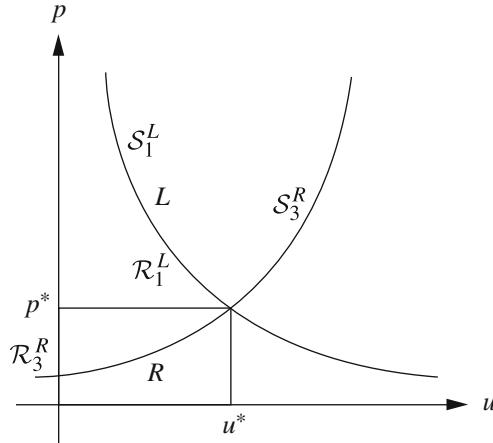
This function is of class  $C^2$  and is strictly increasing from  $u_R + \Psi_R(0)$  to  $+\infty$  when  $p$  increases from 0 to  $+\infty$  (Fig. 3.6 (ii)).

In order to solve the Riemann problem, we have to find the intersection of these two curves so as to determine  $(u^*, p^*)$ .

Geometrically (see Fig. 3.6), it is clear that a necessary and sufficient condition for  $(u^*, p^*)$  to be uniquely defined is given by

$$u_R + \Psi_R(0) < u_L - \Psi_L(0)$$

or, equivalently, by (3.15). Moreover, in that case, the nature of the 1-wave and the 3-wave is easily determined (a 1-rarefaction wave and a 3-shock wave in the case of Figs. 3.1 and 3.7).



**Fig. 3.7** Intersection of the 1-wave ( $L$ ) and 3-wave ( $R$ ) curves

It remains to compute the densities  $\rho_I$  and  $\rho_{II}$  of the constant states  $I$  and  $II$  located on both sides of the contact discontinuity. Let us determine  $\rho_I$ . Assume first that the 1-wave is a shock wave. Then, we have by (3.2)

$$\tau_I = \frac{1}{\rho_I} = h_L(p^*).$$

On the other hand, if the 1-wave is a rarefaction wave, using the function  $\rho = \rho(p, s)$ , since in that case  $s^* = s_L$ , we obtain

$$\rho_I = \rho(p^*, s_L).$$

Moreover, we have (see (2.30) in Example 2.3, Chap. II)

$$c_L^2 = \frac{\partial p}{\partial \rho}(\rho_L, s_L), \quad c_I^2 = \frac{\partial p}{\partial \rho}(\rho_I, s_L).$$

Remember that the 1-rarefaction wave is contained in the fan

$$\lambda_1(L) = u_L - c_L \leq \frac{x}{t} \leq u^* - c_I = \lambda_1(I).$$

We proceed in an analogous way for the 3-wave. □

*Example 3.3.* Let us go back to the case of a polytropic ideal gas (see Example 3.2). The condition (3.15) can be written

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

If the 1-wave is a shock wave, we obtain

$$\tau_I = \tau_L \frac{\mu^2 p^* + p_L}{p^* + \mu^2 p_L}.$$

On the other hand, if the 1-wave is a rarefaction wave, we may write

$$p_L = A(s_L) \rho_L^\gamma, \quad p^* = A(s_L) \rho_I^\gamma,$$

so that

$$\rho_I = \left( \frac{p^*}{p_L} \right)^{1/\gamma} \rho_L.$$

Again, in this model case, the calculations are simple.  $\square$

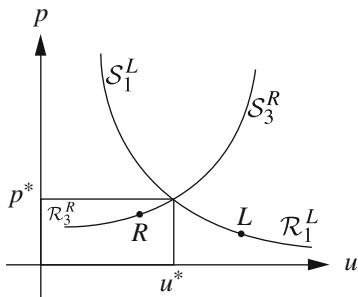
In fact, we have four cases for the solution of the Riemann problem, depicted in Fig. 3.8.

*Remark 3.3.* If the initial states  $L$  and  $R$  do not satisfy the condition (3.15), the Riemann problem has no solution in the above sense. However, one can yet define a solution by introducing a vacuum. The solution consists of two rarefaction waves separated by a vacuum where  $\rho = 0$  and the other dependent variables are left undefined (see Fig. 3.9). Then Theorem 3.1 is a global existence theorem in that the two initial states are not required to be close to each other.  $\square$

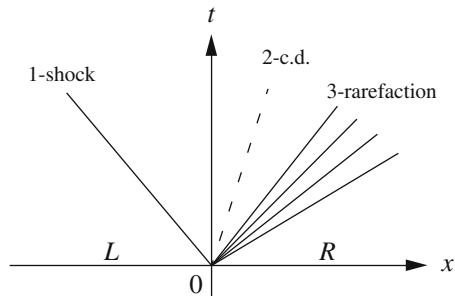
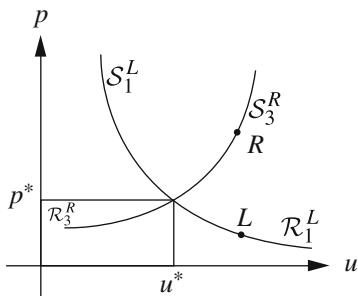
*Example 3.4.* As an example, we consider the case of a shock tube that corresponds to the initial condition of a gas at rest with two different states on each side of  $x = 0$ :  $u_L = u_R = 0$ ,  $\rho_L > \rho_R$ ,  $p_L > p_R$ . The associated Riemann problem is easily solved by looking at Fig. 3.8 (we are in case 3 of Fig. 3.8). Now, the values of  $\rho$ ,  $u$ , and  $p$  are as depicted in Fig. 3.10, which gives the behavior but not the precise values of the density, velocity, and pressure. This shock tube problem serves very frequently as a test for numerical schemes [1068], and more precise values can be found in many references (Einfeldt [457], for instance). Other tests corresponding to different initial conditions are often proposed (see, e.g., Montagné et al. [873]). One also finds the example of two mirror states ( $\mathbf{U}_L = (\rho, -u, p)$  and  $\mathbf{U}_R = (\rho, u, p)$ ) [458], which corresponds to case 4 if  $u > 0$  or case 1 if  $u < 0$ , with a 2-contact discontinuity on  $x = 0$ .  $\square$

We shall not detail the solution of the Riemann problem in Lagrangian coordinates: it has exactly the same structure as in Eulerian coordinates. The jump conditions are given in (2.15) (see Remark 2.1). If  $x$  is the mass variable, in the  $(x, t)$  plane the 2-contact discontinuity is the axis  $x = 0$ , across which  $u$  and  $p$  are constant.

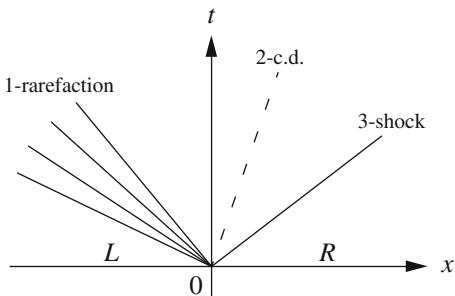
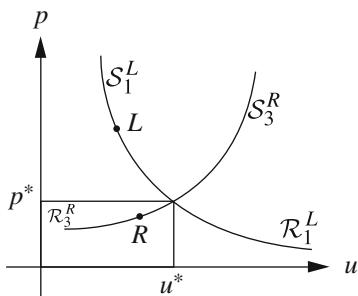
1st case: two shocks



2nd case: 1-shock and 3-rarefaction



3rd case: 1-rarefaction and 3-shock



4th case: two rarefactions

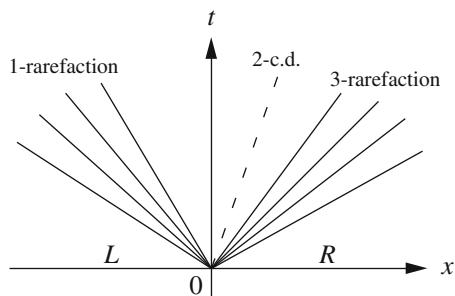
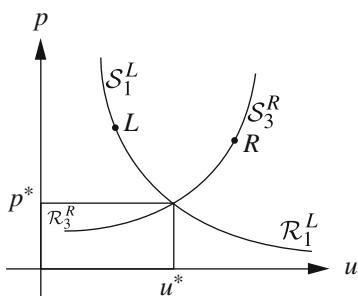


Fig. 3.8 Solution of the Riemann problem

*Remark 3.4.* Note the following property: on a 1-rarefaction wave, the  $C_-$ -characteristics are the straight lines

$$\frac{x}{t} = -\frac{c}{\tau} = -g$$

(see Chap. II, (5.6)). Therefore, the  $C_+$  characteristics on this 1-rarefaction wave satisfy at each point  $(x, t)$

$$\frac{dx}{dt} = -\frac{x}{t}$$

which yields

$$xt = \text{constant},$$

and the cross characteristics  $C_+$  are thus hyperbolas (we have an analogous property for a 3-rarefaction). This property can be used to parametrize the rarefaction wave by characteristic coordinates  $(\alpha, \beta)$  defined by

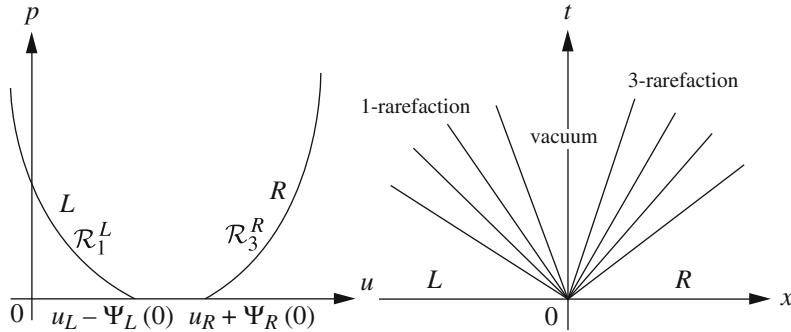
$$\alpha = -(-g_L xt)^{1/2}, \quad \beta = -g_L^{-1} \frac{x}{t}.$$

$\beta$  is a normalized slope of the  $C_-$  line, while  $\alpha$  is the  $x$ -coordinate of the intersection of the  $C_+$ -curve with the head characteristic (the slopes of the rarefaction fan extend from  $-g_L$  to  $-g_L^*$ , so that  $\beta = 1$  at the head, and  $\beta = \beta^* = \frac{g_L^*}{g_L}$  at the tail) (see Ben Artzi and Falcovitz [103] for details).  $\square$

*Remark 3.5.* We can also solve the Riemann problem for a flow consisting of a mixture of two components, for example, reacting species; see Example 2.6 in Chap. I. For simplicity, we consider two species. We add to the system of three equations (2.1) (which now states conservation of mass, momentum, and energy for the mixture) a fourth equation corresponding to the conservation of one species with mass fraction  $Y$  (the mass fraction of the other species is thus  $1 - Y$ )

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = 0, \\ \frac{\partial}{\partial t}(\rho Y) + \frac{\partial}{\partial x}(\rho Yu) = 0. \end{cases}$$

In fact, for reacting species, the full system contains source terms from chemical production that we have not written in the right-hand side. For simplicity, we also assume for each species a perfect gas  $\gamma$ -law  $p_i = (\gamma_i - 1)\rho Y_i \varepsilon_i$ ,  $\varepsilon_i = C_{v_i} T$ , and Dalton's law for the pressure of the mixture  $p = p_1 + p_2$  (see Remark 1.2). The total energy of the mixture is  $\rho e = \rho Y \varepsilon_1 + \rho(1 - Y) \varepsilon_2 + \frac{1}{2} \rho u^2$ . We have  $p = (\gamma - 1)(\rho e - \frac{1}{2} \rho u^2)$ , where we find that  $\gamma$  is equal to



**Fig. 3.9** Vacuum

$$\gamma = \frac{YC_{v_1}\gamma_1 + (1-Y)C_{v_2}\gamma_2}{YC_{v_1} + (1-Y)C_{v_2}} = \frac{YC_{p_1} + (1-Y)C_{p_2}}{YC_{v_1} + (1-Y)C_{v_2}}.$$

We obtain a system in conservative variables  $(\rho, \rho u, \rho e, \rho Y)$  that is not strictly hyperbolic, with eigenvalues  $u - c, u, u, u + c$ , where  $c^2 = \frac{\partial p(\rho, s, Y)}{\partial \rho} = \frac{\gamma p}{\rho}$ . The field corresponding to the double eigenvalue is linearly degenerate (see Remark 6.1, Chap. II), and the solution of the Riemann problem consists of four constant states separated by a 1-wave, a 2- (or 3-) contact discontinuity, and a 4-wave, as depicted in Fig. 3.11.

It is worthwhile to note that the mass fraction  $Y$  is only discontinuous across a contact discontinuity (with the notations of Fig. 3.11,  $Y_I = Y_L$  and  $Y_{II} = Y_R$ ). Indeed,  $Y$  is a 1- and 4-Riemann invariant and is thus constant on a rarefaction wave. Across a shock, it follows from the Rankine-Hugoniot jump conditions (2.3) together with  $\sigma[\rho Y] = [\rho u Y]$  that  $[Y] = 0$  and  $Y$  is again constant. This justifies the fact that the last equation can also be written in nonconservative form

$$\frac{\partial Y}{\partial t} + u \frac{\partial Y}{\partial x} = 0,$$

which means that  $Y$  is purely convected by the flow. Indeed, since it is obtained by combining the first and fourth conservation equations, it holds for smooth solutions, while for a solution such that  $Y$  is discontinuous, the discontinuity must be a contact discontinuity across which  $u$  is constant, so that the nonconservative product  $u \frac{\partial Y}{\partial x}$  is well-defined.  $\square$

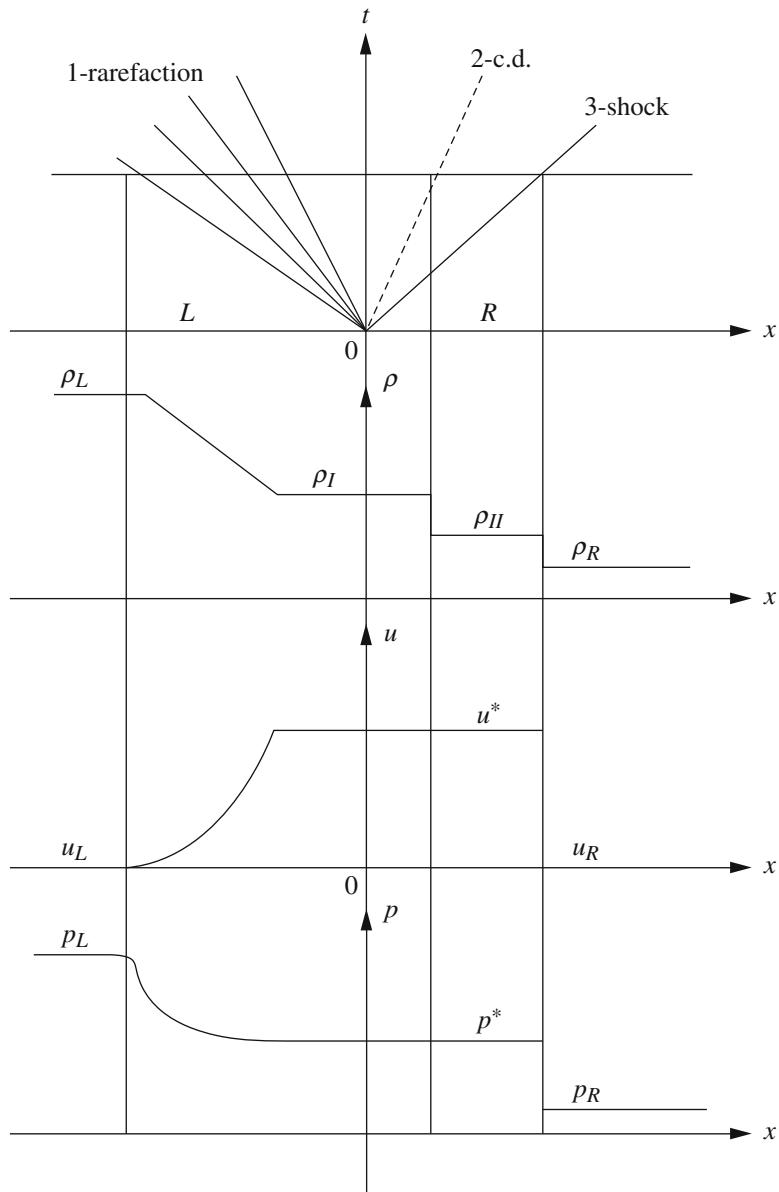
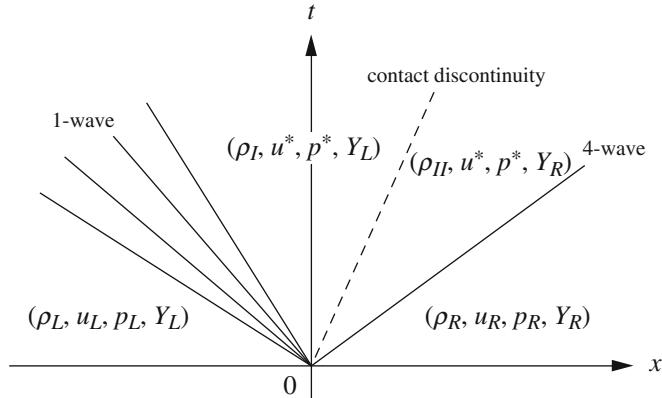


Fig. 3.10 Solution of the shock tube problem



**Fig. 3.11** Solution of the Riemann problem for a mixture of two species

## 4 Reacting Flows: The Chapman-Jouguet Theory

Let us come now to the more complex example of reacting gas flow (Example 2.6 in Chap. I). We restrict ourselves to the plane one-dimensional flow involving a single exothermic reaction between two species, the unburnt gas and the burnt gas. Neglecting transport effects, the equations to be considered are the Euler equations plus a chemical reaction equation

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = 0, \\ \frac{\partial}{\partial t}(\rho z) + \frac{\partial}{\partial x}(\rho zu) = \rho r, \end{cases} \quad (4.1)$$

where  $z$  denotes the mass fraction of the burnt gas (so that  $1 - z$  is the mass fraction of the unburnt gas) and  $r$  is the reaction rate, which we assume to be of the form

$$r = r(\rho, p, z).$$

Note that now we must include in the internal energy  $\varepsilon$  a term corresponding to the heat of reaction of the mixture.

*Example 4.1.* Assuming that both gases are ideal polytropic with the same  $\gamma$ -law and with constant energy of formation  $Q_u$  (resp.  $Q_b$ ) for the unburnt (resp. burnt) gas, we have

$$\varepsilon = \varepsilon(\rho, p, z) = (\gamma - 1)^{-1} \frac{p}{\rho} + (1 - z)Q_u + zQ_b.$$

For an exothermic reaction,  $Q_b < Q_u$ . □

Let us begin by studying the even simpler case of an *infinite reaction rate*. This amounts to supposing that the reaction is completed instantaneously, so that we replace the chemical reaction equation by the following description: the reacting gas consists of two components, the unburnt gas and the burnt gas, each in thermodynamic equilibrium, and separated by an infinitely thin reaction front. The equations of state of the unburnt gas and the burnt gas, which are denoted, respectively, by  $\varepsilon_u(\tau, p)$  and  $\varepsilon_b(\tau, p)$ , are assumed to be related by

$$\varepsilon_b(\tau, p) = \varepsilon_u(\tau, p) - Q, \quad (4.2)$$

where  $Q$  is the (constant) heat of complete reaction. In the case of an exothermic reaction, which we shall study later, we have

$$Q > 0.$$

In this model, we thus consider the usual gas dynamics equations, but with different equations of state for the unburnt and the burnt components of the gas. We now investigate the possible elementary waves that are allowed by this model. Indeed, in addition to the shock waves, the rarefaction waves, and the contact discontinuities, we shall have to introduce discontinuity waves that separate the unburnt gas from the burnt gas. These discontinuities will be referred to in the following as *combustion waves*.

If the reaction propagates with speed  $\sigma$ , the equations relating the properties on each side of the front are derived in the same way as the jump conditions for a discontinuity wave as in Chap. I, Sect. 4.2:

$$\begin{cases} \sigma[\rho] = [\rho u], \\ \sigma[\rho u] = [\rho u^2 + p], \\ \sigma[\rho e] = [(\rho e + p)u]. \end{cases} \quad (4.3)$$

We perform now the same analysis as in Sect. 2, with the only difference (with equations (2.3)) that, in the last relation (4.3), the equation of state differs on each side of the front.

Let us introduce the following convention: the state (0) will refer to the *unburnt* gas, while the state (1) refers to the *burnt* gas. As before, we denote the velocity of the gases relative to the front by

$$v_i = u_i - \sigma, \quad i = 0, 1.$$

Then, relations (4.3) give after a computation similar to that of Lemma 2.1

$$\begin{cases} \rho_0 v_0 = \rho_1 v_1, \\ \rho_0 v_0^2 + p_0 = \rho_1 v_1^2 + p_1, \\ \left( \rho_0 \left( \varepsilon_0 + \frac{v_0^2}{2} \right) + p_0 \right) v_0 = \left( \rho_1 \left( \varepsilon_1 + \frac{v_1^2}{2} \right) + p_1 \right) v_1. \end{cases} \quad (4.4)$$

Next, we set

$$M = \rho_0 v_0 = \rho_1 v_1$$

so that the second equation in (4.4) yields

$$M v_0 + p_0 = M v_1 + p_1$$

or

$$M = -\frac{p_1 - p_0}{v_1 - v_0} = -\frac{p_1 - p_0}{u_1 - u_0}. \quad (4.5)$$

Also, since  $v_i = M\tau_i$ , we obtain

$$M^2 = -\frac{p_1 - p_0}{\tau_1 - \tau_0}. \quad (4.6)$$

Now, eliminating the velocities in the third equation (4.2) gives

$$\varepsilon_1 - \varepsilon_0 + \frac{1}{2}(p_1 + p_0)(\tau_1 - \tau_0) = 0. \quad (4.7)$$

This is the analog of the Hugoniot equation (2.12). Again, it is worthwhile to notice that  $\varepsilon_0 = \varepsilon_u(\tau_0, p_0)$  and  $\varepsilon_1 = \varepsilon_b(\tau_1, p_1)$  correspond to different equations of state.

In order to characterize all possible burnt states that can be connected to a given unburnt state (0) by a combustion wave, we define in the  $(\tau, p)$ -plane the *Crussard curve* (or Hugoniot curve)  $\mathcal{C}$  with *center*  $(\tau_0, p_0)$  by

$$\varepsilon(\tau, p) - \varepsilon_0 + \frac{1}{2}(p + p_0)(\tau - \tau_0) = 0, \quad (4.8)$$

where

$$\varepsilon(\tau, p) = \varepsilon_b(\tau, p)$$

denotes from now on the equation of state of the *burnt* gas. Due to the heat of reaction  $Q$ , the pole  $A_0 = (\tau_0, p_0)$  does not belong to  $\mathcal{C}$ .

*Example 4.2.* Assuming as in Example 4.1 that both gases are ideal polytropic with the same  $\gamma$ -law, we find that the Crussard curve  $\mathcal{C}$  is the rectangular hyperbola

$$\frac{1}{2\mu^2}((\tau - \mu^2\tau_0)p - (\tau_0 - \mu^2\tau)p_0) - Q = 0,$$

where  $\mu^2 = \frac{(\gamma-1)}{(\gamma+1)}$  (see (2.19), Example 2.1). □

More generally, we can state the following result.

*Lemma 4.1*

Assume that the function  $(\tau, s) \mapsto p(\tau, s)$  satisfies the conditions (2.21),

$$\frac{\partial p}{\partial \tau} < 0, \quad \frac{\partial p}{\partial s}(\tau, s) > 0,$$

and that the reaction is exothermic,

$$\varepsilon(\tau_0, p_0) < \varepsilon_0.$$

Then, the Crussard curve with center  $A_0 = (\tau_0, p_0)$  lies above  $A_0$ .

*Proof.* Let  $B_0 = (\tau_0, p_1)$  be the point of  $\mathcal{C}$  with abscissa  $\tau_0$  (see Fig. 4.1). From (4.8) and (4.9), we deduce

$$\varepsilon(\tau_0, p_1) = \varepsilon_0 > \varepsilon(\tau_0, p_0).$$

Now, according to (2.21), the function  $p \mapsto \varepsilon(\tau_0, p)$  is strictly increasing since (see (1.1) and (2.25))

$$\frac{\partial \varepsilon}{\partial p}(\tau, p)|_\tau = \frac{\partial \varepsilon}{\partial s}(\tau, s) \frac{\partial s}{\partial p}(\tau, p) = T \left( \frac{\partial p}{\partial s}(\tau, p) \right)^{-1} > 0.$$

Thus

$$\varepsilon(\tau_0, p_1) = \varepsilon_0 > \varepsilon(\tau_0, p_0) \iff p_1 > p_0,$$

and  $B_0$  lies above  $A_0$ .

Notice that when (4.2) holds, (4.9) is equivalent to

$$\varepsilon(\tau_0, p_0) < \varepsilon_0 \iff Q > 0.$$

In the same way, let  $C_0 = (\tau_1, p_0)$  be the point on  $\mathcal{C}$  with  $p = p_0$ . Arguing now on the specific enthalpy defined by

$$h = \varepsilon + p\tau,$$

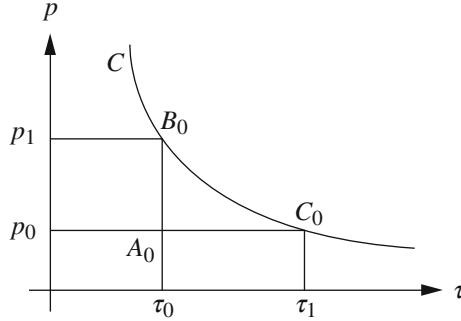
we have

$$\begin{aligned} h(\tau_0, p_0) &= \varepsilon(\tau_0, p_0) + p_0\tau_0, \\ h(\tau_1, p_0) &= \varepsilon(\tau_1, p_0) + p_0\tau_1 = \varepsilon_0 - p_0(\tau_1 - \tau_0) + p_0\tau_1 = \varepsilon_0 + p_0\tau_0, \end{aligned}$$

which implies by (4.9)

$$h(\tau_1, p_0) > h(\tau_0, p_0).$$

Now (see (1.1) and (2.25))



**Fig. 4.1** Crussard curve  $\mathcal{C}$  with center  $A_0$

$$\begin{aligned}\frac{\partial h}{\partial \tau}(\tau, p)|_{\tau} &= \frac{\partial \varepsilon}{\partial \tau}(\tau, p) + p \\ &= \frac{\partial \varepsilon}{\partial \tau}(\tau, s) + \frac{\partial \varepsilon}{\partial s}(\tau, s) \frac{\partial s}{\partial \tau}(\tau, p) + p = T \frac{\partial s}{\partial \tau}(\tau, p) > 0\end{aligned}$$

so that the function  $\tau \mapsto h(\tau, p_0)$  is strictly increasing and

$$h(\tau_1, p_0) > h(\tau_0, p_0) \iff \tau_1 > \tau_0,$$

which ends the proof.  $\square$

From now on, we shall assume that (2.21) holds. In view of (4.6), the part of the Crussard curve between  $B_0$  and  $C_0$  is not admissible since  $(\tau, p)$  must satisfy

$$\frac{p - p_0}{\tau - \tau_0} < 0.$$

The remaining part of the curve consists of two distinct branches: the upper branch with  $p > p_0$ ,  $\tau < \tau_0$ , called the *detonation branch*, and the lower one  $p < p_0$ ,  $\tau > \tau_0$ , called the *deflagration branch*.

For a given final state  $B = (\tau, p)$  of  $\mathcal{C}$ , the negative slope of the *Rayleigh line* joining  $A_0$  and  $B$

$$\frac{p - p_0}{\tau - \tau_0} = -M^2$$

enables us to determine the wave velocity  $\sigma = u_i - M\tau_i$  through (4.4). Conversely, for a given combustion velocity  $\sigma$ , we see that the corresponding Rayleigh line may or may not intersect the Crussard curve. Let us assume moreover that (2.29) holds, which implies (see Remark 2.6) the property that in the  $(\tau, p)$ -plane, the Rayleigh line drawn from a point  $\bar{B} = (\bar{\tau}, \bar{p})$  with slope

$$-M^2 = \frac{p - \bar{p}}{\tau - \bar{\tau}}$$

intersects the Hugoniot curve  $\mathcal{H}$  with pole  $\bar{B}$  at one and only one other point. We can then prove the following lemma.

*Lemma 4.2*

Assume that the conditions (2.21) and (2.29) hold. For a given  $\sigma$ , there may exist 0 or 2 (possibly coalescing) burnt states that can be connected to the (unburnt) state (0) by a combustion wave with speed  $\sigma$ . These two states belong to the same Hugoniot shock curve for the burnt gas.

*Proof.* Consider the Rayleigh line  $\Delta$  passing through  $A_0$  with slope  $-M^2$ , and suppose that  $\Delta$  intersects  $\mathcal{C}$  at some point  $B' = (\tau', p')$  (see Fig. 4.2). Then, let us see that  $\Delta$  intersects  $\mathcal{C}$  at one and only one other point  $B''$  (possibly coalescing with  $B'$ ). Since  $B'$  belongs to  $\mathcal{C}$ , we have

$$\varepsilon(\tau', p') - \varepsilon_0 + \frac{1}{2}(p' + p_0)(\tau' - \tau_0) = 0. \quad (4.9)$$

Now, let us introduce the shock Hugoniot curve  $\mathcal{H}'$  with center  $B'$  relative to the burnt gas,

$$\varepsilon(\tau, p) - \varepsilon(\tau', p') + \frac{1}{2}(p' + p)(\tau - \tau') = 0,$$

which we recall is the set of all (burnt) states that can be connected to the (burnt) state  $B'$  by a shock wave. If (2.29) holds and if the pressure may increase indefinitely along  $\mathcal{H}'$ , the Rayleigh line intersects  $\mathcal{H}'$  at exactly one other point  $B'' = (\tau'', p'')$  (possibly coalescing with  $B'$  if  $\Delta$  is tangent to  $\mathcal{H}'$ ), which thus satisfies

$$\varepsilon(\tau'', p'') - \varepsilon(\tau', p') + \frac{1}{2}(p' + p'')(\tau'' - \tau') = 0. \quad (4.10)$$

Let us check that  $B''$  also belongs to the Crussard curve  $\mathcal{C}$  with center  $A_0$ . Since  $A_0$ ,  $B'$ , and  $B''$  lie on the same Rayleigh line, we can write

$$-M^2 = \frac{p' - p_0}{\tau' - \tau_0} = \frac{p'' - p_0}{\tau'' - \tau_0} = \frac{p'' - p'}{\tau'' - \tau'}. \quad (4.11)$$

Setting

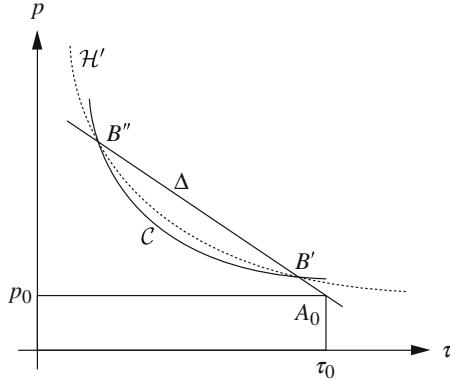
$$\varepsilon' = \varepsilon(\tau', p'), \quad \varepsilon'' = \varepsilon(\tau'', p''),$$

it follows from (4.9) and (4.11) that

$$\varepsilon' - \varepsilon_0 + \frac{(p'^2 - p_0^2)(\tau'' - \tau_0)}{2(p'' - p_0)} = 0, \quad (4.12)$$

and from (4.10) and (4.11) that

$$\varepsilon'' - \varepsilon' + \frac{(p''^2 - p'^2)(\tau'' - \tau_0)}{2(p'' - p_0)} = 0. \quad (4.13)$$



**Fig. 4.2** Rayleigh line  $\Delta$  through  $A_0$

Adding (4.12) and (4.13), we get

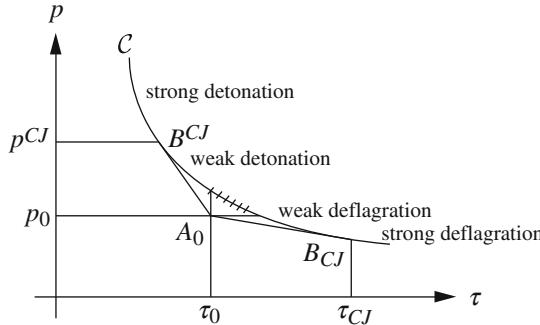
$$\varepsilon' - \varepsilon_0 + \frac{1}{2}(p'' + p_0)(\tau'' - \tau_0) = 0,$$

which means exactly that  $B''$  lies on the Crussard curve.  $\square$

According to Lemma 2, for  $M^2$  large enough, the Rayleigh line with slope  $-M^2$  intersects the Crussard curve at two points on the detonation branch. When  $M^2$  decreases, it reaches a minimum value for which the two points coalesce when the corresponding Rayleigh line is tangent to  $C$ . This point  $B^{CJ}$  represents the so-called Chapman-Jouguet detonation and separates the detonation branch of the Crussard curve into two parts: a detonation represented by a point on the upper part is called a *strong detonation*, while one represented by a point on the lower part is called a *weak detonation*.

For smaller values of  $M^2$ , the corresponding Rayleigh line  $\Delta$  does not intersect  $C$  until  $M^2$  reaches the value for which  $\Delta$  is tangent to the deflagration branch. This point of tangency  $B_{CJ}$  corresponds to the *Chapman-Jouguet deflagration*. It separates the deflagration branch into two parts: the upper part represents *weak deflagrations* and the lower one *strong deflagrations* (see Fig. 4.3).

*Remark 4.1.* Let us show that a detonation wave may be viewed as a precompression shock propagating into the unburnt gas followed by a deflagration wave, both processes having the same velocity. Indeed, let a shock connecting the state  $(0)$  to a state  $(\rho^*, u^*, p^*)$  be followed by a deflagration connecting  $(\rho^*, u^*, p^*)$  to  $(\rho_1, u_1, p_1)$  in such a way that the fluxes through both discontinuities are the same,



**Fig. 4.3** Detonation and deflagration branches on  $\mathcal{C}$

$$M = \rho_0 v_0 = \rho^* v^* = \rho_1 v_1.$$

Then, writing the remaining conditions (2.11) and (2.12) for the shock, and (4.5) and (4.6) for the deflagration, we obtain the following system, which is equivalent to the Rankine-Hugoniot relations:

$$\begin{aligned} M^2 &= -\frac{p^* - p_0}{\tau^* - \tau_0} = -\frac{p_1 - p}{\tau_1 - \tau_0}, \\ \varepsilon^* - \varepsilon_0 + \frac{(p^* + p_0)(\tau^* - \tau_0)}{2} &= \varepsilon_1 - \varepsilon^* + \frac{(p_1 + p^*)(\tau_1 - \tau^*)}{2} = 0. \end{aligned}$$

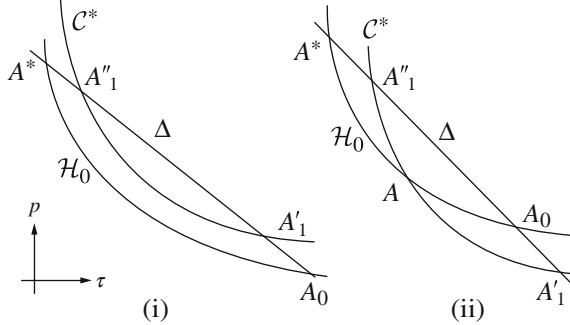
Using the same argument as in the proof of Lemma 2, we get, since the three states (0), (\*), (1) lie on the same Rayleigh line,

$$\begin{aligned} M^2 &= -\frac{p_1 - p_0}{\tau_1 - \tau_0}, \\ \varepsilon_1 - \varepsilon_0 + \frac{1}{2}(p_1 + p_0)(\tau_1 - \tau_0) &= 0. \end{aligned}$$

Thus, the process connecting the state (0) to the state (1) is indeed equivalent to a single process that is a detonation (see Fig. 4.4(i)).

Note that the Crussard curve  $\mathcal{C}^*$  with center  $A^* = (\tau^*, p^*)$  lies above the shock Hugoniot curve  $\mathcal{H}_0$  with center  $A_0 = (\tau_0, p_0)$ . Otherwise, the two curves would intersect at a point  $A = (\tau, p)$  (see Fig. 4.4 (ii)). By the above computations,  $A$  would belong to the same Rayleigh line as  $A_0$  and  $A^*$ , which is obviously impossible.

Moreover, we see in Fig. 4.4 (i) that two cases are possible that correspond to the points  $A'_1$  and  $A''_1$ . The transition from  $A_0$  to  $A''_1$  (resp. from  $A_0$  to  $A'_1$ ) is a strong (resp. weak) detonation representing a shock followed by a weak (resp. strong) deflagration. We shall come back to this interpretation



**Fig. 4.4** Case (ii) is not possible

in the next section when we introduce the Z.N.D. model (for Zeldovich-von Neumann-Döring).  $\square$

We shall now study in some detail the properties of the Chapman-Jouguet detonation or deflagration waves.

*Lemma 4.3*

*Along the Crussard curve, the entropy (of the burnt gas) is stationary at the Chapman-Jouguet (C.J.) points and only at these points. For a C.J. detonation, the entropy  $s$  is a relative minimum; for the C.J. deflagration, the entropy  $s$  is a relative maximum.*

*Proof.* We follow the argument of the proof of Lemma 2.4, since the expressions defining the Crussard and the Hugoniot curves are identical. From (2.27), we deduce that along  $\mathcal{C}$  (parametrized by  $\tau$ )

$$T\dot{s} = \frac{(\tau - \tau_0)}{2} \left\{ \frac{p - p_0}{\tau - \tau_0} - \dot{p} \right\},$$

where the dot  $\cdot$  denotes differentiation along  $\mathcal{C}$ . Hence  $\dot{s} = 0$  only at the points where the Rayleigh line is tangent to  $\mathcal{C}$ .

Now, if a Rayleigh line crosses the *detonation* branch at two points  $B'$  and  $B''$  as in Fig. 4.5, comparing the slopes gives

$$\dot{p}(\tau') > \frac{p' - p_0}{\tau' - \tau_0}, \quad \dot{p}(\tau'') < \frac{p'' - p_0}{\tau'' - \tau_0},$$

and we deduce that

$$\dot{s}(\tau') > 0 = \dot{s}(\tau^{CJ}) > \dot{s}(\tau'').$$

Hence, as  $\tau$  decreases from  $\tau_0$ , the entropy decreases from  $B_0$  to  $B^{CJ}$  and then increases after  $B^{CJ}$ . Moreover, since  $B'$  and  $B''$  are also on the same shock Hugoniot curve, we know from Corollary 2.1 that

$$s(\tau'') > s(\tau') > s(\tau^{CJ}). \quad (4.14a)$$

We check similarly the properties of the *deflagration* branch (see Fig. 4.5) and get

$$s(\tau_{II}) < s(\tau_I) < s(\tau_{CJ}). \quad (4.14b)$$

Note that the isentrope  $s = s^{CJ}$  is tangent to both the Crussard curve and the Rayleigh line at the point  $B^{CJ}$ . Indeed, the slope of the tangent to the isentrope  $s(\tau, p) = s^{CJ}$  is given by

$$-\frac{\partial s}{\partial \tau}(\tau, p)/\frac{\partial s}{\partial p}(\tau, p) = \frac{\partial p}{\partial \tau}(\tau, s),$$

where we have used (2.25). Since  $\dot{s}(\tau^{CJ}) = 0$ , we get

$$\frac{\partial p}{\partial \tau}(\tau, s^{CJ}) = \dot{p}(\tau^{CJ}) = \frac{p^{CJ} - p_0}{\tau^{CJ} - \tau_0}. \quad (4.15)$$

We shall use this property below (Example 4.3; also in (4.22)).  $\square$

#### *Lemma 4.4*

*Along the Crussard curve, the detonation (resp. the deflagration) speed  $|v_0|$  is a local minimum (resp. maximum) for a C.J. detonation (resp. deflagration). Moreover, for a C.J. process, the speed  $|v_1|$  of the burnt gas relative to the front is equal to the local sound speed, i.e.,*

$$|v_1| = c_1 \text{ at C.J. points .} \quad (4.16)$$

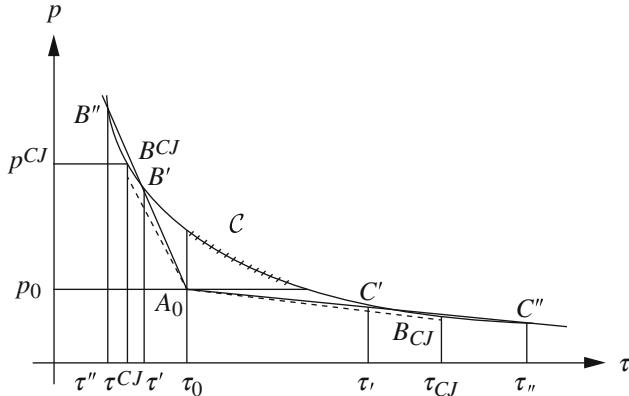
*Proof.* We have characterized the Chapman-Jouguet detonation as the detonation point on the Crussard curve with center  $A_0 = (\tau_0, p_0)$  for which the Rayleigh line

$$M^2 = -\frac{p - p_0}{\tau - \tau_0}$$

has minimal slope  $M^2$  (in absolute value). Since

$$M^2 = \frac{v_0^2}{\tau_0^2},$$

this proves that  $v_0^2$  is also minimum. In particular, if  $u_0 = 0$ , the detonation speed  $|\sigma|$  is minimum at  $B^{CJ}$ . The argument for a deflagration is similar.



**Fig. 4.5** Chapman-Jouguet states on the Crussard curve with center  $A_0$

Now, the local sound speed of the burnt gas satisfies (see (1.4))

$$-\rho^2 c^2 = \frac{\partial p}{\partial \tau}(\tau, s).$$

By (4.15), we have at the point  $\tau^{CJ} = \tau$

$$-\rho^2 c^2 = \dot{p}(\tau) = \frac{p - p_0}{\tau - \tau_0} = -M^2 = -\rho^2 v^2,$$

which implies

$$v^2 = c^2 \text{ at a C.J. point}$$

and gives the result.  $\square$

### Theorem 4.1

The gas flow relative to the reaction front satisfies the following properties:

- (i) For a strong detonation  $|v_0| > c_0, |v_1| < c_1$ ,
- (ii) For a weak detonation  $|v_0| > c_0, |v_1| > c_1$ ,
- (iii) For a weak deflagration  $|v_0| < c_0, |v_1| < c_1$ ,
- (iv) For a strong deflagration  $|v_0| < c_0, |v_1| > c_1$ .

*Proof.* We shall follow the same argument as in the proof of Theorem 2.3. We first consider the gas behind the reaction front. In the case of a detonation, let  $B'$  and  $B''$  be the two points on the intersection of the Rayleigh line with  $C$ . It follows from the proof of Lemma 2 that  $B''$  lies on the Hugoniot curve with center  $B'$ . Thus, applying the argument used in the proof of Theorem 2.3 (replacing  $A_0$  by  $B'$  and  $A_1$  by  $B''$ ), we have along  $\Delta$  parametrized by  $\alpha \geq 0$

$$\frac{ds}{d\alpha} > 0 \text{ at } B', \quad \frac{ds}{d\alpha} < 0 \text{ at } B''$$

and

$$c^2 < v^2 \text{ at } B', \quad c^2 > v^2 \text{ at } B''.$$

The case of a deflagration follows similarly.

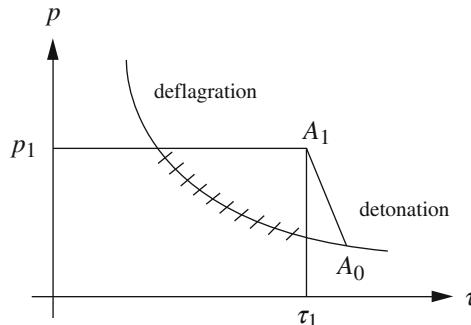
In order to study the property ahead of a reaction front, let us consider the backward Crussard curve of a given burnt state  $A_1 = (\tau_1, p_1)$ , i.e., the locus of all possible unburnt states  $A_0 = (\tau, p)$  that can be connected to  $A_1$  by a combustion wave. These states satisfy

$$\varepsilon_0(\tau, p) - \varepsilon_1 + \frac{1}{2}(p_1 + p)(\tau - \tau_1) = 0,$$

where  $\varepsilon_0(\tau, p)$  is the equation of state of the unburnt gas. Assuming as for Lemma 4.1 that the reaction is exothermic,

$$\varepsilon_0(\tau_1, p_1) > \varepsilon_1,$$

we obtain that the corresponding curve lies below the center  $A_1$  (see Fig. 4.6).

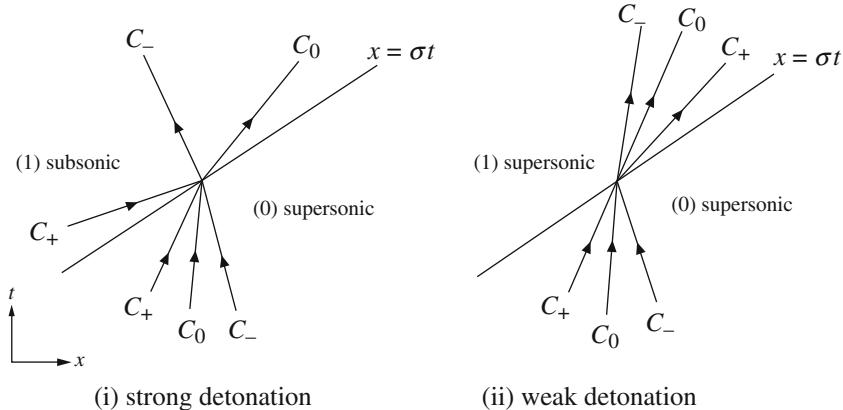


**Fig. 4.6** Backward Crussard curve of the state  $A_1$

We must again exclude the part of the curve corresponding to  $\frac{(p-p_1)}{(\tau-\tau_1)} \geq 0$ , and there remain two branches, the upper (resp. the lower) part representing the states that can be connected to  $A_1$  by a deflagration (resp. detonation) since  $p \geq p_1$  (resp.  $p \leq p_1$ ). Now let  $A_0$  be a state on the “detonation part” and  $\mathcal{H}_0$  be the Hugoniot function relative to the unburnt gas with center  $A_0$  (see (2.18))

$$\mathcal{H}_0(\tau, p) = \varepsilon_0(\tau, p) - \varepsilon_0 + \frac{1}{2}(p_0 + p)(\tau - \tau_0).$$

It follows from (2.31) that along the line  $A_0A_1$  the entropy of the unburnt gas has exactly one critical point, which is a maximum. Thus, if  $\Delta = A_0A_1$



**Fig. 4.7.1** Characteristics in the case of detonations

is parametrized by  $\alpha \geq 0$ ,  $\frac{ds}{d\alpha} > 0$  at  $A_0$ , and, concluding as in the proof of Theorem 2.3 or 4.1, we deduce, since  $\tau_1 < \tau_0$ , that

$$|v_0| > c_0.$$

If  $A_0$  lies on the “deflagration part,”  $\tau_1 < \tau_0$  and we get instead

$$|v_0| < c_0,$$

which ends the proof.

□

*Remark 4.2.* The four possibilities enumerated in Theorem 4.1—weak or strong deflagration or detonation—are mathematically admissible, i.e., are compatible with the conservation laws. But some of them are ruled out by physical considerations: in particular, weak detonations and strong deflagrations are not admissible, at least for the simple model that we have assumed [480, 1189]. The exclusion of weak detonations and strong deflagrations does not mean that this kind of wave does not exist at all. But these require special considerations apart from the simple model presented here and are believed to represent some particular phenomena (see again Williams [1189, Section 6.1.3, 6.2.2]).

Let us see how studying the characteristics can enhance these considerations. For instance, assume that we have a steady ( $\sigma$  constant) reaction front, moving to the right, the unburnt gas being on the right and the burnt gas on the left. Then, due to Theorem 4.1, the  $C_{\pm}$  characteristics with slope  $u \pm c$  (see Chap. II, Example 5.3) are as depicted in Figs. 4.7.1 and 4.7.2.

The considerations of characteristics and boundary conditions preceding Definition 5.1 of the Lax entropy conditions show that in case (i), which is the analog of an admissible 3-shock, the gas flow is completely determined by

the Rankine-Hugoniot conditions (the inequalities (5.12) and (5.13) in Chap. II hold with  $j = k - 1 = 2$ ). In cases (ii) and (iii), there remains one degree of indeterminacy, for instance, the velocity of the shock front can be chosen arbitrarily (in case (ii),  $j = k = 3$ , while in case (iii)  $j = k = 2$ ). In case (iv), there are two degrees of freedom ( $j = 3, k = 2$ ). Note in particular that the  $C_+$  characteristics diverge from the discontinuity.  $\square$

As in Sect. 3, it is also convenient to study the projection on the  $(u, p)$  plane of the Rankine-Hugoniot conditions. For specificity, we shall study the case where the combustion front propagates to the right: we suppose  $M < 0$ , and accordingly  $v_0$  and  $v_1$  are negative. For instance, if  $u_0 = 0$  the unburnt gas is at rest and  $\sigma$  is  $> 0$  (Fig. 4.8).

Assume that the two parts of the Crussard curve may be parametrized by  $p$ ; more precisely, assume that  $\tau = \tau(p)$  with  $p \geq p^0 - p_{B_0}$  for a detonation and  $p \leq p_0$  for a deflagration (see Figs. 4.1 and 4.3). Then, from equations (4.4) and (4.5), we get

$$u - u_0 = -\frac{(p - p_0)}{M(p)} = (p - p_0)\sqrt{(\tau - \tau_0)/(p - p_0)}.$$

In the case of a detonation,  $p \geq p^0 = p_{B_0} > p_0$  and  $\tau \leq \tau_0$ , we have

$$u - u_0 = \sqrt{(\tau_0 - \tau(p))(p - p_0)}, \quad (4.17)$$

and  $u$  increases from  $u_0$  to  $+\infty$  with, moreover,

$$\lim_{p \rightarrow p_+^0} \frac{du}{dp} = +\infty.$$

For a deflagration,  $p \leq p_0$  and  $\tau > \tau_0$ , we get

$$u - u_0 = \sqrt{(\tau_0 - \tau(p))(p - p_0)} \quad (4.18)$$

with

$$\lim_{p \rightarrow p_-^0} \frac{du}{dp} = +\infty.$$

We obtain the curve depicted in Fig. 4.9.

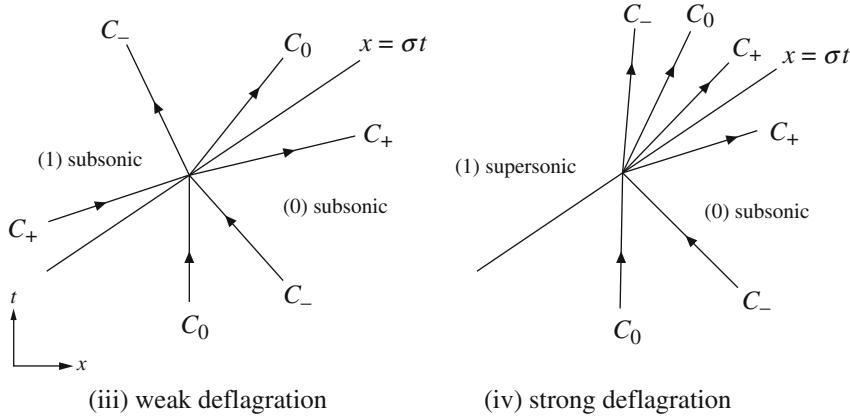
It is easily seen that there is a unique straight line passing through the point  $(u_0, p_0)$  that is tangent to the curve at both C.J. points. Indeed, differentiating

$$(u - u_0)^2 = -(p - p_0)(\tau - \tau_0), \quad (4.19)$$

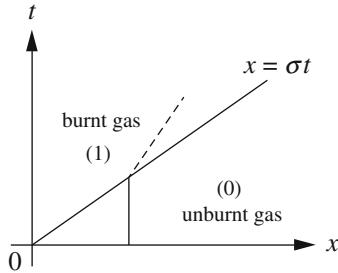
we get

$$2(u - u_0) \frac{du}{dp} = -(p - p_0) \frac{d\tau}{dp} - (\tau - \tau_0).$$

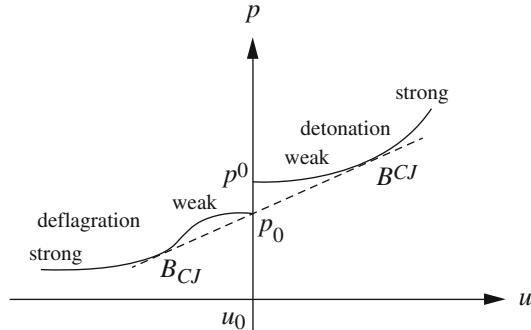
Thus, a line issuing from  $(u_0, p_0)$  is tangent at the point  $(u, p)$  if



**Fig. 4.7.2** Characteristics in the case of deflagrations



**Fig. 4.8** Combustion front



**Fig. 4.9** States that can be connected to the unburnt state (0) by a combustion wave

$$\frac{-(p - p_0)d\tau/dp - (\tau - \tau_0)}{2(u - u_0)} = \frac{u - u_0}{p - p_0}.$$

Together with (4.19), this yields

$$\frac{d\tau}{dp} = \frac{\tau - \tau_0}{p - p_0},$$

which characterizes the C.J. points. In particular, the value  $u^{CJ}$  at the C.J. detonation point is given by

$$u^{CJ} = u_0 + \sqrt{(\tau^{CJ} - \tau_0)(p_0 - p^{CJ})},$$

where  $p^{CJ}$  and  $\tau^{CJ}$  are the solutions with  $p > p_0, \tau < \tau_0$ , of

$$\begin{cases} \varepsilon(\tau, p) - \varepsilon_0 + \frac{1}{2}(p_0 + p)(\tau - \tau_0) = 0, \\ \frac{p - p_0}{\tau - \tau_0} = -\frac{c^2}{\tau^2}, \end{cases} \quad (4.20)$$

and the C.J. detonation speed is

$$\sigma^{CJ} = c^{CJ} + u^{CJ}. \quad (4.21)$$

For the C.J. deflagration, we consider the solutions of (4.20) with  $p < p_0, \tau \geq \tau_0$ , with (4.18) and (4.21).

*Example 4.3.* Let  $\gamma$  denote the adiabatic exponent, defined as the negative logarithmic slope of the isentrope (for a polytropic ideal gas, this definition coincides with the definitions of Sect. 1.2)

$$\gamma = -\left(\frac{\tau}{p}\right) \frac{\partial p(\tau, s)}{\partial \tau} \Bigg|_s = -\left(\frac{\partial \text{Log } p}{\partial \text{Log } \tau}\right) \Bigg|_s.$$

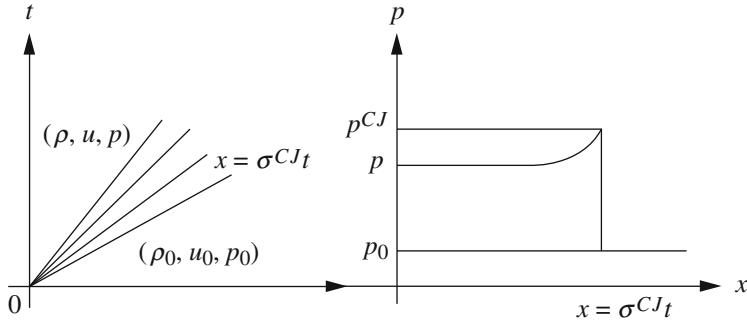
Since the Crussard curve and the isentrope are both tangent to the Rayleigh line at point C.J., equating the slopes in the  $(\text{Log } p, \text{Log } \tau)$ -plane gives

$$\gamma^{CJ} = -\left(\frac{\partial \text{Log } p}{\partial \text{Log } \tau}\right) \Bigg|_{s=s^{CJ}} = \frac{1 - (p_0/p^{CJ})}{(\tau_0/\tau^{CJ} - 1)}.$$

If  $p_0$  can be neglected (a case encountered for a liquid or a solid) and if the unburnt state is at rest ( $u_0 = 0$ ), we obtain by (4.4) and (4.5)

$$\begin{aligned} \sigma^{CJ} &= -v_0, \\ M^{CJ} &= \frac{v^{CJ}}{\tau^{CJ}} = -\frac{p^{CJ}}{u^{CJ}} = -\rho_0 \sigma^{CJ}, \\ (M^{CJ})^2 &= \frac{p^{CJ}}{\tau_0 - \tau^{CJ}}. \end{aligned}$$

We get the simple relations



**Fig. 4.10** Composite wave: C.J. detonation and 3-rarefaction wave

$$\begin{aligned}\tau^{CJ} &= \frac{\tau_0 \gamma^{CJ}}{\gamma^{CJ} + 1}, \quad p^{CJ} = \frac{\rho_0 (\sigma^{CJ})^2}{\gamma^{CJ} + 1}, \\ \frac{v^{CJ}}{v_0} &= \frac{\gamma^{CJ}}{\gamma^{CJ} + 1}, \quad u^{CJ} = \frac{\sigma^{CJ}}{\gamma^{CJ} + 1}, \\ c^{CJ} &= \sigma^{CJ} - u^{CJ} = \gamma^{CJ} u^{CJ}.\end{aligned}$$

For a constant  $\gamma$ -law  $\varepsilon(\tau, p) = \frac{p\tau}{\gamma-1} - q$ , a further computation gives

$$(\sigma^{CJ})^2 = 2q(\gamma^2 - 1)$$

for the C.J. detonation state.  $\square$

Let us study more precisely the flows involving a *detonation* process. On the basis of physical considerations (a closer study of the wave structure, to which we shall return in the next section), we shall assume that only C.J. detonations and strong detonations are admissible. Instead of weak detonations, we consider composite waves involving a Chapman-Jouguet process and a non-reacting rarefaction wave. For instance, in the  $(u, p)$ -plane, the state  $(u_0, p_0)$  can be connected to a state  $(u, p)$  with  $u < u^{CJ}$  by a C.J. detonation followed by a 3-rarefaction wave (see Fig. 4.10).

This is indeed possible since the fan of the rarefaction wave is bordered on the right by the line with slope  $\lambda = u + c$  and we know by (4.21) that, at the C.J. detonation point,

$$\lambda = \sigma^{CJ} = c^{CJ} + u^{CJ}.$$

Thus, in Fig. 4.11, the detonation part of the curve in Fig. 4.9 is replaced for  $u < u^{CJ}$  by

$$u = u^{CJ} + \Psi^{CJ}(p), \quad p < p^{CJ},$$

which is the equation of the 3-rarefaction curve passing through the C.J. detonation (see (3.10)). Note that the rarefaction curve through  $B^{CJ}$  is an

isentrope (see (3.8)) that is tangent to the detonation curve, as we have already noticed in the proof of Lemma 4.3.

The locus of burnt states that can be connected to the state (0) by a strong detonation or a C.J. detonation followed by a 3-rarefaction is then

$$u = \begin{cases} u_0 + \sqrt{(\tau(p) - \tau_0)(p_0 - p)}, & p \geq p^{CJ}, \\ u_0 + \sqrt{(\tau^{CJ} - \tau_0)(p_0 - p^{CJ})} + \Psi^{CJ}(p), & p < p^{CJ}. \end{cases} \quad (4.22)$$

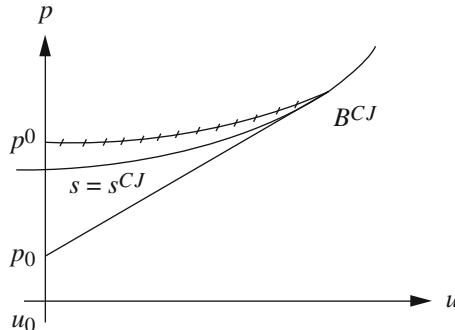
The function in (4.22) is easily seen to be increasing.

This enables us to solve the Riemann problem for a reacting gas in the particular case of a flow involving a detonation. We assume that the right state is the unburnt state (0) and the left state is the burnt state (1):

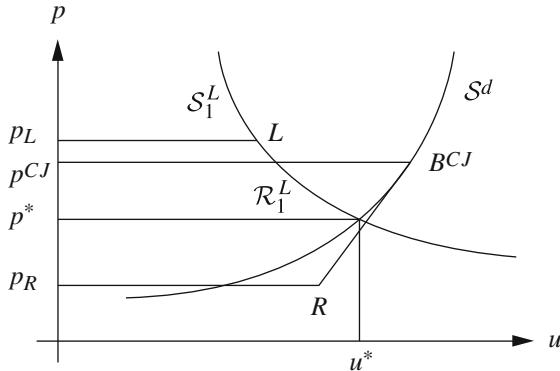
$$(\rho, u, p)(x, 0) = \begin{cases} (\rho_L, u_L, p_L) = (\rho_1, u_1, p_1), & x < 0, \\ (\rho_R, u_R, p_R) = (\rho_0, u_0, p_0), & x > 0. \end{cases}$$

By analogy with the general theory of Chap. II, we look for a solution where the left state is connected to the right state by a 1-wave (shock or rarefaction propagating in the burnt gas), a 2-contact discontinuity and a strong detonation, or a composite wave involving a C.J. detonation and a 3-rarefaction wave (see Fig. 4.12).

Let  $p^{CJ}$  be the pressure at the C.J. detonation point, which is entirely determined by the right state and the equation of state of the burnt gas. As in Sect. 3, we look at the intersection in the  $(u, p)$  plane of the curve (3.16) (the locus of states to which the left state can be connected by a 1-wave) with the curve  $\mathcal{S}^d$  (4.22), which we have described above. This intersection  $(u^*, p^*)$  gives the velocity and the pressure at the two intermediate constant states. It remains to determine  $\rho_I$  and  $\rho_{II}$ , which is done following the lines of Sect. 3.



**Fig. 4.11** Isentrope through the C.J. detonation state

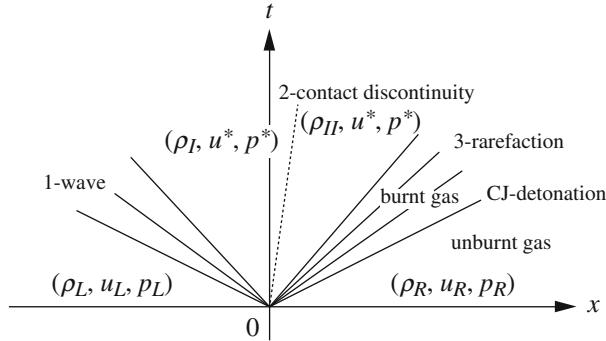


**Fig. 4.12** Solution of the Riemann problem in the  $(u, p)$ -plane

There are four cases according to the relative positions of  $p_L = p_1, p_R = p_0$ , and  $p^{CJ}$ . For instance, if we assume that  $p_L > p^{CJ} > p_R$ , a look at Fig. 4.12 shows that the corresponding solution of the Riemann problem is as depicted in Fig. 4.13.

*Remark 4.3.* We shall not investigate in the same way the case of a flow involving a deflagration. A first reason is, as noticed in Remark 4.2, that in this simple model deflagration processes have a higher degree of indeterminacy because the gas flow relative to the front is subsonic ahead of a deflagration front. A possible remedy would be to fix, for instance, the reaction rate and then search the states that can be connected to a right state by a weak deflagration preceded by a nonreacting 3-wave (shock or rarefaction). Strong deflagrations are excluded on the same basis as weak detonation waves (see Courant and Friedrichs [371, Section 93]; Williams [1189, Section 6.1.2.3]). The intersection with the curve (3.16) will then determine a possible wave pattern (see Teng et al. [1116]).  $\square$

*Remark 4.4.* In fact, weak deflagrations propagate with a definite wave speed that depends on the reaction rate, and they are nearly isobaric (see Williams [1189, Sections 6.1.2.3, 6.1.3]). Note also that the exclusion of weak detonations and strong deflagrations does not mean that this kind of wave does not exist at all. But they require special considerations apart from the simple model presented here and are believed to represent some particular phenomena (see Williams [1189, Sections 6.1.3, 6.2.2]).  $\square$



**Fig. 4.13** Solution of the Riemann problem in the  $(x, t)$ -plane

## 5 Reacting Flows: The Z.N.D. Model for Detonations

The preceding theory is now extended to include a finite reaction rate. Thus, we do not assume anymore that the reaction takes place instantaneously. Instead, we make the assumption that a detonation process can be modeled as a nonreacting shock wave propagating in the unburnt gas initiating a chemical reaction. Hence, we suppose that the reaction is irreversible and the reaction rate is zero ahead of the shock and finite behind (the reaction is complete at the end of the reaction zone). Again, the flow is supposed to be planar and one-dimensional, and transport effects (heat conduction, viscosity) are neglected, and we consider the simplest possible chemical process  $R \mapsto P$ . Recall that  $z$  is the mass fraction of the burnt gas. Then, the equations to be considered are the system (4.1) with

$$\varepsilon = \varepsilon(\rho, p, z), \quad r = r(\rho, p, z), \quad (5.1)$$

or equivalently

$$\varepsilon = \varepsilon(\tau, p, z), \quad r = r(\tau, p, z)$$

when it is more convenient. We rewrite the system in the form

$$\begin{cases} \frac{\partial \Phi}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\Phi, z) = 0, \\ \frac{\partial}{\partial t} (\rho z) + \frac{\partial}{\partial x} (\rho z u) = \rho r, \end{cases} \quad (5.2)$$

where

$$\Phi = \begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix}, \quad \mathbf{f}(\Phi, z) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ (\rho e + p)u \end{pmatrix}.$$

We will show that there exists a family of travelling wave solutions of (5.2), i.e., solutions of the form

$$(\Phi, z)(x, t) = (\hat{\Phi}, \hat{z})(\xi), \quad \xi = x - \sigma t,$$

depending on the parameter  $\sigma$ , which is the constant velocity of the travelling wave, connecting the unburnt state  $z = 0$  to the burnt state  $z = 1$ , and therefore satisfying

$$\begin{cases} (\hat{\Phi}, \hat{z})(\xi) = (\Phi_0, 0) & \text{for } \xi > 0, \\ (\hat{\Phi}, \hat{z})(0) = (\Phi_N, 0) & \text{at } \xi = 0, \\ \lim_{\xi \rightarrow -\infty} (\hat{\Phi}, \hat{z})(\xi) = (\Phi_1, 1) & \text{(total reaction),} \end{cases} \quad (5.3)$$

where  $\Phi_0$  is a given constant initial state. Thus, the problem is to determine the function  $(\hat{\Phi}, \hat{z})(\xi)$ ,  $\xi < 0$ , for a given parameter  $\sigma$ .

We begin by writing the usual Rankine-Hugoniot jump condition for the ordinary shock discontinuity in the unburnt gas (dropping the “^”)

$$\sigma[(\Phi, z)] = [f(\Phi, z)] \quad \text{at } \xi = 0.$$

The shock is supposed to be inert,  $z = 0$  across the shock, so that

$$\sigma(\Phi_N - \Phi_0) = \mathbf{f}(\Phi_N, 0) - \mathbf{f}(\Phi_0, 0). \quad (5.4)$$

Now, assuming that the solution is smooth for  $\xi < 0$ , we have

$$\begin{cases} -\sigma \frac{d\Phi}{d\xi} + \frac{d}{d\xi} \mathbf{f}(\Phi, z) = 0, \\ -\sigma \frac{d}{d\xi}(\rho z) + \frac{d}{d\xi}(\rho z u) = \rho r. \end{cases} \quad (5.5)$$

Integrating the first equation (5.5) gives

$$\sigma(\Phi(\xi) - \Phi_N) = \mathbf{f}(\Phi(\xi), z(\xi)) - \mathbf{f}(\Phi_N, 0),$$

which together with (5.4) yields

$$\sigma(\Phi(\xi) - \Phi_0) = \mathbf{f}(\Phi(\xi), z(\xi)) - \mathbf{f}(\Phi_0, 0),$$

or equivalently

$$\begin{cases} \sigma(\rho(\xi) - \rho_0) = (\rho u)(\xi) - (\rho u)_0, \\ \sigma((\rho u)(\xi) - (\rho u)_0) = (\rho u^2 + p)(\xi) - (\rho u^2 + p)_0, \\ \sigma((\rho e)(\xi) - (\rho e)_0) = ((\rho e + p)u(\xi) - ((\rho e + p)u)_0. \end{cases}$$

We find the analog of the Rankine-Hugoniot relations. Setting

$$M = \rho_0(u_0 - \sigma) = \rho(\xi)(u(\xi) - \sigma),$$

we get as in the nonreacting case (2.10)–(2.12) or as in the Chapman-Jouguet case (4.4)–(4.6)

$$\begin{aligned} M &= \frac{u(\xi) - u_0}{\tau(\xi) - \tau_0} = -\frac{p(\xi) - p_0}{u(\xi) - u_0}, \\ M^2 &= -\frac{p(\xi) - p_0}{\tau(\xi) - \tau_0}, \end{aligned} \quad (5.6)$$

$$\varepsilon(\tau(\xi), p(\xi), z(\xi)) - \varepsilon(\tau_0, p_0, 0) + \frac{1}{2}(p(\xi) + p_0)(\tau(\xi) - \tau_0) = 0. \quad (5.7)$$

Let us next consider the second equation (5.5), which can be written in the form

$$z \left\{ -\sigma \frac{d\rho}{d\xi} + \frac{d}{d\xi}(\rho u) \right\} + \frac{dz}{d\xi} \{\rho(-\sigma + u)\} = \rho r.$$

Since

$$\sigma \frac{d\rho}{d\xi} = \frac{d}{d\xi}(\rho u),$$

we obtain

$$\frac{dz}{d\xi} = \frac{1}{M} r(\tau(\xi), p(\xi), z). \quad (5.8)$$

Therefore, the problem amounts to finding the triple  $(\tau(\xi), p(\xi), z(\xi))$  for  $\xi < 0$  which is the solution of equations (5.6)–(5.8) with the initial condition  $z(0) = 0$ . We solve the equations (5.6)–(5.8) in two steps. First, considering  $z$  as a parameter, we solve (5.6) and (5.7), which gives  $\tau$  and  $p$  as functions of  $z$ . Then, replacing  $\tau$  and  $p$  in (5.8) by their values, we solve the ordinary differential problem

$$\begin{cases} \frac{dz(\xi)}{d\xi} = \frac{1}{M} r(\tau(z), p(z), z)(\xi), \\ z(0) = 0. \end{cases} \quad (5.9)$$

Indeed, we consider the equation of a Rayleigh line

$$M^2 = -\frac{p - p_0}{\tau - \tau_0}, \quad (5.10)$$

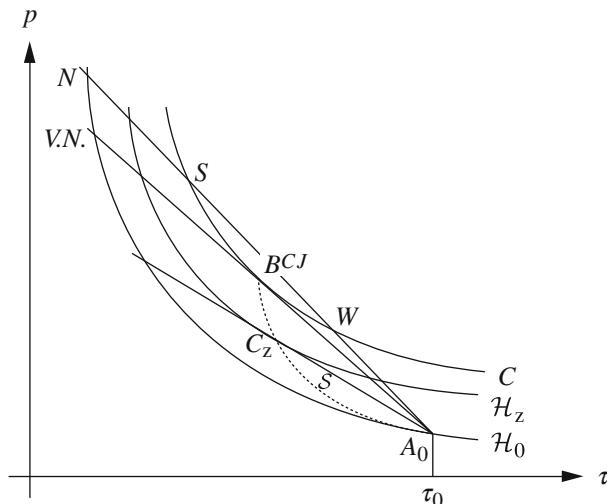
and we introduce for  $0 \leq z \leq 1$  the family of Hugoniot curves  $\mathcal{H}_z$  depending on  $z$ :

$$\varepsilon(\tau, p, z) - \varepsilon(\tau_0, p_0, 0) + \frac{1}{2}(p + p_0)(\tau - \tau_0) = 0.$$

For  $z = 0$ , we find the shock Hugoniot curve  $\mathcal{H}_0$  with center  $A_0 = (\tau_0, p_0)$ , whereas for  $z = 1$ ,  $\mathcal{H}_1$  corresponds to the Crussard curve  $\mathcal{C}$  with center  $A_0$  (Fig. 5.1).

Solving (5.6) and (5.7) amounts to finding the intersection of  $\mathcal{H}_z$  with the Rayleigh line  $\Delta$ . Due to (5.6), for a given value of  $\sigma$  and thus of  $M$ , all the partial reaction states lie on  $\Delta$ . By Lemma 4.2, the line intersects the Hugoniot curve  $\mathcal{H}_z$  at 0 or 2 (possibly coalescing) points. At the point  $C_z$  where  $\Delta$  is tangent to  $\mathcal{H}_z$ , the Rayleigh line is also tangent to an isentrope (see Lemma 4.3), and the flow is sonic (Lemma 4.4). The locus  $\mathcal{S}$  of all such points  $C_z$  is called the *sonic locus*. The curve starts from the center  $A_0$  on  $\mathcal{H}_0$  and ends on the Crussard curve at the C.J. detonation point. Moreover, the sonic locus  $\mathcal{S}$  separates the region lying between  $\mathcal{H}_0$  and  $\mathcal{C}$  in the  $(\tau, p)$ -plane into two zones: one (on the left of  $\mathcal{S}$ ) where the flow is subsonic by Theorem 4.1 and the other where it is supersonic (on the right) (Fig. 5.1).

Provided an explicit equation of state  $\varepsilon = \varepsilon(\tau, p, z)$  is known, we can obtain the function  $p = p_\sigma(z)$ , depending on  $\sigma$ , by eliminating  $\tau$  from the equations (5.9) and (5.10). Each function  $p_\sigma$  has two branches: one subsonic, which for  $z = 1$  corresponds to a strong detonation, and the other supersonic, corresponding to a weak detonation. We plot in Fig. 5.2 three typical curves depending on the position of  $\sigma$  with respect to  $\sigma^{CJ}(\sigma^{CJ}$  is indeed the value of  $\sigma$  for which  $\Delta$  is tangent to the Crussard curve).

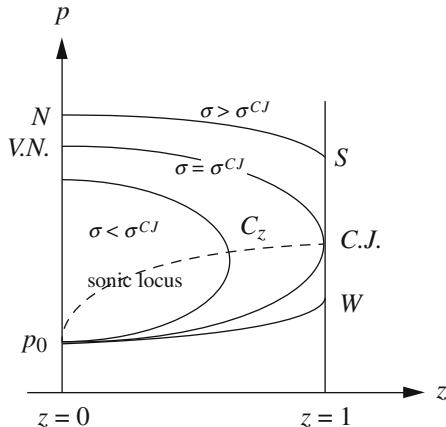


**Fig. 5.1** Hugoniot curves  $\mathcal{H}_z$

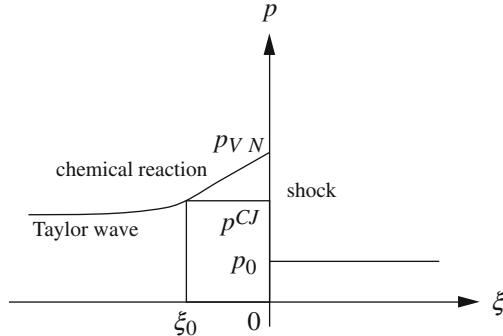
Now, there exist two possibilities:

- (i)  $\sigma \geq \sigma^{CJ}$ : The Rayleigh line  $\Delta$  intersects the Crussard curve  $\mathcal{C}$  at two points  $S$  and  $W$  (coalescing at  $B^{CJ}$  if  $\sigma = \sigma^{CJ}$ ). The only admissible final state is  $S$ , as follows from thermodynamical considerations (see Remark 5.1).
- (ii)  $\sigma < \sigma^{CJ}$ : The Rayleigh line  $\Delta$  does not intersect the Crussard curve  $\mathcal{C}$ . We see in Fig. 5.2 that there may be a solution ending on the sonic locus  $\mathcal{S}$  that corresponds to a maximum value of  $z$ . However, at such a point  $r > 0$  (since  $z < 1$ ) and the reaction is irreversible so that (5.8) cannot be satisfied.

For  $\sigma \geq \sigma^{CJ}$  we have obtained the existence of a combustion profile of the form (5.3), travelling with speed  $\sigma$ . It consists of a nonreacting shock wave (represented by  $A_0 N$  in the  $(\tau, p)$ -plane; see Fig. 5.1), followed by a chemical reaction (between  $N$  and  $S$ ). The reaction is achieved in finite time, and the length of the reaction zone is finite if there exists a finite  $\xi_0 > -\infty$  such that  $z(\xi_0) = 1$ . In this model, we have assumed that the shock initiates the chemical reaction. Thus, through the shock, pressure and density are raised instantaneously from  $A_0$  to a point  $N$  on the shock Hugoniot curve  $\mathcal{H}_0$ . Then, the reaction proceeds and the state is described in the  $(\tau, p)$ -plane by a point on the Rayleigh line which is the intersection with the curve  $\mathcal{H}_z$ . If it were  $W$  (see Fig. 5.1), the gas would have before reached the state  $S$  where it is in chemical ( $z = 1$ ) and thermodynamical equilibrium ( $S \in \mathcal{C}$ ). Since by (4.14) the entropy in  $W$  is less than in  $S$ , we cannot find an admissible transformation that leads from  $S$  to  $W$  on the Rayleigh line.



**Fig. 5.2** The curves  $p = p_\sigma(z)$



**Fig. 5.3** C.J. detonation and Taylor wave

*Remark 5.1.* If the length of the reaction zone is finite (i.e., if there exists a finite  $\xi_0 > -\infty$  such that  $z(\xi_0) = 1$ ), a rarefaction wave may follow the reaction zone to join a final state with given pressure less than  $p^{CJ}$  as is indeed the case in the experiments. However, this supposes that the speed  $\sigma$  is equal to  $\sigma^{CJ}$  (so that the flow is sonic), this because the speed of the head of the rarefaction wave is equal to the characteristic speed  $u + c$ . However at point  $S$ , the flow is subsonic and  $u + c > \sigma$ . In that case, the rarefaction and detonation waves will interact until the speed of the head of the rarefaction wave equals that of the detonation wave, which supposes

$$\sigma = \sigma^{CJ}.$$

In short, this justifies the famous *Chapman-Jouguet hypothesis* that only C.J. detonations are possible. The point *V.N.* on the Hugoniot curve from which the chemical reaction starts is called the *von Neumann* state. The final state is the C.J. point. The following rarefaction wave is called the *Taylor wave*. We have depicted in Fig. 5.3 the corresponding pressure curve. We note that the above model excludes the possibility of weak detonations.  $\square$

These last two sections have given but an introduction to the important subject of reacting flows. The following references will convince the reader that the subject is far too wide for us to pretend to have been exhaustive. We have tried to develop just the simplest approaches in the spirit of the preceding sections.

## Notes

For a better understanding of gas dynamics, we recommend the basic book of Courant and Friedrichs [371] which contains a section on reacting flows and those of Anderson [40], Whitham [1188], and Lighthill [800], also [651]. We mention again the references already quoted in the previous chapters

[278, 650, 1105], in particular Smoller [1066], Chapter 18, and the references therein, and Serre [1038], Chapter 4, also the important paper of Smith [1064], a very complete study of general equations of state in Menikoff and Plohr [862], and the paper of Wagner [1175], where the equivalence of the Euler and Lagrangian equations for weak solutions is proved (on this subject, see also Dafermos [383] and Wagner [1176] and, more recent, Peng [939]). A more precise study of the rarefaction shock curves (the nonadmissible part of the shock curves) can be found in Smoller et al. [1067].

Concerning the convexity of the entropy (Sect. 1), see the work of Dubois [432], and Croisille and Delorme [375], and Croisille and Villedieu [377] for a kinetic approach.

The solution of the Riemann problem in the case of materials with nonconvex equations of state can be found in Wendroff [1184] and Hattori [596] for a van der Waals fluid (see Kawashima and Matsumura [684] for the stability of shock profiles). For multicomponent gas dynamics, see, for instance, Abgrall [1] or Larrouy-Tourou [739]; the influence of the source term is shown in Fey et al. [475]; see also [832].

Concerning combustion, we refer to the book of Courant and Friedrichs [371] already mentioned, and those of Chorin and Marsden (3rd edition 1993) [305], Fickett and Davis [480], Williams [1189], Oran and Boris [916], Chéret [297], to the nice unpublished report of Thouvenin [1122], and to the papers of Chorin [304], Majda [839], Teng et al. [1116], Chen and Wagner [294], Embid et al. [459], Colella et al. [331], Bukiet [217], Ben Artzi [101], Tan and Zhang [1097], Gasser and Szmolyan [510], Sheng and Tan [1051], and Liu and Ying [827].

#### *Note Added in the Second Edition*

Since the first edition, several books have been published, among which the already cited books of D. Serre (now translated) [1039, 1041] and C. Dafermos [384] and the handbook of mathematical fluid dynamics (Ed. by S. Friedlander and D. Serre [493]) which offer a much deeper theoretical insight. We also mention some textbooks more oriented toward modeling (for instance, multicomponent flow [515]), or the construction of numerical schemes: R.J. LeVeque [777], also [774], E.F. Toro (3rd edition) [1127], and B. Després [416]. Now more theoretical results concerning uniqueness and asymptotic stability of Riemann solutions for the compressible Euler equations can be found in Chen and Frid [285]. Let us also mention a few references concerning nonconvex equation of state [385, 881], and other interesting topics in the research report of the DFG priority research program [1183], the reacting compressible Euler equations [295, 790], the reacting Riemann problem, [636, 749, 1219] (associated with evaporation front), [98, 728], and references therein, combustion [287, 636, 790] and also concerning Chapman-Jouguet detonation seen as *phase transition* [336].



# IV

## Finite Volume Schemes for One-Dimensional Systems

### 1 Generalities on Finite Volume Methods for Systems

Let us consider again the Cauchy problem for a general system of conservation laws

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = \mathbf{0}, & x \in \mathbb{R}, t > 0, \\ \mathbf{u}(x, 0) = \mathbf{u}_0(x), \end{cases} \quad (1.1)$$

where  $\mathbf{u} = (u_1, \dots, u_p)^T$  is a  $p$ -vector. As usual, we assume that this system is hyperbolic, i.e., the Jacobian matrix  $\mathbf{A}(\mathbf{u}) = \mathbf{f}'(\mathbf{u})$  of  $\mathbf{f}(\mathbf{u})$  has  $p$  real eigenvalues ranked in increasing order,

$$a_1(\mathbf{u}) \leq a_2(\mathbf{u}) \leq \dots \leq a_p(\mathbf{u}),$$

and a complete set of eigenvectors. Moreover, we shall assume that for all  $1 \leq k \leq p$ , the  $k$ th characteristic field is either genuinely nonlinear or linearly degenerate.

In a finite volume approach, we consider a partition of the domain  $\mathbb{R}$ , with cells  $C_j$  which are naturally intervals,  $C_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ ,  $j \in \mathbb{Z}$ , with center  $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$  and length  $|C_j|$ . The solution is approximated at a discrete time  $t_n$  by one unknown  $\mathbf{v}_j^n$  on each cell  $C_j$ , and we approximate the mean value of  $\mathbf{u}$  over the cell  $C_j$ . Thus, we look for  $\mathbf{v}_j^n \in \mathbb{R}^p$

$$\mathbf{v}_j^n \approx \frac{1}{|C_j|} \int_{C_j} \mathbf{u}(x, t_n) d\mathbf{x},$$

and the scheme satisfied by the  $(\mathbf{v}_j^n)$ s is obtained in a first step by integrating the conservation law (1.1) over  $C_j$ , which gives

$$\frac{d}{dt} \int_{C_j} \mathbf{u}(x, t) d\mathbf{x} + \mathbf{f}(\mathbf{u}(x_{j+\frac{1}{2}}-, t)) - \mathbf{f}(\mathbf{u}(x_{j-\frac{1}{2}}+, t)) = \mathbf{0},$$

and then by approximating the different terms. In particular, the interface fluxes  $\mathbf{f}(\mathbf{u}(x_{j\pm\frac{1}{2}}, t))$  are approximated by *numerical fluxes*  $\mathbf{g}_{j\pm\frac{1}{2}}(t)$ . This step results in the “method of lines”

$$\frac{d}{dt}\mathbf{v}_j(t) + \frac{1}{\Delta x}(\mathbf{g}_{j+\frac{1}{2}}(t) - \mathbf{g}_{j-\frac{1}{2}}(t)) = \mathbf{0},$$

where

$$\mathbf{v}_j(t) \approx \frac{1}{|C_j|} \int_{C_j} \mathbf{u}(x, t) dx.$$

The numerical fluxes should depend only on the unknowns  $\mathbf{v}_j$ 's; they are thus defined from *one* numerical flux function  $\mathbf{g}$  and

$$\mathbf{g}_{j+\frac{1}{2}}(t) = \mathbf{g}(\mathbf{v}_{j-k+1}(t), \dots, \mathbf{v}_{j+k}(t)),$$

if we choose that it depends on  $k$  values on each side of the interface.

In a last step, one uses a classical numerical method to approximate the resulting system of ODE, setting  $\mathbf{v}_j^n \approx \mathbf{v}_j(t_n)$ . In this last step, if we use the Euler explicit method, the resulting scheme writes

$$|C_j|\mathbf{v}_j^{n+1} = |C_j|\mathbf{v}_j^n - \Delta t_n (\mathbf{g}_{j+\frac{1}{2}}^n - \mathbf{g}_{j-\frac{1}{2}}^n)$$

with  $\Delta t_n = t_{n+1} - t_n$ ,  $\mathbf{g}_{j+\frac{1}{2}}^n = \mathbf{g}(\mathbf{v}_{j-k+1}^n, \dots, \mathbf{v}_{j+k}^n)$ . Starting from  $\mathbf{v}_j^0 = \frac{1}{|C_j|} \int_{C_j} \mathbf{u}_0(x) dx$ ,  $j \in \mathbb{Z}$ , the above formula enables us to compute explicitly all the values of  $\mathbf{v}_j^n$ ,  $j \in \mathbb{Z}, n > 0$ . Since we have chosen an explicit time scheme, the time step  $\Delta t_n$  will be limited by a stability condition that will be introduced later on.

*Remark 1.1.* In this last step, it is possible to use a higher-order explicit method, such as a Runge–Kutta scheme, for instance, the Heun scheme, or even an implicit method. We will not consider in detail this last possibility which for systems necessitates the inversion of a nonlinear system and in practice is used together with a linearization procedure.  $\square$

We shall mainly consider uniform grids in the present chapter concerning the one-dimensional case, setting  $\Delta x = |C_j|$ . Concerning the time step, since many results concern the updating of  $\mathbf{v}_j^n$  over one time step, it is not a real limitation to consider also a uniform  $\Delta t$ , though, in practice, the limiting stability condition should be written at each time step with  $\Delta t_n$ .

Then, the finite volume approach resembles the finite difference one, and we often use the term *finite difference scheme*, even if we only consider schemes in conservation form (1.2ab). For the notations concerning the difference schemes, we adopt the same as in Chapter 3 of G.R. [539].

Given a uniform grid with time step  $\Delta t$  and spatial mesh size  $\Delta x$ , we define an approximation  $\mathbf{v}_j^n \in \mathbb{R}^p$  of  $\mathbf{u}(x_j, t_n)$  at the point  $(x_j = j\Delta x, t_n = n\Delta t)$  by the formula

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda (\mathbf{g}_{j+\frac{1}{2}}^n - \mathbf{g}_{j-\frac{1}{2}}^n), \quad j \in \mathbb{Z}, n \geq 0 \quad (1.2a)$$

where  $(\mathbf{v}_j^0)_{j \in \mathbb{Z}}$  is given,

$$\lambda = \frac{\Delta t}{\Delta x},$$

and we have

$$\mathbf{g}_{j+\frac{1}{2}}^n = \mathbf{g}(\mathbf{v}_{j-k+1}^n, \dots, \mathbf{v}_{j+k}^n), \quad (1.2b)$$

where the function  $\mathbf{g}: \mathbb{R}^{p \times 2k} \rightarrow \mathbb{R}^p$  is continuous and is called the *numerical flux*. The scheme is said to be *consistent* with (1.1) if the numerical flux  $\mathbf{g}$  is consistent with  $\mathbf{f}$  in the sense that it satisfies

$$\mathbf{g}(\mathbf{u}, \dots, \mathbf{u}) = \mathbf{f}(\mathbf{u}), \quad \forall \mathbf{u} \in \mathbb{R}^p. \quad (1.2c)$$

The scheme is said to be *consistent with the entropy condition* associated with an entropy pair  $(U, F)$  for (1.1) if there exists a continuous function  $G: \mathbb{R}^{p \times 2k} \rightarrow \mathbb{R}$ , consistent with  $F$

$$G(\mathbf{u}, \dots, \mathbf{u}) = F(\mathbf{u}), \quad \forall \mathbf{u} \in \mathbb{R}^p,$$

which satisfies the cell entropy inequality

$$U(\mathbf{v}_j^{n+1}) \leq U(\mathbf{v}_j^n) - \lambda (G_{j+\frac{1}{2}}^n - G_{j-\frac{1}{2}}^n), \quad j \in \mathbb{Z}, n \geq 0,$$

where

$$G_{j+\frac{1}{2}}^n = G(\mathbf{v}_{j-k+1}^n, \dots, \mathbf{v}_{j+k}^n),$$

and  $G$  is called the numerical entropy flux.

The scheme (1.2a) and (1.2b) is a  $(2k+1)$ -point scheme, and when  $k=1$ , we have a 3-point scheme. The scheme is *essentially 3-point* if it satisfies the stronger consistency relation:

$$\mathbf{g}(\mathbf{v}_{-k+1}, \dots, \mathbf{v}_{-1}, \mathbf{u}, \mathbf{u}, \mathbf{v}_2, \dots, \mathbf{v}_k) = \mathbf{f}(\mathbf{u}), \quad \forall \mathbf{v}_{-k+1}, \dots, \mathbf{v}_k \in \mathbb{R}^p, \quad \forall \mathbf{u} \in \mathbb{R}^p.$$

Scheme (1.2) can also be written in the general form

$$\mathbf{v}_j^{n+1} = \mathbf{H}(\mathbf{v}_{j-k}^n, \dots, \mathbf{v}_{j+k}^n) \quad (1.2d)$$

where  $\mathbf{H}: \mathbb{R}^{p \times (k+1)} \rightarrow \mathbb{R}^p$  is the *discrete solution operator*. The converse is not true, and when a difference scheme can be written in the form (1.2a) with a numerical flux  $\mathbf{g}$ , it is called *conservative*, and (1.2a) is the *conservation form*. When we restrict to the scalar case, we note  $H$  this discrete solution operator. For the reader's convenience, we recall some basic notions and results that were introduced in the scalar case in G.R., Chapter 3 [539]. We consider sequences  $v = (v_j)_{j \in \mathbb{Z}}$  in  $\ell^1 \equiv \mathbf{L}^1(\mathbb{Z})$  which means their (discrete)

$\mathbf{L}^1$  norm  $\sum_j |v_j|$  is finite. Then  $H(v)$  denotes the sequence with  $j$ th term  $H(v_{j-k}, \dots, v_{j+k})$ .

*Proposition 1.1*

The scheme (1.2d) can be put in conservation form if and only if we have for any sequence  $v = (v_j) \in \mathbf{L}^1(\mathbb{Z})$  such that  $H(v) \in \mathbf{L}^1(\mathbb{Z})$

$$\sum_{j \in \mathbb{Z}} H(v_{j-k}, \dots, v_{j+k}) = \sum_{j \in \mathbb{Z}} v_j.$$

*Proof.* Set  $D(v_{-k}, \dots, v_k) = -\frac{1}{\lambda}(H(v_{-k}, \dots, v_{+k}) - v_0)$ , then  $D$  satisfies for any sequence

$$\sum_{j \in \mathbb{Z}} D(v_{j-k}, \dots, v_{j+k}) = 0.$$

Let us write the above identity for two particular sequences with compact support. First consider a sequence with  $v_j = 0$  for  $j \leq -k$ , or  $j \geq k+1$ , and arbitrary values for  $-k+1 \leq j \leq k$ . Then consider the sequence such that  $v_j = 0$  for  $j \leq -k-1$ , or  $j \geq k+1$ , with the same arbitrary values for  $-k+1 \leq j \leq k$  and arbitrary  $v_k$ , and write the corresponding identity for this sequence. Now define  $g$  by

$$\begin{aligned} g(-v_{k+1}, \dots, v_k) &= D(0, \dots, 0, 0, v_{k+1}) + D(0, \dots, 0, v_{k+1}, v_{k+2}) + \dots \\ &\quad D(0, v_{-k+1}, \dots, v_k). \end{aligned}$$

It is easy to see that we get from the two identities the relation

$$g(-v_k, \dots, v_{k-1}) + D(v_{-k}, \dots, v_k) - g(-v_{k+1}, \dots, v_k) = 0,$$

which leads to the desired result. The only if part is easy; it says that the discrete solution operator preserves the discrete integral (in  $\mathbf{L}^1(\mathbb{Z})$ ).  $\square$

We will admit the important result, known as the Lax–Wendroff theorem (see the proof after Theorem 1.1 in G.R., Chapter 3, Section 1.1 [539], and also this volume, Chap. V, Sect. 4.2.2), which asserts that, in the scalar case, when a conservative consistent scheme “converges” to a function  $u$  (in some sensible way), the limit  $u$  is a weak solution of (1.1) (with  $p = 1$ ). Roughly speaking, this ensures that a conservative scheme computes discontinuities which propagate at the right speed (satisfying the Rankine–Hugoniot condition). In order to precise the convergence, we need to associate with the sequences  $v^n = (v_j^n)$  a piecewise constant function  $\mathbf{v}_\Delta(x, t)$  defined by

$$v_\Delta(x, t) = v_j^n, \quad x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}}, t \in [t_n, t_{n+1}[.$$

### Theorem 1.1

Let (1.2a) be a scheme in conservation form, with consistent numerical flux (1.2c), and let  $v^0$  be given by  $v_j^0 = \frac{1}{|C_j|} \int_{C_j} u_0(x) dx, j \in \mathbb{Z}$ . Assume that there exists a sequence  $\Delta_k x$  which tends to 0 such that if we set  $\Delta_k t = \lambda \Delta_k x$  ( $\lambda$  kept constant),  $\|v_{\Delta_k}\|_{\mathbf{L}^\infty(\mathbb{R} \times (0, \infty))} \leq C$  and  $v_{\Delta_k}$  converges in  $\mathbf{L}^1_{\text{loc}}(\mathbb{R} \times (0, +\infty))$  and a.e. to a function  $u$ . Then  $u$  is a weak solution of (1.1).

Let us recall here the classical notions of truncation error and order of accuracy.

### Definition 1.1

The order of accuracy of scheme (1.2d) is the largest number  $p \geq 1$  such that we have for any smooth solution  $u$  of (1.1) and for  $\lambda = \frac{\Delta t}{\Delta x}$  kept constant

$$u(x, t + \Delta t) - H(u(x - k\Delta x, t), \dots, u(x + k\Delta x, t)) = \mathcal{O}(\Delta t^{p+1}), \text{ as } \Delta t \rightarrow 0.$$

The left-hand side (once divided by  $\Delta t$ ) is called the *truncation error*. Note that this definition links the time and space increments through  $\lambda$ ; this is due to the fact that we are approximating transport equations whose transport speeds are supposed to be finite. The expression “ $\lambda = \frac{\Delta t}{\Delta x}$  kept constant” implies that  $\Delta t$  and  $\Delta x$  are of the same order,  $\lambda = \mathcal{O}(1)$ , and does not go to zero.

The order (of accuracy) of a scheme is evaluated as usual by Taylor expansions, assuming  $u$  is smooth enough. Note that in order to have  $p \geq 0$ , it is necessary that the discrete solution operator satisfies  $H(v, \dots, v) = v$ ,  $\forall v \in \mathbb{R}$ , which means that *constant states are solution* of the scheme. This is satisfied as soon as the scheme is in conservation form. Now, if the numerical flux  $g$  is consistent with the flux  $f$ , one can easily prove that the scheme is (at least) first-order accurate. More precisely, one can prove the following result (see G.R., Chapter 3, Proposition 1.2 [539]).

### Proposition 1.2

Let (1.2d) be a difference scheme which can be put in conservative form (1.2a), with consistent numerical flux, and assume that  $H$  is a  $C^3$  function. Then for any smooth solution  $u$  of (1.1) and for  $\lambda = \frac{\Delta t}{\Delta x}$  kept constant, the truncation error has the following expression:

$$\begin{aligned} u(x, t + \Delta t) - H(u(x - k\Delta x, t), \dots, u(x + k\Delta x, t)) = \\ - \Delta t^2 \partial_x (\beta(u, \lambda) \partial_x u) + \mathcal{O}(\Delta t^3) \text{ as } \Delta t \rightarrow 0, \end{aligned}$$

where

$$\beta(u, \lambda) = \frac{1}{2\lambda^2} \sum_{j=-k}^k j^2 \partial_{v_j} H(u, \dots, u) - \frac{1}{2} f'(u)^2.$$

*Remark 1.2.* The above result ensures that a conservative scheme with consistent flux is consistent in the sense of finite difference schemes. In the case of a nonuniform mesh, one must be more cautious. Indeed, the ratio of increments  $\Delta x_j = |C_j|$  and  $\Delta x_{j+\frac{1}{2}} = x_{j+1} - x_j$  is now involved in the truncation error, and for a general mesh, only consistency in the sense of finite volume schemes holds (see [466], Section 5.3.1, for details).  $\square$

When studying the *stability* of the scheme, we can use the (discrete)  $L^p$  norms for the sequences  $v^n = (v_j^n)$ , in particular the  $L^\infty$  stability which is often obtained by simple arguments, such as convex combination, and the  $L^2$  stability that will be studied below in the linear case, using Fourier transform. In the scalar case, we have also introduced more specific tools, monotonicity and the “total variation diminishing” property, which are important but purely *scalar* notions. We recall the definitions just for the sake of completeness. Scheme (1.2) is *monotone* if given two sequences  $v^0 = (v_j^0)$  and  $w^0 = (w_j^0)$ ,

$$v^0 \geq w^0 \Rightarrow v^1 \geq w^1,$$

where  $v \geq w$  means that  $\forall j$ ,  $v_j \geq w_j$ , and  $v^1 = (v_j^1)$  is the sequence obtained after one step,  $v_j^1 = H(v_{j-k}^0, \dots, v_{j+k}^0)$ . If the function  $H$  is differentiable, the monotone character of a scheme can be analyzed via the sign of the partial derivatives of  $\partial_{v_j} H$ .

Concerning BV stability, it is characterized by what is usually called the *total variation diminishing* (in short TVD) property. Scheme (1.2) is TVD if

$$\forall v^0 = (v_j^0), \quad TV(v^1) \leq TV(v^0),$$

where  $TV(v) = \sum_{j \in \mathbb{Z}} |v_{j+1} - v_j|$  is the total variation of a sequence  $v = (v_j)$ .

TVD schemes are attractive for various reasons. A TVD scheme is monotonicity preserving, i.e., it transforms a monotone sequence, say a nondecreasing one, into a monotone (nondecreasing) sequence, and oscillations cannot occur. But while a monotone scheme is at most first-order accurate (Harten et al. [594]), it is possible to derive “high-order” TVD schemes (at least away from sonic extrema of the solution; see [1007]). Moreover, there exist simple sufficient conditions that ensure that a scalar scheme is TVD when it can be written in incremental or viscous form (which is always the case for all 3-point or essentially 3-point schemes). We say that scheme (1.2a) can be put in *incremental* (resp. *viscous*) form if there exist coefficients  $C, D$  (resp.  $Q$ ) that are functions of  $2k$  variables ( $C, D, Q : \mathbb{R}^{2k} \rightarrow \mathbb{R}$ ) such that setting

$$\Delta v_{j+1/2} = v_{j+1} - v_j, \quad C_{j+\frac{1}{2}} = C(v_{j-k+1}, \dots, v_{j+k})$$

and so on, we have

$$v_j^{n+1} = v_j^n + C_{j+1/2}^n \Delta v_{j+1/2}^n - D_{j-1/2}^n \Delta v_{j-1/2}^n, \quad (1.2e)$$

respectively

$$v_j^{n+1} = v_j^n - \frac{\lambda}{2}(f_{j+1}^n - f_{j-1}^n) + \frac{1}{2}(Q_{j+1/2}^n \Delta v_{j+1/2}^n - Q_{j-1/2}^n \Delta u_{j-1/2}^n). \quad (1.2f)$$

with  $f_j^n = f(v_j^n)$ . Harten's criteria are given by the following proposition.

*Proposition 1.3*

Assume that the scheme (1.2) can be put in incremental (resp. viscous) form and that the incremental (resp. viscosity) coefficients satisfy for any  $j \in \mathbb{Z}$  and  $n \geq 0$

$$C_{j+1/2}^n \geq 0, \quad D_{j+1/2}^n \geq 0, \quad C_{j+1/2}^n + D_{j+1/2}^n \leq 1,$$

(resp.

$$\lambda |\Delta f_{j+1/2}^n / \Delta v_{j+1/2}^n| \leq Q_{j+1/2}^n \leq 1).$$

Then the scheme is TVD.

These conditions happen to be necessary for a 3-point scheme. Note that conservative 3-point second-order accurate methods cannot satisfy a local entropy inequality [1022].

*Remark 1.3.* These notions were introduced to mimic the properties of the exact solution operator, which is order preserving and TVD (see G.R., Chapter 2, Theorem 5.2 [539]). These last properties do not hold for systems, which explains why there is no simple extension of the above notions to numerical schemes for systems. However, since these properties are easily characterized and moreover lead in general to satisfactory results, they are still considered as useful criteria for selecting relevant schemes.  $\square$

We present now the most usual ways of extending to a nonlinear hyperbolic system a finite difference scheme derived in the scalar case. Then, we shall study the  $L^2$  stability of general linear schemes of approximation of a linear hyperbolic system. The following sections will be devoted to a deeper study of the most usual difference schemes, such as Godunov's, Roe's, Osher's, and van Leer's, followed by an introduction to the more recent kinetic and relaxation schemes.

## 1.1 Extension of Scalar Schemes to Systems: Some Examples

Consider the usual 3-point schemes derived in the scalar case. Some of them, the Lax–Friedrichs and the Lax–Wendroff schemes [746], can be generalized immediately to systems.

*Example 1.1. The Lax–Friedrichs scheme.* This scheme is given by

$$\mathbf{v}_j^{n+1} = \frac{1}{2}(\mathbf{v}_{j+1}^n + \mathbf{v}_{j-1}^n) - \frac{\lambda}{2}(\mathbf{f}(\mathbf{v}_{j+1}^n) - \mathbf{f}(\mathbf{v}_{j-1}^n)). \quad (1.3)$$

It is associated with the numerical flux

$$\mathbf{g}^{L.F.}(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v})) - \frac{1}{2\lambda}(\mathbf{v} - \mathbf{u}).$$

It is first-order accurate, and in the scalar case, under the so-called *CFL stability condition* (for Richard Courant, Kurt Friedrichs, and Hans Lewy)

$$\lambda \max_u |f'(u)| \leq 1, \quad (1.4)$$

it is monotone since  $\partial_{v_0} H = 0$  and  $\partial_{v_{\pm 1}} H(v_{-1}, v_0, v_1) = \frac{1}{2}(1 \mp \lambda f'(v_{\pm 1}) \geq 0$  (see G.R., Chapter 3, Examples 2.1 and 3.1 [539]).  $\square$

*Example 1.2. The Lax–Wendroff scheme.* This scheme can be written as

$$\begin{cases} \mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \frac{\lambda}{2}(\mathbf{f}(\mathbf{v}_{j+1}^n) - \mathbf{f}(\mathbf{v}_{j-1}^n)) \\ \quad + \frac{\lambda^2}{2}\{\mathbf{A}_{j+\frac{1}{2}}^n(\mathbf{f}(\mathbf{v}_{j+1}^n) - \mathbf{f}(\mathbf{v}_j^n)) - \mathbf{A}_{j-\frac{1}{2}}^n(\mathbf{f}(\mathbf{v}_j^n) - \mathbf{f}(\mathbf{v}_{j-1}^n))\}. \end{cases} \quad (1.5)$$

Here  $\mathbf{A}_{j+\frac{1}{2}}^n$  is either the Jacobian matrix of  $\mathbf{f}$  evaluated at some average state, for instance,

$$\mathbf{A}_{j+\frac{1}{2}}^n = \mathbf{A}\left(\frac{\mathbf{v}_{j+1}^n + \mathbf{v}_j^n}{2}\right),$$

or

$$\mathbf{A}_{j+\frac{1}{2}}^n = \mathbf{A}(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n),$$

where  $\mathbf{A} = \mathbf{A}(\mathbf{u}, \mathbf{v})$  is a  $p \times p$  matrix satisfying

$$\begin{cases} \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) = \mathbf{A}(\mathbf{u}, \mathbf{v})(\mathbf{v} - \mathbf{u}), \\ \mathbf{A}(\mathbf{u}, \mathbf{u}) = \mathbf{f}(\mathbf{u}). \end{cases} \quad (1.6)$$

Such a matrix  $\mathbf{A}(\mathbf{u}, \mathbf{v})$  is called a *Roe matrix*. The construction and properties of Roe matrices will be discussed in Sect. 3.

The Lax–Wendroff scheme is second-order accurate and is associated with the numerical flux

$$\mathbf{g}^{L.W.}(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v})) - \frac{\lambda}{2} \begin{cases} \mathbf{A}\left(\frac{\mathbf{u} + \mathbf{v}}{2}\right)(\mathbf{f}(\mathbf{v}) - (\mathbf{f}(\mathbf{u}))) \\ \mathbf{A}^2(\mathbf{u}, \mathbf{v})(\mathbf{v} - \mathbf{u}), \end{cases}$$

according to the choice of  $\mathbf{A}_{j+1/2}^n$ . In the scalar case, it is not TVD since its viscosity coefficient  $\lambda^2 A_{j+1/2}^2$  does not satisfy Harten's criteria.

Let us mention a variant of the above scheme, known as the two-step Richtmyer's version, that avoids the computation of the matrices  $\mathbf{A}(\mathbf{u}, \mathbf{v})$  and the products  $\mathbf{Af}$ . This scheme reads

$$\begin{cases} \mathbf{v}_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \frac{1}{2}(\mathbf{v}_{j+1}^n + \mathbf{v}_j^n) - \frac{\lambda}{2}(\mathbf{f}(\mathbf{v}_{j+1}^n) - \mathbf{f}(\mathbf{v}_j^n)), \\ \mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda(\mathbf{f}(\mathbf{v}_{j+\frac{1}{2}}^{n+\frac{1}{2}}) - \mathbf{f}(\mathbf{v}_{j-\frac{1}{2}}^{n+\frac{1}{2}})). \end{cases} \quad (1.7)$$

One can check easily that (1.7) is a second-order scheme and that

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{f}\left(\frac{\mathbf{u} + \mathbf{v}}{2} - \frac{\lambda}{2}(\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}))\right)$$

corresponds to the numerical flux.  $\square$

*Example 1.3. The upwind scheme in the linear case.* First, for a linear scalar advection equation

$$\partial_t u + a \partial_x u = 0,$$

the upwind scheme consists in choosing for numerical flux  $g_{j+\frac{1}{2}}$  the exact upwind flux, i.e.,  $au_j$  if  $a > 0$  and  $au_{j+1}$  if  $a < 0$ . The resulting scheme writes

$$\begin{aligned} v_j^{n+1} &= v_j^n - \lambda a(v_j^n - v_{j-1}^n), \text{ if } a > 0, \\ v_j^{n+1} &= v_j^n - \lambda a(v_{j+1}^n - v_j^n), \text{ if } a < 0. \end{aligned}$$

This choice relies on the idea that the information is propagated along characteristics; indeed the exact solution satisfies  $u(x, t + \Delta t) = u(x - a\Delta x, t)$ . Then  $v_j^{n+1} \approx u(x_j, t_{n+1}) = u(x_j - a\Delta t, t_n)$ , and under CFL condition (1.4), i.e.,  $\lambda|a| \leq 1$ ,  $v_j^{n+1}$  appears as a convex combination of two values, at  $x_j$  and at the node taken on the upstream side, which means that it is obtained by interpolating linearly the value at  $x_j - a\Delta t$  by the two values  $v_{j-1}^n, v_j^n$  when  $a > 0$  and  $v_j^n, v_{j+1}^n$  when  $a < 0$ . This scheme is also called the Courant–Isaacson–Rees scheme. The term  $\nu = a \frac{\Delta t}{\Delta x} = \lambda a$  is called the *Courant number*.

The upwind scheme for a linear system

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0},$$

is not so directly written. A natural way consists in using a scalar scheme for the  $p$  characteristic quantities.

First, let us introduce some notations, which apply in the general case of nonconstant matrices. Denote by  $\mathbf{r}_1(\mathbf{u}), \dots, \mathbf{r}_p(\mathbf{u})$  (respectively  $\mathbf{l}_1(\mathbf{u}), \dots, \mathbf{l}_p(\mathbf{u})$ ) a complete system of (right) eigenvectors of  $\mathbf{A}(\mathbf{u})$  (resp.  $\mathbf{A}^T(\mathbf{u})$ ), forming a dual basis of  $\mathbb{R}^p$

$$\mathbf{l}_j^T \cdot \mathbf{r}_k = \delta_j^k.$$

The matrices  $\mathbf{T}(\mathbf{u})$  with columns  $(\mathbf{r}_1(\mathbf{u}), \dots, \mathbf{r}_p(\mathbf{u}))$  and  $\mathbf{T}^{-1}(\mathbf{u})$  with rows  $(\mathbf{l}_1^T(\mathbf{u}), \dots, \mathbf{l}_p^T(\mathbf{u}))$  satisfy

$$\mathbf{T}^{-1}(\mathbf{u})\mathbf{A}(\mathbf{u})\mathbf{T}(\mathbf{u}) = \text{diag}(a_i(\mathbf{u})) \equiv \mathbf{A}(\mathbf{u}). \quad (1.8)$$

Now, let us set

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

so that

$$a = a^+ + a^-, \quad |a| = a^+ - a^-,$$

and denote

$$\mathbf{A}^\pm(\mathbf{u}) = \text{diag}(a_k^\pm(\mathbf{u})).$$

If we now define  $\mathbf{A}^+$  and  $\mathbf{A}^-$  by

$$\mathbf{A}^\pm(\mathbf{u}) = \mathbf{T}(\mathbf{u})\mathbf{A}^\pm(\mathbf{u})\mathbf{T}^{-1}(\mathbf{u}), \quad (1.9)$$

clearly we have

$$\mathbf{A}(\mathbf{u}) = \mathbf{A}^+(\mathbf{u}) + \mathbf{A}^-(\mathbf{u}), \quad |\mathbf{A}(\mathbf{u})| = \mathbf{A}^+(\mathbf{u}) - \mathbf{A}^-(\mathbf{u}).$$

Assume now that  $\mathbf{A}$  is constant; introducing as in Chap. II, Sect. 1, the *characteristic* variables

$$\mathbf{w} = \mathbf{T}^{-1}\mathbf{u}, \quad \text{i.e., } \mathbf{w}_k = \mathbf{1}_k^T \mathbf{u},$$

we get a system of  $p$  decoupled scalar equations,

$$\frac{\partial \mathbf{w}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{w}}{\partial x} = \mathbf{0}.$$

This gives a natural way of extending the scalar upwind scheme to a (linear) system by applying it to each scalar characteristic equation. The upwind scheme applied to each equation gives

$$\mathbf{w}_j^{n+1} = \mathbf{w}_j^n - \frac{\lambda}{2} \mathbf{A}(\mathbf{w}_{j+1}^n - \mathbf{w}_{j-1}^n) + \frac{\lambda}{2} |\mathbf{A}|(\mathbf{w}_{j+1}^n - 2\mathbf{w}_j^n + \mathbf{w}_{j-1}^n).$$

Now, we need to write the scheme in the original conservative variables, and setting  $\mathbf{v}_j^n = \mathbf{T}\mathbf{w}_j^n$ , we check that we get the formula

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \frac{\lambda}{2} \mathbf{A}(\mathbf{v}_{j+1}^n - \mathbf{v}_{j-1}^n) + \frac{\lambda}{2} |\mathbf{A}|(\mathbf{v}_{j+1}^n - 2\mathbf{v}_j^n + \mathbf{v}_{j-1}^n). \quad (1.10a)$$

We can also write it in upwind form

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \{ \mathbf{A}^+(\mathbf{v}_{j+1}^n - \mathbf{v}_j^n) + \mathbf{A}^-(\mathbf{v}_j^n - \mathbf{v}_{j-1}^n) \} \quad (1.10b)$$

with numerical flux

$$\mathbf{g}^u(\mathbf{u}, \mathbf{v}) = \mathbf{A}^+\mathbf{u} + \mathbf{A}^-\mathbf{v}.$$

This scheme has various extensions in the nonlinear case such as Godunov's and Roe's scheme, which we shall detail in the next sections.  $\square$

For second-order schemes with flux limiters using the Lax–Wendroff and the above upstream scheme as building block, such as Sweby's or Davis' (see G.R., Chapter 4, Section 2 [539] and [392, 877, 1079]), they can be extended in two ways. One way consists in replacing the ratio of consecutive increments  $r$ , for instance, by

$$\begin{aligned} r_{j+\frac{1}{2}}^{n+} &= \frac{(\Delta\mathbf{v}_{j-\frac{1}{2}}^n, \Delta\mathbf{v}_{j+\frac{1}{2}}^n)}{(\Delta\mathbf{v}_{j+\frac{1}{2}}^n, \Delta\mathbf{v}_{j+\frac{1}{2}}^n)}, \\ r_{j+\frac{1}{2}}^{n-} &= \frac{(\Delta\mathbf{v}_{j-\frac{1}{2}}^n, \Delta\mathbf{v}_{j+\frac{1}{2}}^n)}{(\Delta\mathbf{v}_{j-\frac{1}{2}}^n, \Delta\mathbf{v}_{j-\frac{1}{2}}^n)}, \end{aligned}$$

where  $(.,.)$  denotes the Euclidean inner product in  $\mathbb{R}^p$  (see Davis [392], for details). Another way consists in modifying the flux componentwise (relative to the eigenvector decomposition), as we shall now explain.

*Example 1.4. Schemes with flux limiters.* Assume first that  $\mathbf{A}$  is constant. Then, the Lax–Wendroff scheme (1.5) corresponds to the numerical flux

$$\mathbf{g}^{L.W}(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\mathbf{A}(\mathbf{u} + \mathbf{v}) - \frac{\lambda}{2}\mathbf{A}^2(\mathbf{v} - \mathbf{u}).$$

Since  $\mathbf{A} = \mathbf{A}^+ + \mathbf{A}^-$ , we can write

$$\frac{1}{2}\mathbf{A}(\mathbf{u} + \mathbf{v}) = \mathbf{A}^+\mathbf{u} + \mathbf{A}^-\mathbf{v} + \frac{1}{2}|\mathbf{A}|(\mathbf{v} - \mathbf{u})$$

and

$$\mathbf{g}^{L.W}(\mathbf{u}, \mathbf{v}) = \mathbf{A}^+\mathbf{u} + \mathbf{A}^-\mathbf{v} + \frac{1}{2}(\mathbf{I} - \lambda|\mathbf{A}|)|\mathbf{A}|(\mathbf{v} - \mathbf{u}).$$

Then, we notice that we have  $\mathbf{A}^+\mathbf{A}^- = \mathbf{0}$ , which implies

$$(\mathbf{I} - \lambda|\mathbf{A}|)|\mathbf{A}| = (\mathbf{I} - \lambda\mathbf{A}^+)\mathbf{A}^+ - (\mathbf{I} + \lambda\mathbf{A}^-)\mathbf{A}^-.$$

Hence, the Lax–Wendroff flux can be written in terms of the upwind flux (see (1.10)),

$$\mathbf{g}^{L.W}(\mathbf{u}, \mathbf{v}) = \mathbf{g}^u(\mathbf{u}, \mathbf{v}) + \frac{1}{2}(\mathbf{I} - \lambda\mathbf{A}^+)\mathbf{A}^+(\mathbf{v} - \mathbf{u}) - \frac{1}{2}(\mathbf{I} + \lambda\mathbf{A}^-)\mathbf{A}^-(\mathbf{v} - \mathbf{u}).$$

Setting

$$\nu_k = \lambda a_k, \quad \sigma_k = \operatorname{sgn} a_k$$

and writing  $\Delta\mathbf{v}_{j+\frac{1}{2}}$  in the basis of eigenvectors

$$\Delta\mathbf{v}_{j+\frac{1}{2}} = \sum_{k=1}^p \alpha_{k,j+\frac{1}{2}} \mathbf{r}_{k,j+\frac{1}{2}},$$

we get

$$\begin{aligned}\mathbf{g}_{j+\frac{1}{2}}^{L.W.} &= \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2} \left( \sum_{a_k \geq 0} (1 - \lambda a_k) a_k \alpha_{k,j+\frac{1}{2}} \mathbf{r}_k \right. \\ &\quad \left. - \sum_{a_k < 0} (1 + \lambda a_k) a_k \alpha_{k,j+\frac{1}{2}} \mathbf{r}_k \right)\end{aligned}$$

or

$$\mathbf{g}_{j+\frac{1}{2}}^{L.W.} = \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2} \sum_{k=1}^p (\sigma_k - \nu_k) a_k \alpha_{k,j+\frac{1}{2}} \mathbf{r}_k.$$

We then limit the flux in each characteristic direction

$$\mathbf{g}_{j+\frac{1}{2}} = \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2} \sum_{k=1}^p \varphi_{k,j+\frac{1}{2}} (\sigma_k - \nu_k) a_k \alpha_{k,j+\frac{1}{2}} \mathbf{r}_k, \quad (1.11)$$

where the limiter  $\varphi_{k,j+\frac{1}{2}}$  satisfies

$$\varphi_{k,j+\frac{1}{2}} = \varphi(r_{k,j+\frac{1}{2}}), \quad r_{k,j+\frac{1}{2}} = \begin{cases} \frac{\alpha_{k,j-\frac{1}{2}}}{\alpha_{k,j+\frac{1}{2}}} & \text{if } \sigma_k = \operatorname{sgn} a_k > 0, \\ \frac{\alpha_{k,j+\frac{3}{2}}}{\alpha_{k,j+\frac{1}{2}}} & \text{if } \sigma_k = \operatorname{sgn} a_k < 0, \end{cases}$$

and  $\varphi$  is, for instance,

$$\varphi(r) = \max\{0, \min(\Phi r, 1), \min(r, \Phi)\}, \quad \Phi \in [1, 2].$$

For  $\Phi = 1$ , this is the usual minmod limiter since  $\varphi_{j+\frac{1}{2}} r_{j+\frac{1}{2}} = \operatorname{minmod}(\Delta v_{j+\frac{1}{2}}, \Delta v_{j-\frac{1}{2}})$ , where

$$\operatorname{minmod}(a, b) = \begin{cases} s \min(|a|, |b|) & \text{if } s = \operatorname{sgn}(a) = \operatorname{sgn}(b), \\ 0 & \text{otherwise.} \end{cases}$$

For  $\Phi = 2$ , we get the so-called superbee limiter (see G.R., Chapter 4, Section 2.3 [539]).

In the general nonlinear case (following the ideas of G.R., Chapter 4, Section 2.2), we set

$$\Delta \mathbf{f}_{j+\frac{1}{2}}^+ = \mathbf{f}_{j+1} - \mathbf{g}_{j+\frac{1}{2}}^u, \quad \Delta \mathbf{f}_{j+\frac{1}{2}}^- = \mathbf{g}_{j+\frac{1}{2}}^u - \mathbf{f}_j,$$

where  $\mathbf{g}^u$  is the numerical flux of Roe's or Godunov's scheme, which we shall discuss in the next sections. Then, from (1.5), we obtain

$$\begin{aligned}\mathbf{g}_{j+\frac{1}{2}}^{L.W.} &= \frac{1}{2}(\mathbf{f}_{j+1} + \mathbf{f}_j) - \frac{\lambda}{2}\mathbf{A}_{j+\frac{1}{2}}(\mathbf{f}_{j+1} - \mathbf{f}_j) \\ &= \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2}(\Delta\mathbf{f}_{j+\frac{1}{2}}^+ - \Delta\mathbf{f}_{j+\frac{1}{2}}^-) \\ &\quad - \frac{\lambda}{2}(\mathbf{A}_{j+\frac{1}{2}}^+ + \mathbf{A}_{j+\frac{1}{2}}^-)(\Delta\mathbf{f}_{j+\frac{1}{2}}^+ + \Delta\mathbf{f}_{j+\frac{1}{2}}^-).\end{aligned}$$

If we assume

$$\mathbf{A}^+\Delta\mathbf{f}^- = \mathbf{A}^-\Delta\mathbf{f}^+ = 0, \quad (1.12)$$

which indeed holds in the linear case, we get

$$\mathbf{g}_{j+\frac{1}{2}}^{L.W.} = \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2}(\mathbf{I} - \lambda\mathbf{A}_{j+\frac{1}{2}}^+)\Delta\mathbf{f}_{j+\frac{1}{2}}^+ - \frac{1}{2}(\mathbf{I} + \lambda\mathbf{A}_{j+\frac{1}{2}}^-)\Delta\mathbf{f}_{j+\frac{1}{2}}^-.$$

Now, as previously, we decompose  $\Delta\mathbf{v}_{j+\frac{1}{2}}$  on the basis  $\mathbf{r}_{k,j+\frac{1}{2}}$  of eigenvectors of  $\mathbf{A}_{j+\frac{1}{2}}$ ,

$$\begin{aligned}\mathbf{A}_{j+\frac{1}{2}}\mathbf{r}_{k,j+\frac{1}{2}} &= a_{k,j+\frac{1}{2}}\mathbf{r}_{k,j+\frac{1}{2}}, \\ \Delta\mathbf{v}_{j+\frac{1}{2}} &= \sum_{k=1}^p \alpha_{k,j+\frac{1}{2}}\mathbf{r}_{k,j+\frac{1}{2}}\end{aligned}$$

(where the dependence of  $\mathbf{A}, a, \alpha$ , and  $\mathbf{r}$  on  $\mathbf{u}$  is omitted in the notations). We get

$$\begin{aligned}\mathbf{g}_{j+\frac{1}{2}}^{L.W.} &= \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2} \sum_{a_k>0} (1 - \lambda a_{k,j+\frac{1}{2}}) a_{k,j+\frac{1}{2}} \alpha_{k,j+\frac{1}{2}} \mathbf{r}_{k,j+\frac{1}{2}} \\ &\quad - \frac{1}{2} \sum_{a_k<0} (1 + \lambda a_{k,j+\frac{1}{2}}) a_{k,j+\frac{1}{2}} \alpha_{k,j+\frac{1}{2}} \mathbf{r}_{k,j+\frac{1}{2}},\end{aligned}$$

or

$$\mathbf{g}_{j+\frac{1}{2}}^{L.W.} = \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2} \sum_{k=1}^p (\sigma_{k,j+\frac{1}{2}} - \nu_{k,j+\frac{1}{2}}) a_{k,j+\frac{1}{2}} \alpha_{k,j+\frac{1}{2}} \mathbf{r}_{k,j+\frac{1}{2}}.$$

Finally, we limit the flux as in formula (1.11),

$$\mathbf{g}_{j+\frac{1}{2}} = \mathbf{g}_{j+\frac{1}{2}}^u + \frac{1}{2} \sum_{k=1}^p \varphi(r_{k,j+\frac{1}{2}})(\sigma_{k,j+\frac{1}{2}} - \nu_{k,j+\frac{1}{2}}) a_{k,j+\frac{1}{2}} \alpha_{k,j+\frac{1}{2}} \mathbf{r}_{k,j+\frac{1}{2}},$$

where

$$r_{k,j+\frac{1}{2}} = \begin{cases} \frac{((\sigma - \nu)a\alpha)_{k,j-\frac{1}{2}}}{((\sigma - \nu)a\alpha)_{k,j+\frac{1}{2}}} & \text{if } \sigma_{k,j+\frac{1}{2}} > 0, \\ \frac{((\sigma - \nu)a\alpha)_{k,j+\frac{3}{2}}}{((\sigma - \nu)a\alpha)_{k,j+\frac{1}{2}}} & \text{if } \sigma_{k,j+\frac{1}{2}} < 0. \end{cases}$$

Note that we can use the same (Hui and Loh [640, 641]) or a different (Swanson and Turkel [1078], Jorgenson and Turkel [671]) limiter in each characteristic variable (see also Yee [1205]).  $\square$

Similarly we can easily extend the second-order schemes with the “modified flux” of Harten by modifying the flux componentwise following the ideas developed in the scalar case (see G.R., Chapter 4, Section 1 [539]).

*Example 1.5. Schemes with “modified flux.”* As for the Lax–Wendroff scheme, we introduce an average value, more precisely a symmetric averaging operator  $V$  and the corresponding average state

$$\bar{\mathbf{v}}_{j+\frac{1}{2}} = V(\mathbf{v}_j, \mathbf{v}_{j+1}),$$

and set

$$\mathbf{r}_{k,j+\frac{1}{2}} = \mathbf{r}_k(\bar{\mathbf{v}}_{j+\frac{1}{2}}).$$

Then, we define

$$\alpha_{k,j+\frac{1}{2}} = \mathbf{l}_k^T(\bar{\mathbf{v}}_{j+\frac{1}{2}}) \Delta \mathbf{v}_{j+\frac{1}{2}}$$

so that

$$\Delta \mathbf{v}_{j+\frac{1}{2}} = \sum_{k=1}^p \alpha_{k,j+\frac{1}{2}} \mathbf{r}_{k,j+\frac{1}{2}}.$$

Let us set, in complete analogy with the scalar case,

$$a_{k,j+\frac{1}{2}} = a_k(\bar{\mathbf{v}}_{j+\frac{1}{2}}).$$

If using Roe’s extension technique, we shall rather take for  $a_{k,j+\frac{1}{2}}$  and  $\mathbf{r}_{k,j+\frac{1}{2}}$  the eigenvalues and eigenvectors of an average matrix  $\mathbf{A}(\mathbf{v}_j, \mathbf{v}_{j+1})$  satisfying (1.6). As we have already written, such a matrix will be constructed in Sect. 3. Then, the formula (1.6) of G.R., Chapter 4, Section 1 [539], leads us to introduce the scheme

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \frac{\lambda}{2} (\tilde{\mathbf{f}}_{j+1}^n - \tilde{\mathbf{f}}_{j-1}^n) + \frac{1}{2} (\tilde{\mathbf{Q}}_{j+\frac{1}{2}}^n \Delta \mathbf{v}_{j+\frac{1}{2}}^n - \tilde{\mathbf{Q}}_{j-\frac{1}{2}}^n \Delta \mathbf{v}_{j-\frac{1}{2}}^n).$$

In this formula, the modified flux is

$$\tilde{\mathbf{f}}_j = \mathbf{f}(\mathbf{v}_j) + \frac{\mathbf{h}_j}{\lambda},$$

where

$$\mathbf{h}_j = \sum_{k=1}^p h_{k,j} \mathbf{r}_{k,j+\frac{1}{2}},$$

and the viscosity term

$$\tilde{\mathbf{Q}}_{j+\frac{1}{2}} \Delta \mathbf{v}_{j+\frac{1}{2}} = \sum_{k=1}^p Q_k \left( \lambda a_{k,j+\frac{1}{2}} + \frac{\Delta h_{k,j+\frac{1}{2}}}{\alpha_{k,j+\frac{1}{2}}} \right) \alpha_{k,j+\frac{1}{2}} \mathbf{r}_{k,j+\frac{1}{2}}.$$

Here, following formula (1.12) in G.R. [539], Chapter 4,  $h_{k,j}$  is defined by

$$h_{k,j} = \frac{1}{2} s_{k,j} \min \left\{ (Q_k(\lambda a_{k,j+\frac{1}{2}}) - \lambda^2 a_{k,j+\frac{1}{2}}^2) |\alpha_{k,j+\frac{1}{2}}|, (Q_k(\lambda a_{k,j-\frac{1}{2}}) - \lambda^2 a_{k,j-\frac{1}{2}}^2) |\alpha_{k,j-\frac{1}{2}}| \right\},$$

$$s_{k,j} = \begin{cases} \operatorname{sgn} \alpha_{k,j \pm \frac{1}{2}} & \text{if } \alpha_{k,j-\frac{1}{2}} \alpha_{k,j+\frac{1}{2}} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the function  $Q_k$ , still satisfying

$$|x| \leq Q_k(x) \leq 1,$$

may depend on the characteristic field. We obtain by comparison with the Lax–Wendroff scheme an overall second-order accurate scheme. Furthermore, we can define a  $TV$  discrete norm by

$$TV(\mathbf{v}_{j+\frac{1}{2}}) = \sum_j |\Delta \mathbf{v}_{j+\frac{1}{2}}|,$$

where, for instance, the norm is the discrete  $\mathbf{L}^1$  norm

$$|\Delta \mathbf{v}_{j+\frac{1}{2}}| = \sum_{k=1}^p |\alpha_{k,j+\frac{1}{2}}|.$$

Then, it is easy to show that, at least in the linear case, the resulting scheme is TVD under the same CFL restriction as in the scalar case (see Harten [590, 591], Yee et al. [1208]).  $\square$

*Remark 1.4.* Let us note that if a scheme is written in viscous form, extending (1.2f) to systems, we write

$$\begin{cases} \mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \frac{\lambda}{2} (\mathbf{f}(\mathbf{v}_{j+1}^n) - \mathbf{f}(\mathbf{v}_{j-1}^n)) \\ \quad + \frac{1}{2} (\mathbf{Q}_{j+\frac{1}{2}}^n (\mathbf{v}_{j+1}^n - \mathbf{v}_j^n) - \mathbf{Q}_{j-\frac{1}{2}}^n (\mathbf{v}_j^n - \mathbf{v}_{j-1}^n)), \end{cases} \quad (1.13)$$

where the viscosity coefficient is replaced by a viscosity matrix  $\mathbf{Q}$ . The matrix may be diagonal,  $\mathbf{Q} = \mathbf{I}$  for Lax–Friedrichs’ scheme or  $\mathbf{Q} = \operatorname{diag}(Q_k)$  as above when working on the characteristic variables. But for the Lax–Wendroff scheme using Roe’s average, the viscosity matrix is not diagonal, and we have

$$\mathbf{Q}_{j+\frac{1}{2}}^{LW} = \lambda^2 \mathbf{A}^2(\mathbf{v}_{j+1}, \mathbf{v}_j)$$

(see also Einfeldt [457]). Then, for a general scheme, if  $\mathbf{Q}$  is smooth enough, second-order accuracy requires

$$\mathbf{Q}(\mathbf{u}, \dots, \mathbf{u}) = \lambda^2 \mathbf{A}^2(\mathbf{u}).$$

Also, as a generalization of the results of G.R., Chapter 3, Section 4 [539], concerning the schemes satisfying an entropy condition, we can compare viscosity matrices (see Tadmor [1090, 1092] for details and the more recent [1093]).  $\square$

## 1.2 $L^2$ Stability

Consider the linear hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, \quad (1.14)$$

where  $\mathbf{A}$  is a constant  $p \times p$  matrix, and the following linear difference scheme (as in G.R. Chapter 3 Section 1.3):

$$\mathbf{v}_j^{n+1} = \sum_{\ell=-k}^{+k} \mathbf{C}_\ell \mathbf{v}_{j+\ell}^n, \quad (1.15)$$

where the  $\mathbf{C}_\ell, -k \leq \ell \leq k$  are  $p \times p$  constant matrices that are in fact polynomials in  $\lambda \mathbf{A}$ . For the sake of convenience, we extend this scheme to the whole real line by setting

$$\mathbf{v}^{n+1}(x) = \sum_{\ell=-k}^{+k} \mathbf{C}_\ell \mathbf{v}^n(x + \ell \Delta x), \quad (1.16)$$

where

$$\mathbf{v}^n(x) = \mathbf{v}_\Delta(x, t_n) = \mathbf{v}_j^n, \quad x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}},$$

meaning that we have extended the sequence  $\mathbf{v}^n$  as a piecewise constant function  $\mathbf{v}_\Delta(., t_n)$  defined by

$$\mathbf{v}_\Delta(x, t) = \mathbf{v}_j^n, \quad x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}}, t \in [t_n, t_{n+1}[.$$

Then, the scheme (1.15) is  $L^2$ -stable if there exists a constant  $C$  independent of  $\Delta x$  and  $\Delta t$  such that

$$\|\mathbf{v}^n\|_{(\mathbf{L}^2(\mathbb{R}))^p} \leq C \|\mathbf{v}^0\|_{(\mathbf{L}^2(\mathbb{R}))^p}. \quad (1.17)$$

By using the Fourier transform

$$\hat{\varphi}(\xi) = (2\pi)^{-\frac{1}{2}} \int_{\mathbb{R}} e^{-ix\xi} \varphi(x) dx,$$

(1.15) can be equivalently written as

$$\hat{\mathbf{v}}^{n+1}(\xi) = \mathbf{G}^a(\xi)\hat{\mathbf{v}}^n(\xi), \quad (1.18)$$

where the matrix

$$\mathbf{G}^a(\xi) = \sum_{\ell=-k}^{+k} \mathbf{C}_\ell e^{i\ell\Delta x \xi} \quad (1.19)$$

is called the *amplification matrix* of the scheme. Iterating on (1.18) gives

$$\hat{\mathbf{v}}^n(\xi) = (\mathbf{G}^a(\xi))^n \hat{\mathbf{v}}^0(\xi).$$

The  $L^2$  stability property thus requires that the powers of the matrix  $\mathbf{G}^a(\xi)$  are uniformly bounded, i.e.,  $\mathbf{G}^a$  satisfies the “power boundedness condition”

$$\|(\mathbf{G}^a(\xi))^n\|_{\mathcal{L}(\mathbf{L}^2)} \leq C, \quad \forall n. \quad (1.20)$$

Let  $\rho(\mathbf{G})$  denote the spectral radius of a matrix  $\mathbf{G}$ . We have the following necessary and sufficient condition.

*Proposition 1.4*

Assume that  $\mathbf{A}$  is a constant matrix. The linear difference scheme (1.5) is  $L^2$ -stable if and only if its amplification matrix  $\mathbf{G}^a$  defined by (1.19) satisfies

$$\rho(\mathbf{G}^a(\xi)) \leq 1, \quad \forall \xi \in \mathbb{R}. \quad (1.21)$$

*Proof.* Since by assumption the coefficient matrices  $\mathbf{C}_\ell$  are polynomials in  $\lambda\mathbf{A}$ , the matrices  $\mathbf{A}$  and  $\mathbf{C}_\ell$  are all simultaneously diagonalizable by  $\mathbf{T}$  (see (1.8)), and so are the  $\mathbf{G}^a(\xi)^n$ . The result follows easily.  $\square$

*Remark 1.5.* If we assume that  $\mathbf{A}$  is no longer constant but depends on  $(x, t)$ , a general necessary stability condition is given by

$$\rho(\mathbf{G}^a(\xi))^n \leq C, \quad \forall n,$$

or, equivalently, by

$$\rho(\mathbf{G}^a(\xi)) \leq 1 + c\Delta t, \quad 0 < \Delta t < \Delta t_0, \quad \forall \xi. \quad (1.22)$$

This condition is known as the *von Neumann condition*. Condition (1.21), which is also sufficient in our case, is known as the strict von Neumann condition. There exists an extensive literature on the  $L^2$  stability of difference schemes. Indeed the “power boundedness condition” in the more general context has been characterized by Kreiss [710]; see also Lax and Wendroff [747], Richtmyer and Morton [974], and then Tadmor [1086] and LeVeque and Trefethen [783].

Sufficient stability conditions are obtained by introducing dissipation in the difference scheme, as we shall discuss below.  $\square$

*Remark 1.6.* Now assume that we have a 3-point linear scheme written in viscous form (1.13)

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \frac{\lambda}{2} \mathbf{A}(\mathbf{v}_{j+1}^n - \mathbf{v}_{j-1}^n) + \frac{1}{2} \mathbf{Q}(\mathbf{v}_{j+1}^n - 2\mathbf{v}_j^n + \mathbf{v}_{j-1}^n),$$

where  $\mathbf{Q}$  is the constant (viscosity) matrix

$$\mathbf{Q} = \mathbf{C}_{-1} + \mathbf{C}_1 = \mathbf{I} - \lambda \mathbf{C}_0 \mathbf{A}.$$

Then  $\mathbf{Q}$  has the same eigenvectors as  $\mathbf{A}$ ; thus, it is diagonalizable with real eigenvalues  $Q_k$ , and the condition for  $L^2$  stability can be written as

$$(\lambda a_k)^2 \leq Q_k \leq 1, \quad k = 1, \dots, p. \quad (1.23)$$

Indeed, by using the basis of eigenvectors of  $\mathbf{A}$ , the eigenvalues of  $\mathbf{G}^a(\xi)$  are easily computed in terms of  $Q_k$  and  $a_k$ , and the proof then mimics that of G.R., Proposition 1.4, Chapter 3 [539]. For the unique second-order (linear) scheme, the condition becomes

$$\lambda \rho(\mathbf{A}) \leq 1,$$

which is a CFL-like condition.  $\square$

We shall now study the dissipative and dispersive characters of a numerical scheme using Fourier analysis or discrete Fourier modes; we shall also introduce the equivalent system.

### 1.3 Dissipation and Dispersion

Let us study first the differential system (1.14). After applying a Fourier transform, it becomes

$$\frac{\partial \hat{\mathbf{u}}}{\partial t} + i\xi \mathbf{A} \hat{\mathbf{u}} = 0.$$

The exact amplification matrix

$$\mathbf{G}^{ex}(\xi) = \exp(-i\xi \Delta t \mathbf{A}),$$

such that

$$\hat{\mathbf{u}}(\xi, t + \Delta t) = \mathbf{G}^{ex}(\xi) \hat{\mathbf{u}}(\xi, t)$$

obviously satisfies

$$\rho(\mathbf{G}^{ex}(\xi)) = 1.$$

One can equivalently see that (1.14) admits elementary solutions of the form

$$\mathbf{u}(x, t) = \hat{\mathbf{u}} e^{i(kx - \omega t)}$$

iff the phase velocity  $\omega/k$  agrees with an eigenvalue of  $\mathbf{A}$ . Since these eigenvalues are real, the amplitude remains constant.

For a difference scheme, dissipation or diffusion yields the attenuation (damping) of the amplitude of the Fourier components of a discrete solution. Following Kreiss [709], we define more precisely the dissipative character of difference schemes.

*Definition 1.2*

The linear difference scheme (1.15) is called dissipative (in the sense of Kreiss) of order  $2q$  if there exists some constant  $\delta > 0$  such that its amplification matrix (1.19) satisfies

$$\rho(\mathbf{G}^a(\xi)) \leq 1 - \delta|\xi|^{2q}, \quad \forall \xi \in \mathbb{R}, |\xi| \leq \pi. \quad (1.24)$$

In fact, we take for  $q$  the smallest integer such that (1.24) is satisfied.

As for the scalar case, it is easy to prove that a 3-point linear scheme is dissipative with order 4 iff it is  $L^2$ -stable (i.e., (1.22) holds) and

$$0 < Q_k < 1, \quad \forall k = 1, \dots, p,$$

where  $Q_k$  are the eigenvalues of  $\mathbf{Q}$ , and dissipative with order 2 iff

$$(\lambda a_k)^2 < Q_k < 1.$$

Thus, a second-order accurate  $L^2$ -stable linear scheme is never dissipative with order 2 but is dissipative with order 4 iff  $\mathbf{A}$  has no zero eigenvalue and

$$\rho(\lambda \mathbf{A}) < 1.$$

Note also that if a general scheme is linearized (see G.R., Chapter 3, Remark 1.3 [539]) by freezing the coefficients at some given state, it will not be linearly dissipative at a point  $\mathbf{u}$  where the matrix  $\mathbf{A}(\mathbf{u})$  is singular. Again, for a further study of dissipative schemes, we refer to Kreiss and Richtmyer and Morton [974] (see also G.R., Chapter 5, Section 5.1).

Dissipation or diffusion corresponds to an even-order term in the truncation error or in the equivalent system (see G.R., Chapter 3, Remark 1.2 [539]). Indeed, for a first-order scheme, the *equivalent* (or *modified*) system, which the scheme approximates with second-order accuracy, can be written as

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \lambda \Delta x \frac{\partial}{\partial x} \left( \mathbf{B}(\mathbf{u}, \lambda) \frac{\partial \mathbf{u}}{\partial x} \right),$$

where

$$\mathbf{B}(\mathbf{u}, \lambda) = \frac{1}{2\lambda^2} \left\{ \sum_j j^2 \frac{\partial \mathbf{H}}{\partial v_j}(\mathbf{u}, \mathbf{u}, \dots) - \lambda^2 \mathbf{A}^2(\mathbf{u}) \right\}.$$

In this formula,  $\mathbf{H} : \mathbb{R}^{p \times (k+1)} \rightarrow \mathbb{R}^p$  is the discrete solution operator in (1.2d), and  $\mathbf{B}(\mathbf{u}, \lambda)$  is the analog for system of the function  $\beta$  introduced in Proposition 1.2. For instance, for the Lax–Friedrichs (linear) scheme, we find the diffusive system

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \frac{\Delta x}{2\lambda} (1 - \lambda^2 \mathbf{A}^2) \frac{\partial^2 \mathbf{u}}{\partial x^2}.$$

For a second-order scheme, the equivalent system involves third-order differential terms (see Lerat [761]), and we can then study the dispersive character of the scheme. A scheme is *dispersive* if different Fourier components of the solution travel at different speeds. Consider first the continuous problem, i.e., a partial differential equation that stands for either the conservation law or the equivalent equation of a numerical scheme (see [1181]). By linearity, it is sufficient as above to consider a unique “Fourier mode”

$$\mathbf{u}(x, t) = \hat{\mathbf{u}} e^{i(kx - \omega t)},$$

where  $\varphi = kx - \omega t$  is the phase and  $\omega = \omega(k)$  is the frequency, which is a function of the *wave number*  $k$ . This function  $\omega(k)$  is obtained by substituting this mode  $\mathbf{u}$  in the equation, which gives the “dispersion relation” between  $\omega$  and  $k$ . If this function is imaginary, it does not correspond to the description of wave phenomena, and this analysis is irrelevant. The *phase velocity* is  $\frac{\omega(k)}{k}$ , and the waves are called dispersive if the phase velocity is not constant but depends on  $k$ . For instance, if  $u$  is a solution of the linear scalar advection equation

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

we find  $\omega(k) = ak$ , the phase velocity is constant, and the amplitude does not decay, as we have already seen above. For a diffusive equation

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} - b \frac{\partial^2 u}{\partial x^2} = 0,$$

we get  $\omega(k) = ak - ibk^2$ . The imaginary part yields a damping of the amplitude by a factor  $\exp(-bk^2)$  but does not change the phase velocity, whereas if  $u$  is solution of a dispersive equation with an odd-order term,

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + c \frac{\partial^3 u}{\partial x^3} = 0,$$

we have  $\omega(k) = ak - ck^3$  and  $\frac{\omega(k)}{k} = a - ck^2$ . In fact, when there is a superposition of wave trains of different wave numbers  $k$ , it is the *group velocity* defined by  $\omega'(k) = \frac{d\omega(k)}{dk}$  that is the propagation velocity for the wave number  $k$  (see Whitham (1974), Chapter 11 [1188]).

For example, the equivalent equation for the Lax–Wendroff scheme,

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = -\frac{1}{6} a (\Delta x)^2 (1 - \lambda^2 a^2) \frac{\partial^3 u}{\partial x^3},$$

is dispersive with  $c = \frac{1}{6} a (\Delta x)^2 (1 - \lambda^2 a^2) > 0$  if we assume  $a > 0$  and a CFL condition  $\lambda a < 1$ , and we have  $\omega(k)/k = a - ck^2 < a$ , i.e., a lagging phase error opposite to a leading phase error when  $\omega(k)/k > a$ . In general, the lower-order odd (resp. even) term gives the leading dispersive (resp. diffusive) error term. For details, we refer to Richtmyer and Morton [974] and LeVeque [774]; concerning group velocity in finite difference schemes, see Trefethen [1132]; for  $S_\alpha^\beta$  schemes and the study of  $k$ -dispersion relative to the  $k$ th characteristic field, see Lerat [761]; for some examples, see Song and Tang [1070], Desideri et al. [413], and Beux et al. [145].

We can also observe the phase error directly on the equation of the difference scheme. We can either follow the lines of the preceding section (1.16)–(1.19) or look for elementary (discrete Fourier modes) solutions of the form

$$u_j^n = \hat{u} e^{i(kj\Delta x - \omega n \Delta t)}.$$

We obtain a discrete dispersion relation  $\omega = \omega(k, \Delta x, \Delta t)$  by substituting this single Fourier mode in the formula defining the scheme. On the one hand,

$$u_j^{n+1} = \hat{u} e^{i(kj\Delta x - \omega(n+1)\Delta t)} = e^{-i\omega\Delta t} u_j^n,$$

and on the other hand,

$$u_j^{n+1} = \hat{u} \sum_{\ell} C_{\ell} e^{i((j+\ell)k\Delta x - \omega n \Delta t)} = g^a(k) u_j^n,$$

which yields the discrete dispersion relation  $e^{-i\omega\Delta t} = g^a(k)$ , where  $g^a$  is the (scalar) amplification factor (1.19). The real part of  $e^{-i\omega\Delta t}$  (which yields a diffusion or dissipation error) is the modulus of the amplification factor  $g^a$  and has already been considered above. The study of the imaginary part, i.e., the argument of  $g^a$ , reveals the existence of a “phase error” (or dispersion error). For instance, the amplification factor of the Lax–Wendroff scheme is

$$e^{-i\omega\Delta t} = g^a(k) = 1 - \lambda^2 a^2 (1 - \cos(k\Delta x)) + i\lambda a \sin(k\Delta x),$$

which gives, if  $\omega = (a' + ib')k$ ,  $\tan a' k \Delta t$  and after some Taylor expansions (for small  $\Delta t = \lambda \Delta x$ )

$$a' = a - ck^2 = a - \frac{1}{6} a (k \Delta x)^2 (1 - \lambda^2 a^2) < a$$

(within  $\Delta x^3$ ). The discrete Fourier modes present a lagging phase error when  $a > 0$  as noted previously (see Sod [1069], Chapter 3, Section 3.6).

## 2 Godunov's Method

The most natural finite volume method to solve the Cauchy problem (1.1) is Godunov's method ([542]; see G.R., Chapter 3, Example 2.3 [539]). Let us recall its main features.

### 2.1 Godunov's Method for Systems

With the usual assumptions of Sect. 1, we know that the Riemann problem

$$\begin{cases} \frac{\partial \mathbf{w}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{w}) = \mathbf{0}, & x \in \mathbb{R}, t > 0, \\ \mathbf{w}(x, 0) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0 \end{cases} \end{cases} \quad (2.1)$$

has an entropy solution

$$\mathbf{w}(x, t) = \mathbf{w}_R \left( \frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R \right) \quad (2.2)$$

that consists of at most  $(p + 1)$  constant states separated by shock waves, rarefaction waves, or contact discontinuities, at least when  $|\mathbf{u}_L - \mathbf{u}_R|$  is small enough (see Chap. II, Sect. 6). A solution of this form is unique.

Given an approximation  $\mathbf{v}^n = (\mathbf{v}_j^n)_{j \in \mathbb{Z}}$  of  $\mathbf{u}(\cdot, t_n)$  (where  $\mathbf{v}_j^n$  is now a column vector of  $\mathbb{R}^p$ ), we define the approximation  $\mathbf{v}^{n+1} = (\mathbf{v}_j^{n+1})_{j \in \mathbb{Z}}$  of  $\mathbf{u}(\cdot, t_{n+1})$  as follows:

- (i) We extend the sequence  $\mathbf{v}^n$  as a piecewise constant function (reconstruction step)  $\mathbf{v}_\Delta(\cdot, t_n)$  defined by

$$\mathbf{v}_\Delta(\cdot, t_n) = \mathbf{v}_j^n, \quad x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}}. \quad (2.3)$$

- (ii) We solve the Cauchy problem (evolution step)

$$\begin{cases} \frac{\partial \mathbf{w}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{w}) = \mathbf{0}, & x \in \mathbb{R}, t > 0, \\ \mathbf{w}(x, 0) = \mathbf{v}_\Delta(\cdot, t_n), \end{cases} \quad (2.4)$$

- (iii) We project ( $L^2$  projection) the solution  $\mathbf{w}(\cdot, \Delta t)$  onto the piecewise constant functions, i.e., we set

$$\mathbf{v}_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \mathbf{w}(x, \Delta t) dx. \quad (2.5)$$

This step may also be called an averaging step. Provided we assume the CFL condition

$$\lambda \max |a_k(\mathbf{v}_j^n)| \leq \frac{1}{2}, \quad 1 \leq k \leq p, \quad (2.6)$$

so that the waves issued from the points  $x_{j-\frac{1}{2}}$  and  $x_{j+\frac{1}{2}}$  do not interact, the solution of (2.4) is in fact obtained by solving a juxtaposition of local Riemann problems and

$$\mathbf{w}(x, t) = \mathbf{w}_R \left( \frac{x - x_{j+\frac{1}{2}}}{\Delta t}; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n \right), \quad x_j < x < x_{j+1}, \quad j \in \mathbb{Z}. \quad (2.7)$$

In order to derive a more explicit form of the scheme, let us integrate Eq. (2.4) over the rectangle  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times (0, \Delta t)$ . Since the function is piecewise smooth, we obtain

$$\begin{aligned} & \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} (\mathbf{w}(x, \Delta t) - \mathbf{w}(x, 0)) dx \\ & + \int_0^{\Delta t} (\mathbf{f}(\mathbf{w}(x_{j+\frac{1}{2}} - 0, t)) - \mathbf{f}(\mathbf{w}(x_{j-\frac{1}{2}} + 0, t))) dt = 0. \end{aligned}$$

Using (2.3) and (2.5), we get

$$\Delta x (\mathbf{v}_j^{n+1} - \mathbf{v}_j^n) + \int_0^{\Delta t} (\mathbf{f}(\mathbf{w}(x_{j+\frac{1}{2}} - 0, t)) - \mathbf{f}(\mathbf{w}(x_{j-\frac{1}{2}} + 0, t))) dt = 0.$$

Hence, we have by (2.7)

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \{ \mathbf{f}(\mathbf{w}_R(0-; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n)) - \mathbf{f}(\mathbf{w}_R(0+; \mathbf{v}_{j-1}^n, \mathbf{v}_j^n)) \}.$$

Now, the function  $\xi \mapsto \mathbf{f}(\mathbf{w}_R(\xi; \mathbf{u}_L, \mathbf{u}_R))$  is continuous at the origin because of the Rankine–Hugoniot condition, even if  $\mathbf{w}_R$  may be discontinuous at 0, so that Godunov's method can be written in the form

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \{ \mathbf{f}(\mathbf{w}_R(0\pm; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n)) - \mathbf{f}(\mathbf{w}_R(0\pm; \mathbf{v}_{j-1}^n, \mathbf{v}_j^n)) \}, \quad (2.8)$$

and its numerical flux is given by

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{w}_R(0; \mathbf{u}, \mathbf{v})). \quad (2.9)$$

Hence, in the finite volume approach, Godunov's scheme consists in approximating the interface flux by the exact flux taken on the solution of the Riemann problem between the two states on each side of the interface. We have seen in the scalar case that the above derivation remains valid with a relaxed CFL condition, which can be written here as

$$\lambda \max |a_k(\mathbf{v}_j^n)| \leq 1, \quad 1 \leq k \leq p. \quad (2.10)$$

Let us notice that since

$$\begin{cases} \mathbf{w}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) = \mathbf{u}_L, & \text{if } \frac{x}{t} \leq a_1(\mathbf{u}_L), \\ \mathbf{w}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) = \mathbf{u}_R, & \text{if } \frac{x}{t} \geq a_p(\mathbf{u}_R), \end{cases}$$

the numerical flux satisfies

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \begin{cases} \mathbf{f}(\mathbf{v}), & \text{if } a_p < 0, \\ \mathbf{f}(\mathbf{u}), & \text{if } a_1 > 0, \end{cases}$$

and thus coincides with that of the upwind scheme when all eigenvalues have the same sign.

Let us consider the *linear* case (1.14)

$$\mathbf{f}(\mathbf{u}) = \mathbf{A}\mathbf{u}$$

and  $\mathbf{A}$  is a  $p \times p$  matrix whose eigenvalues are real and distinct,

$$a_1 < a_2 < \dots < a_p.$$

As previously, we denote by  $\mathbf{r}_1, \dots, \mathbf{r}_p$  (respectively  $\mathbf{l}_1, \dots, \mathbf{l}_p$ ) corresponding (right) eigenvectors of  $\mathbf{A}$  (resp.  $\mathbf{A}^T$ ) forming a dual basis of  $\mathbb{R}^P$ . Then (see Chap. II, Sect. 1), the Cauchy problem

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, & x \in \mathbb{R}, t > 0, \\ \mathbf{u}(x, 0) = \mathbf{u}_0(x) \end{cases} \quad (2.11)$$

can be solved explicitly. Indeed, the system can be decomposed in  $p$  decoupled scalar equations involving the characteristic variables, and we get

$$\mathbf{u}(x, t) = \sum_{k=1}^p \mathbf{l}_k^T \mathbf{u}_0(x - a_k t) \mathbf{r}_k.$$

Now, we consider the Riemann problem (2.1) in this *linear* case. Setting

$$\mathbf{u}_L = \sum_{k=1}^p \alpha_{kL} \mathbf{r}_k, \quad \mathbf{u}_R = \sum_{k=1}^p \alpha_{kR} \mathbf{r}_k, \quad (2.12)$$

we obtain for  $a_m < \xi < a_{m+1}$ ,  $0 \leq m \leq p$ , with the convention  $a_0 = -\infty$ ,  $a_{p+1} = +\infty$ ,

$$\mathbf{w}_R(\xi; \mathbf{u}_L, \mathbf{u}_R) = \mathbf{w}_m = \sum_{k=1}^m \alpha_{kR} \mathbf{r}_k + \sum_{k=m+1}^p \alpha_{kL} \mathbf{r}_k. \quad (2.13)$$

Hence, we have, if the index  $m$  is such that  $a_m < 0 < a_{m+1}$ ,

$$\mathbf{A}\mathbf{w}_R(0; \mathbf{u}_L, \mathbf{u}_R) = \sum_{k=1}^m a_k \alpha_{kR} \mathbf{r}_k + \sum_{k=m+1}^p a_k \alpha_{kL} \mathbf{r}_k. \quad (2.14)$$

Clearly (2.14) remains valid when  $a_m = 0$  or  $a_{m+1} = 0$ .

Let us give a more compact expression for (2.14). As in (1.8), we introduce a  $p \times p$  nonsingular matrix  $\mathbf{T}$  such that

$$\mathbf{A} = \mathbf{T}\Lambda\mathbf{T}^{-1}, \quad \Lambda = \text{diag}(a_k).$$

We can suppose that the  $k$ th column of  $\mathbf{T}$  is indeed the right eigenvector  $\mathbf{r}_k$ , while the  $k$ th row of  $\mathbf{T}^{-1}$  is the left eigenvector  $\mathbf{l}_k^T$ . Then, we define  $\mathbf{A}^+$  and  $\mathbf{A}^-$  by (1.9),

$$\mathbf{A}^\pm = \mathbf{T}\Lambda^\pm\mathbf{T}^{-1}, \quad \Lambda^\pm = \text{diag}(a_k^\pm).$$

Now, it follows at once from (2.14) that

$$\mathbf{A}\mathbf{w}_R(0; \mathbf{u}_L, \mathbf{u}_R) = \mathbf{A}^-\mathbf{u}_R + \mathbf{A}^+\mathbf{u}_L. \quad (2.15)$$

Hence, in the linear case, Godunov's method becomes

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \{ (\mathbf{A}^-(\mathbf{v}_{j+1}^n + \mathbf{A}^+\mathbf{v}_j^n) - (\mathbf{A}^-\mathbf{v}_j^n + \mathbf{A}^+\mathbf{v}_{j-1}^n)) \}$$

and resumes to the standard upwind or C.I.R. scheme (see (1.10b), Example 1.3)

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \{ \mathbf{A}^+(\mathbf{v}_j^n - \mathbf{v}_{j-1}^n) + \mathbf{A}^-(\mathbf{v}_{j+1}^n - \mathbf{v}_j^n) \}, \quad (2.16)$$

which can also be written as in (1.10b)

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \frac{\lambda}{2} \mathbf{A}(\mathbf{v}_{j+1}^n - \mathbf{v}_{j-1}^n) + \frac{\lambda}{2} |\mathbf{A}|(\mathbf{v}_{j+1}^n - 2\mathbf{v}_j^n + \mathbf{v}_{j-1}^n).$$

The numerical flux is (see (2.9))  $\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{A}\mathbf{w}_R(0; \mathbf{u}, \mathbf{v})$  and thus (as already obtained in Example 1.3)

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \mathbf{A}(\mathbf{u} + \mathbf{v}) - \frac{1}{2} |\mathbf{A}|(\mathbf{v} - \mathbf{u}) = \mathbf{A}^+\mathbf{u} + \mathbf{A}^-\mathbf{v}. \quad (2.17)$$

In the following, we shall need these expressions related to the linear Riemann problem:

$$\int_0^{\frac{\Delta x}{2}} \mathbf{w}_R \left( \frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R \right) dx, \quad \int_{-\frac{\Delta x}{2}}^0 \mathbf{w}_R \left( \frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R \right) dx.$$

*Lemma 2.1*

When  $\mathbf{f}(\mathbf{u}) = \mathbf{A}\mathbf{u}$ , we have if  $\lambda \max_k |a_k| \leq \frac{1}{2}$ ,

$$\frac{2}{\Delta x} \int_0^{\frac{\Delta x}{2}} \mathbf{w}_R \left( \frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R \right) dx = \mathbf{u}_R - \lambda \mathbf{A}(\mathbf{u}_R - \mathbf{u}_L) - \lambda |\mathbf{A}|(\mathbf{u}_R - \mathbf{u}_L), \quad (2.18a)$$

$$\frac{2}{\Delta x} \int_{-\frac{\Delta x}{2}}^0 \mathbf{w}_R \left( \frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R \right) dx = \mathbf{u}_L - \lambda \mathbf{A}(\mathbf{u}_R - \mathbf{u}_L) + \lambda |\mathbf{A}|(\mathbf{u}_R - \mathbf{u}_L). \quad (2.18b)$$

*Proof.* These expressions have already been computed in the scalar case (see G.R., Chapter 3, (2.10) and (4.12)). They are obtained, respectively, by integrating (2.11) on the domains  $(0, \frac{\Delta x}{2}) \times (0, \Delta t)$  and  $(-\frac{\Delta x}{2}, 0) \times (0, \Delta t)$ .  $\square$

In order to apply the scheme to gas dynamics, we first recall the equations.

## 2.2 The Gas Dynamics Equations in a Moving Frame

We start from the gas dynamics equations in Eulerian coordinates (see Chap. II, Example 2.4):

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial}{\partial x} \mathbf{F}(\mathbf{U}) = \mathbf{0}, \quad (2.19a)$$

$$\mathbf{U} = \begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ (\rho e + p)u \end{pmatrix}. \quad (2.19b)$$

Let us derive the gas dynamics equations in a moving frame. Denoting by  $v = v(x, t)$  the frame velocity, we consider the differential system

$$\frac{dx}{dt} = v(x, t), \quad (2.20)$$

and, for all  $\xi \in \mathbb{R}$ , we denote by  $t \mapsto x(\xi, t)$  the solution of (2.20) that satisfies the initial condition

$$x(0) = \xi.$$

Then  $(\xi, t)$  are the coordinates associated with the velocity field  $v$ . If we set

$$J(\xi, t) = \frac{\partial x}{\partial \xi}(\xi, t), \quad (2.21)$$

we have

$$\frac{\partial J}{\partial t}(\xi, t) = \frac{\partial}{\partial \xi} \left( \frac{\partial x}{\partial t}(\xi, t) \right) = \frac{\partial}{\partial \xi} v(x, t) = \frac{\partial v}{\partial x} \frac{\partial x}{\partial \xi}(\xi, t) = J \frac{\partial v}{\partial x}.$$

Now, given a function  $\varphi = \varphi(x, t)$  expressed in Eulerian coordinates, we denote by  $\bar{\varphi} = \bar{\varphi}(\xi, t)$  this function expressed in the moving frame coordinates, i.e.,

$$\bar{\varphi}(\xi, t) = \varphi(x(\xi, t), t).$$

Then, an easy computation shows that

$$\begin{aligned}\frac{\partial}{\partial t}(\bar{\varphi}J) &= J\left(\overline{\frac{\partial \varphi}{\partial t}} + \overline{\frac{\partial}{\partial x}(\varphi v)}\right) \\ &= J\left\{\overline{\frac{\partial \varphi}{\partial t}} + \overline{\frac{\partial}{\partial x}(\varphi u)} + \overline{\frac{\partial}{\partial x}(\varphi(v-u))}\right\}.\end{aligned}\quad (2.22)$$

We obtain

$$\begin{cases} \frac{\partial}{\partial t}(\bar{\rho}J) + J\overline{\frac{\partial}{\partial x}(\rho(u-v))} = 0, \\ \frac{\partial}{\partial t}(\bar{\rho}uJ) + J\left\{\overline{\frac{\partial}{\partial x}(\rho u(u-v))} + \overline{\frac{\partial p}{\partial x}}\right\} = 0, \\ \frac{\partial}{\partial t}(\bar{\rho}eJ) + J\left\{\overline{\frac{\partial}{\partial x}(\rho e(u-v))} + \overline{\frac{\partial pu}{\partial x}}\right\} = 0. \end{cases} \quad (2.23a)$$

Since

$$\frac{\partial}{\partial \xi} = \left(\frac{\partial x}{\partial \xi}\right) \frac{\partial}{\partial x} = J \frac{\partial}{\partial x},$$

(2.23a) gives by suppressing the bars for simplicity

$$\begin{cases} \frac{\partial}{\partial t}(\rho J) + \frac{\partial}{\partial \xi}(\rho(u-v)) = 0, \\ \frac{\partial}{\partial t}(\rho u J) + \frac{\partial}{\partial \xi}(\rho u(u-v)) + \frac{\partial p}{\partial \xi} = 0, \\ \frac{\partial}{\partial t}(\rho e J) + \frac{\partial}{\partial \xi}(\rho e(u-v)) + \frac{\partial}{\partial \xi}(pu) = 0. \end{cases} \quad (2.23b)$$

For  $v = 0$ , system (2.23) takes the form of the equations in Eulerian coordinates, while for  $v = u$ , we obtain the *Lagrangian* formulation (see Chap. II, Example 2.3)

$$\begin{cases} \frac{\partial}{\partial t}(\bar{\rho}J) = 0, \\ \frac{\partial}{\partial t}(\bar{\rho}uJ) + J\overline{\frac{\partial p}{\partial x}} = 0, \\ \frac{\partial}{\partial t}(\bar{\rho}eJ) + J\overline{\frac{\partial}{\partial x}(pu)} = 0. \end{cases} \quad \text{or} \quad \begin{cases} \frac{\partial}{\partial t}(\rho J) = 0, \\ \frac{\partial}{\partial t}(\rho u J) + \frac{\partial p}{\partial \xi} = 0, \\ \frac{\partial}{\partial t}(\rho e J) + \frac{\partial}{\partial \xi}(pu) = 0. \end{cases} \quad (2.24)$$

In the case of Lagrangian coordinates, we can put the equations in a more classical form (see Chap. II, Example 2.3). If we introduce the specific volume  $\tau$  and a mass variable  $m$  such that

$$dm = \rho_0 d\xi,$$

then the equations in Lagrangian coordinates can be written as

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial}{\partial m}(pu) = 0, \end{cases} \quad (2.25)$$

with  $p = p(\tau, \varepsilon) = p(\tau, e - \frac{u^2}{2})$ . By setting

$$\mathbf{V} = \begin{pmatrix} \tau \\ u \\ e \end{pmatrix}, \quad \mathbf{F}(\mathbf{V}) = \begin{pmatrix} -u \\ p \\ pu \end{pmatrix}, \quad (2.26)$$

$\Omega = \{\mathbf{V}; \tau > 0, u \in \mathbb{R}, e - \frac{u^2}{2} > 0\}$ , we can write the equations in the general form (1.1)

$$\frac{\partial \mathbf{V}}{\partial t} + \frac{\partial}{\partial m} \mathbf{F}(\mathbf{V}) = \mathbf{0}.$$

### 2.3 Godunov's Method in Lagrangian Coordinates

The gas is divided into slabs  $(\xi_{j-\frac{1}{2}}, \xi_{j+\frac{1}{2}})$  with thickness  $\Delta\xi_j$  that is not necessarily uniform (see Fig. 2.1). As usual, the initial condition for a quantity  $v$  is given by the slab averages

$$v_j^0 = (\Delta\xi_j)^{-1} \int_{\xi_{j-1/2}}^{\xi_{j+1/2}} v(\xi, 0) d\xi.$$

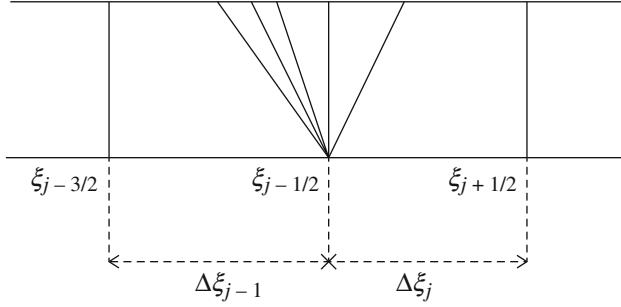
In view of (2.9), we have to compute  $\mathbf{F}(\mathbf{w}_R(0; \mathbf{U}_j^n, \mathbf{U}_{j+1}^n))$ . Remember that the eigenvalues satisfy (see Chap. II, (2.32))

$$a_1 < a_2 = 0 < a_3.$$

Let  $u_{j+\frac{1}{2}}^n$  and  $p_{j+\frac{1}{2}}^n$  be the values of  $u$  and  $p$  at the contact discontinuity between  $\mathbf{V}_j^n$  and  $\mathbf{V}_{j+1}^n$ . Indeed, the second field is linearly degenerate, so that (Chap. II, Theorem 4.2) the two Riemann invariants, and in particular  $p$  and  $u$ , are constant. Hence

$$\mathbf{F}(\mathbf{w}_R(0; \mathbf{V}_j^n, \mathbf{V}_{j+1}^n)) = (-u_{j+\frac{1}{2}}^n, p_{j+\frac{1}{2}}^n, (pu)_{j+\frac{1}{2}}^n)^T.$$

Godunov's scheme applied to the system (2.25) can be written as



**Fig. 2.1** Grid in Lagrangian coordinates

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \frac{\Delta t}{\Delta m_j} (u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n), \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta m_j} (p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n), \\ e_j^{n+1} = e_j^n - \frac{\Delta t}{\Delta m_j} ((pu)_{j+\frac{1}{2}}^n - (pu)_{j-\frac{1}{2}}^n), \end{cases} \quad (2.27)$$

where the mass variable increment is defined by

$$\Delta m_j = \rho_j^0 \Delta \xi_j \quad (2.28)$$

and

$$p_j^n = p(\tau_j^n, \varepsilon_j^n), \quad \varepsilon_j^n = e_j^n - \frac{(u_j^n)^2}{2}.$$

Let us derive another equivalent form for Godunov's scheme. We have not yet specified the movement of the grid. We use an approximation of  $u = \frac{dx}{dt}$ , and we set

$$x_{j+\frac{1}{2}}^0 = \xi_{j+\frac{1}{2}}. \quad (2.29)$$

Then  $x_{j+\frac{1}{2}}^n$ , which is the Eulerian coordinate of the interface  $\xi_{j+\frac{1}{2}}$  at time  $t_n$ , is updated according to

$$x_{j+\frac{1}{2}}^{n+1} = x_{j+\frac{1}{2}}^n + \Delta t u_{j+\frac{1}{2}}^n. \quad (2.30)$$

We can check by induction that

$$\rho_j^n (x_{j+\frac{1}{2}}^n - x_{j-\frac{1}{2}}^n) = \Delta m_j. \quad (2.31)$$

First, this is true for  $n = 0$  by assumption. Suppose that it holds for some  $n$ ; we have

$$\begin{aligned} \Delta m_j \tau_j^{n+1} &= \Delta m_j \tau_j^n + \Delta t (u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n) \\ &= (x_{j+\frac{1}{2}}^n - x_{j-\frac{1}{2}}^n) + \Delta t (u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n) = x_{j+\frac{1}{2}}^{n+1} - x_{j-\frac{1}{2}}^{n+1}. \end{aligned}$$

Hence

$$\rho_j^{n+1}(x_{j+\frac{1}{2}}^{n+1} - x_{j-\frac{1}{2}}^{n+1}) = \Delta m_j,$$

which proves the desired result.

Godunov's method in Lagrangian coordinates can thus be written as

$$\begin{cases} \Delta m_j = \rho_j^0(x_{j+\frac{1}{2}}^0 - x_{j-\frac{1}{2}}^0), \\ x_{j+\frac{1}{2}}^{n+1} = x_{j+\frac{1}{2}}^n + \Delta t u_{j+\frac{1}{2}}^n \end{cases} \quad (2.32a)$$

$$\begin{cases} \rho_j^{n+1} = (x_{j+\frac{1}{2}}^{n+1} - x_{j-\frac{1}{2}}^{n+1})^{-1} \Delta m_j, \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta m_j} (p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n), \\ e_j^{n+1} = e_j^n - \frac{\Delta t}{\Delta m_j} ((pu)_j^n - (pu)_{j-\frac{1}{2}}^n). \end{cases} \quad (2.32b)$$

*Remark 2.1.* Godunov's method in Lagrangian coordinates can be interpreted as a finite volume method. Indeed, Eqs. (2.32b) can be written as

$$\frac{\partial}{\partial t}(\varphi J) + \frac{\partial \mathbf{f}}{\partial \xi} = 0.$$

Let us integrate these equations on  $(\xi_{j-\frac{1}{2}}, \xi_{j+\frac{1}{2}})$ ; we get

$$\frac{d}{dt} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \varphi J d\xi + \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \frac{\partial \mathbf{f}}{\partial \xi} d\xi = 0$$

or

$$\frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \varphi d\xi + (\mathbf{f}_{j+\frac{1}{2}} - \mathbf{f}_{j-\frac{1}{2}}) = 0,$$

where the dependence of  $\mathbf{f}, \varphi$ , and  $x$  on  $t$  is omitted. Assuming that  $\varphi$  is constant in each cell  $(x_{j-\frac{1}{2}}^n, x_{j+\frac{1}{2}}^n)$ , we integrate by the Euler explicit method

$$\Delta x_j^{n+1} \varphi_j^{n+1} = \Delta x_j^n \varphi_j^n - \Delta t (\mathbf{f}_{j+\frac{1}{2}}^n - \mathbf{f}_{j-\frac{1}{2}}^n).$$

This gives (in the case  $v = u$ ), if  $(\rho, u, e)$  are constant in each cell,

$$\begin{cases} \Delta x_j^{n+1} \rho_j^{n+1} = \Delta x_j^n \rho_j^n = \dots = \Delta m_j, \\ \Delta x_j^{n+1} (\rho u)_j^{n+1} = \Delta x_j^n (\rho u)_j^n - \Delta t \{ p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n \}, \\ \Delta x_j^{n+1} (\rho e)_j^{n+1} = \Delta x_j^n (\rho e)_j^n - \Delta t \{ (pu)_{j+\frac{1}{2}}^n - (pu)_{j-\frac{1}{2}}^n \}, \end{cases}$$

and thus

$$\begin{cases} \Delta x_j^n \rho_j^n = \Delta m_j, \\ \Delta m_j^{n+1} u_j^{n+1} = \Delta m_j^n u_j^n - \Delta t \{ p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n \}, \\ \Delta m_j^{n+1} e_j^{n+1} = \Delta m_j^n e_j^n - \Delta t \{ (\rho u)_{j+\frac{1}{2}}^n - (\rho u)_{j-\frac{1}{2}}^n \}. \end{cases}$$

This is Godunov's scheme provided  $(u_{j+\frac{1}{2}}^n, p_{j+\frac{1}{2}}^n)$  are defined by the solution of the Riemann problem. This finite volume method can be easily extended to multidimensional problems and is widely used in practice (see Chap. V, Sect. 4).  $\square$

## 2.4 Godunov's Method in Eulerian Coordinates (Direct Method)

We consider now the system (2.19), with

$$\mathbf{U} = (\rho, \rho u, \rho e)^T, \quad \mathbf{F}(\mathbf{U}) = (\rho u, \rho u^2 + p, (\rho e + p)u)^T.$$

Let us recall that the eigenvalues of  $\mathbf{F}'(\mathbf{U})$  are  $a_1 = u - c < a_2 = u < a_3 = u + c$  (see Chap. II, Example 2.4). One has to solve the Riemann problem at the point  $x_{j+\frac{1}{2}}$ , between the states  $\mathbf{U}_j^n$  and  $\mathbf{U}_{j+1}^n$ . Godunov's scheme can be written as

$$\begin{cases} \rho_j^{n+1} = \rho_j^n - \frac{\Delta t}{\Delta x_j} \{ (\rho u)_{j+\frac{1}{2}}^n - (\rho u)_{j-\frac{1}{2}}^n \}, \\ (\rho u)_j^{n+1} = (\rho u)_j^n - \frac{\Delta t}{\Delta x_j} \{ (\rho u^2 + p)_{j+\frac{1}{2}}^n - (\rho u^2 + p)_{j-\frac{1}{2}}^n \}, \\ (\rho e)_j^{n+1} = (\rho e)_j^n - \frac{\Delta t}{\Delta x_j} \{ ((\rho e + p)u)_{j+\frac{1}{2}}^n - ((\rho e + p)u)_{j-\frac{1}{2}}^n \}. \end{cases} \quad (2.33)$$

The point  $x_{j+\frac{1}{2}}$  no longer corresponds to the contact discontinuity, and we must test the position of the line  $x = x_{j+\frac{1}{2}}$  with respect to the different waves. For instance, in the case of Fig. 2.2 (see also Chap. III, Figs. 3.1 and 3.10), the line  $x = x_{j+\frac{1}{2}}$  lies between the 1-rarefaction fan and the contact discontinuity, so that

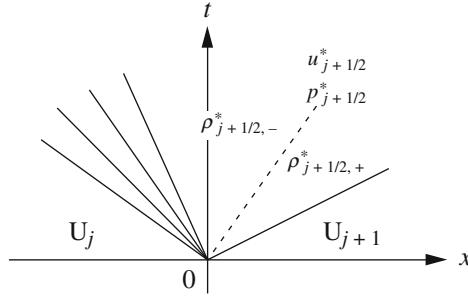
$$(u_{j+\frac{1}{2}}^n, p_{j+\frac{1}{2}}^n) = (u_{j+\frac{1}{2}}^*, p_{j+\frac{1}{2}}^*),$$

$$\rho_{j+\frac{1}{2}}^n = \rho_{j+\frac{1}{2},-}^*,$$

and

$$e_{j+\frac{1}{2}}^n = e_{j+\frac{1}{2},-}^*,$$

with obvious notations (which are slightly different from those of Chap. III).



**Fig. 2.2** Example of Riemann problem

## 2.5 Godunov's Method in Eulerian Coordinates (Lagrangian Step + Projection)

As we have just seen, Godunov's method in a Lagrangian grid is easier to handle. However, many physical models use the Eulerian coordinates, and this justifies the present method, which couples a Lagrangian step followed by a projection back onto the Eulerian grid, which we suppose is fixed, with nodes  $x_{j+\frac{1}{2}}$ .

### (i) The Lagrangian step

At the beginning of each Lagrangian step, the Lagrangian and Eulerian grids coincide, and  $\xi_{j+\frac{1}{2}} = x_{j+\frac{1}{2}}$ . The piecewise constant approximate values of  $(\rho, u, p)$  on the Lagrangian grid at time  $t_n$  are thus  $(\rho_j^n, u_j^n, e_j^n)$ . We compute  $u_{j+\frac{1}{2}}^n, p_{j+\frac{1}{2}}^n$  values of  $u$  and  $p$  on the contact discontinuity  $x = \xi_{j+\frac{1}{2}}$ . Following (2.30), we define the updated Eulerian coordinates of the Lagrangian zone (Fig. 2.3) by

$$x_{j+\frac{1}{2}}^* = x_{j+\frac{1}{2}} + \Delta t u_{j+\frac{1}{2}}^n, \quad (2.34)$$

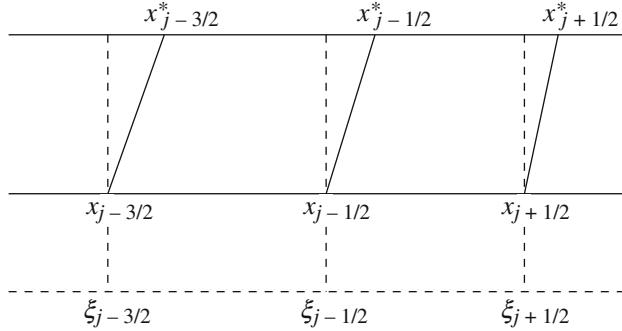
where the dependence of  $x_{j+\frac{1}{2}}^*$  on  $n$  is omitted. In other words, if  $x_{j+\frac{1}{2}} = x(\xi_{j+\frac{1}{2}}, t_n)$ , then  $x_{j+\frac{1}{2}}^* \approx x(\xi_{j+\frac{1}{2}}, t_{n+1})$ .

Now, following (2.32), we compute  $\bar{\rho}_j^{n+1}, \bar{u}_j^{n+1}, \bar{e}_j^{n+1}$ , which are the constant values of  $\bar{\rho}^{n+1}, \bar{u}^{n+1}, \bar{e}^{n+1}$  in each cell  $(x_{j-\frac{1}{2}}^*, x_{j+\frac{1}{2}}^*)$ :

$$\begin{cases} \bar{\rho}_j^{n+1} = (x_{j+\frac{1}{2}}^* - x_{j-\frac{1}{2}}^*)^{-1} \Delta m_j^n, & \text{where } \Delta m_j^n = \rho_j^n \Delta x_j, \\ \bar{u}_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta m_j^n} \{p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n\}, \\ \bar{e}_j^{n+1} = e_j^n - \frac{\Delta t}{\Delta m_j^n} \{(pu)_j^n - (pu)_{j-\frac{1}{2}}^n\}, \end{cases} \quad (2.35)$$

### (ii) The projection step

The values  $\bar{\rho}^{n+1}, \bar{u}^{n+1}, \bar{e}^{n+1}$  are projected back onto the Eulerian grid; denoting any of these quantities by  $v$ , we thus define



**Fig. 2.3** Updating of the Lagrangian grid

$$v_j^{n+1} = (\Delta x_j)^{-1} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \bar{v}^{n+1}(x) dx. \quad (2.36)$$

We break up the integral into three parts and compute separately the corresponding integrals

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \bar{v}^{n+1}(x) dx = \int_{x_{j-\frac{1}{2}}}^{x_{j-\frac{1}{2}}^*} + \int_{x_{j-\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}^*} + \int_{x_{j+\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}}$$

On the one hand, we have

$$\int_{x_{j-\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}} \bar{v}^{n+1}(x) dx = \Delta x_j^* \bar{v}_j^{n+1} \quad (2.37)$$

and, on the other hand,

$$\int_{x_{j+\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}} \bar{v}^{n+1}(x) dx = (x_{j+\frac{1}{2}} - x_{j+\frac{1}{2}}^*) \begin{cases} \bar{v}_{j+1}^{n+1}, & \text{if } u_{j+\frac{1}{2}}^n < 0, \\ \bar{v}_j^{n+1}, & \text{if } u_{j+\frac{1}{2}}^n \geq 0, \end{cases}$$

which we can write

$$\int_{x_{j+\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}} \bar{v}^{n+1}(x) dx = -\Delta t u_{j+\frac{1}{2}}^n v_{j+\frac{1}{2}}^{n+1} + \varepsilon(j,n), \quad (2.38)$$

where

$$\varepsilon(j,n) = \begin{cases} -\frac{1}{2}, & \text{if } u_{j+\frac{1}{2}}^n \geq 0, \\ \frac{1}{2}, & \text{if } u_{j+\frac{1}{2}}^n < 0. \end{cases}$$

In the same way, we have

$$\int_{x_{j-\frac{1}{2}}}^{x_{j-\frac{1}{2}}^*} \bar{v}^{n+1}(x) dx = (x_{j-\frac{1}{2}}^* - x_{j-\frac{1}{2}}) \begin{cases} \bar{v}_j^{n+1}, & \text{if } u_{j-\frac{1}{2}}^n < 0, \\ \bar{v}_{j-1}^{n+1}, & \text{if } u_{j-\frac{1}{2}}^n \geq 0, \end{cases}$$

so that

$$\int_{x_{j-\frac{1}{2}}}^{x_{j-\frac{1}{2}}^*} \bar{v}^{n+1}(x) dx = \Delta t u_{j-\frac{1}{2}}^n \bar{v}_{j-\frac{1}{2}}^{n+1} + \varepsilon(j-1, n).$$

We get

$$\begin{aligned} \Delta x_j v_j^{n+1} &= \Delta x_j^* \bar{v}_j^{n+1} \\ &\quad - \Delta t (u_{j+\frac{1}{2}}^n \bar{v}_{j+\frac{1}{2}+\varepsilon(j, n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{v}_{j-\frac{1}{2}+\varepsilon(j-1, n)}^{n+1}). \end{aligned} \quad (2.39)$$

In particular, we obtain

$$\begin{aligned} \Delta m_j^{n+1} &= \Delta x_j \rho_j^{n+1} = \Delta x_j^* \bar{\rho}_j^{n+1} \\ &\quad - \Delta t (u_{j+\frac{1}{2}}^n \bar{\rho}_{j+\frac{1}{2}+\varepsilon(j, n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho}_{j-\frac{1}{2}+\varepsilon(j-1, n)}^{n+1}), \end{aligned}$$

or

$$\Delta m_j^{n+1} = \Delta m_j^n - \Delta t (u_{j+\frac{1}{2}}^n \bar{\rho}_{j+\frac{1}{2}+\varepsilon(j, n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho}_{j-\frac{1}{2}+\varepsilon(j-1, n)}^{n+1}). \quad (2.40)$$

Also, we have

$$\begin{aligned} \Delta x_j (\rho u)_j^{n+1} &= \Delta x_j^* \bar{\rho} \bar{u}_j^{n+1} \\ &\quad - \Delta t (u_{j+\frac{1}{2}}^n \bar{\rho} \bar{u}_{j+\frac{1}{2}+\varepsilon(j, n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho} \bar{u}_{j-\frac{1}{2}+\varepsilon(j-1, n)}^{n+1}), \end{aligned}$$

and hence

$$\begin{aligned} \Delta m_j^{n+1} u_j^{n+1} &= \Delta m_j^n \bar{u}_j^{n+1} \\ &\quad - \Delta t (u_{j+\frac{1}{2}}^n \bar{\rho} \bar{u}_{j+\frac{1}{2}+\varepsilon(j, n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho} \bar{u}_{j-\frac{1}{2}+\varepsilon(j-1, n)}^{n+1}), \end{aligned}$$

and by (2.35)

$$\left\{ \begin{array}{l} \Delta m_j^{n+1} u_j^{n+1} = \Delta m_j^n u_j^n - \Delta t (p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n) \\ \quad - \Delta t \{ u_{j+\frac{1}{2}}^n \bar{\rho} \bar{u}_{j+\frac{1}{2}+\varepsilon(j, n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho} \bar{u}_{j-\frac{1}{2}+\varepsilon(j-1, n)}^{n+1} \}. \end{array} \right. \quad (2.41)$$

We have similarly

$$\left\{ \begin{array}{l} \Delta m_j^{n+1} e_j^{n+1} = \Delta m_j^n e_j^n - \Delta t \{ (pu)_{j+\frac{1}{2}}^n - (pu)_{j-\frac{1}{2}}^n \} \\ \quad - \Delta t \{ u_{j+\frac{1}{2}}^n \bar{\rho} e_{j+\frac{1}{2}+\varepsilon(j, n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho} e_{j-\frac{1}{2}+\varepsilon(j-1, n)}^{n+1} \}. \end{array} \right. \quad (2.42)$$

The scheme is thus defined by

$$\left\{ \begin{array}{l} \rho_j^{n+1} = \rho_j^n - \frac{\Delta t}{\Delta x_j} \{ u_{j+\frac{1}{2}}^n \bar{\rho}_{j+\frac{1}{2}+\varepsilon(j,n)}^{n+1} \\ \quad - u_{j-\frac{1}{2}}^n \bar{\rho}_{j-\frac{1}{2}+\varepsilon(j-1,n)}^{n+1} \}, \\ \rho_j^{n+1} u_j^{n+1} = \rho_j^n u_j^n - \frac{\Delta t}{\Delta x_j} (p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n) \\ \quad - \frac{\Delta t}{\Delta x_j} \{ u_{j+\frac{1}{2}}^n \bar{\rho} u_{j-\frac{1}{2}+\varepsilon(j,n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho} u_{j-\frac{1}{2}+\varepsilon(j-1,n)}^{n+1} \}, \\ \rho_j^{n+1} e_j^{n+1} = \rho_j^{n+1} e_j^n - \frac{\Delta t}{\Delta x_j} \{ (pu)_{j+\frac{1}{2}}^n - (pu)_{j-\frac{1}{2}}^n \} \\ \quad - \frac{\Delta t}{\Delta x_j} \{ u_{j+\frac{1}{2}}^n \bar{\rho} e_{j+\frac{1}{2}+\varepsilon(j,n)}^{n+1} - u_{j-\frac{1}{2}}^n \bar{\rho} e_{j-\frac{1}{2}+\varepsilon(j-1,n)}^{n+1} \}, \end{array} \right. \quad (2.43)$$

together with (2.35). The main difference between this scheme (Lagrangian + remapping) and the direct Eulerian scheme lies in the way the convection terms are handled. It is simpler to implement but introduces some extra dissipation.

## 2.6 Godunov's Method in a Moving Grid

Starting now from equations (2.23), we use the finite-volume formulation. Solving the Riemann problem for Godunov's method amounts to computing the solution on the line  $\frac{x}{t} = v_{j+\frac{1}{2}}^n$ . This gives  $u_{j+\frac{1}{2}}^n, p_{j+\frac{1}{2}}^n, \rho_{j+\frac{1}{2}}^n, e_{j+\frac{1}{2}}^n$ . As in Remark 2.1, we have to discretize an equation of the form

$$\frac{\partial}{\partial t}(\varphi J) + \frac{\partial \mathbf{f}}{\partial \xi} = \mathbf{0},$$

which by integration yields

$$\Delta x_j^{n+1} \varphi_j^{n+1} = \Delta x_j^n \varphi_j^n - \Delta t (\mathbf{f}_{j+\frac{1}{2}}^n - \mathbf{f}_{j-\frac{1}{2}}^n).$$

In the present case, we obtain

$$\left\{ \begin{array}{l} \Delta x_j^{n+1} \rho_j^{n+1} = \Delta x_j^n \rho_j^n - \Delta t \{ (\rho(u-v))_{j+\frac{1}{2}}^n - (\rho(u-v))_{j-\frac{1}{2}}^n \}, \\ \Delta x_j^{n+1} (\rho u)_j^{n+1} = \Delta x_j^n (\rho u)_j^n - \Delta t \{ p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n \} \\ \quad - \Delta t \{ (\rho u(u-v))_{j+\frac{1}{2}}^n - (\rho u(u-v))_{j-\frac{1}{2}}^n \}, \\ \Delta x_j^{n+1} (\rho e)_j^{n+1} = \Delta x_j^n (\rho e)_j^n - \Delta t \{ (pu)_{j+\frac{1}{2}}^n - (pu)_{j-\frac{1}{2}}^n \} \\ \quad - \Delta t \{ (\rho e(u-v))_{j+\frac{1}{2}}^n - (\rho e(u-v))_{j-\frac{1}{2}}^n \}, \end{array} \right.$$

which gives

$$\left\{ \begin{array}{l} \Delta m_j^{n+1} = \Delta m_j^n - \Delta t \{ (\rho(u-v))_{j+\frac{1}{2}}^n - (\rho(u-v))_{j-\frac{1}{2}}^n \}, \\ \Delta m_j^{n+1} u_j^{n+1} = \Delta m_j^n u_j^n - \Delta t \{ p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n \} \\ \quad - \Delta t \{ (\rho u(u-v))_{j+\frac{1}{2}}^n - (\rho u(u-v))_{j-\frac{1}{2}}^n \}, \\ \Delta m_j^{n+1} e_j^{n+1} = \Delta m_j^n e_j^n - \Delta t \{ (pu)_{j+\frac{1}{2}}^n - (pu)_{j-\frac{1}{2}}^n \} \\ \quad - \Delta t \{ (\rho e(u-v))_{j+\frac{1}{2}}^n - (\rho e(u-v))_{j-\frac{1}{2}}^n \}, \end{array} \right. \quad (2.44)$$

with

$$\Delta m_j^n = \rho_j^n \Delta x_j, \quad x_{j+\frac{1}{2}}^{n+1} = x_{j+\frac{1}{2}}^n + \Delta t v_{j+\frac{1}{2}}^n.$$

For  $v = u$ , scheme (2.44) coincides with the Lagrangian scheme (2.32), while for  $v = 0$ , we find the direct Eulerian scheme (2.34). The strategy for choosing  $v_{j+\frac{1}{2}}^n$  is a crucial step in the method.

*Remark 2.2.* We might also consider a scheme combining a Lagrangian step and a projection on a moving grid. The details are left to the reader.  $\square$

In Godunov's method, we need to solve one Riemann problem per mesh point  $x_{j+\frac{1}{2}}$  at each time step, and this must be done iteratively (for the numerical techniques to obtain this solution, see, for instance, Colella and Glaz [329] and Loh and Hui [830]). This may be considered as expensive when the problem to solve does not present strong shock discontinuities; see also [1006]. In the following, we aim at developing approximate Riemann solvers that are simpler to implement and also cheaper to use.

### 3 Godunov-Type Methods

#### 3.1 Approximate Riemann Solvers and Godunov-Type Methods

We have already noticed that Godunov's method requires the exact solution of the Riemann problem at each point  $x_{j+1/2}$ , which may be a fairly complicated and computationally expensive procedure. Moreover, the projection step in Godunov's method is an averaging procedure that does not make use of all the information contained in the exact solution of the Riemann problem and limits its order of accuracy. Hence it makes sense to study numerical methods which are simpler to construct but retain the essential features of Godunov's method. In *Godunov-type methods*, the exact Riemann solver  $\mathbf{w}_R$  is replaced by an approximate one  $\tilde{\mathbf{w}}_R$ .

### 3.1.1 Consistency Condition for an Approximate Riemann Solver

Let  $\tilde{\mathbf{w}}_R(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$  be an approximation of the entropy solution of the Riemann problem (2.1). We require this approximate Riemann solver to possess some consistency properties. On the one hand, we require that

$$\tilde{\mathbf{w}}_R(\frac{x}{t}; \mathbf{u}, \mathbf{u}) = \mathbf{u}. \quad (3.1)$$

On the other hand, integrating the equation of (2.1) over the rectangle  $(-\frac{\Delta x}{2}, +\frac{\Delta x}{2}) \times (0, \Delta t)$  and assuming the condition

$$\lambda \max(|a_k(\mathbf{u}_L)|, |a_k(\mathbf{u}_R)|) \leq \frac{1}{2}, \quad 1 \leq k \leq p, \quad (3.2)$$

we obtain

$$\frac{1}{\Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \mathbf{w}_R(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R) dx = \frac{1}{2}(\mathbf{u}_L + \mathbf{u}_R) - \lambda(\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)).$$

Thus we require the following property called *consistency with the integral form of the system of conservation laws*: for  $\lambda$  “small enough” (namely, we assume that some constraint of the form (3.2) holds), we have

$$\frac{1}{\Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}_R(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R) dx = \frac{1}{2}(\mathbf{u}_L + \mathbf{u}_R) - \lambda(\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)). \quad (3.3)$$

In the same way, by integrating the entropy inequality associated with an entropy pair  $(U, F)$  over the rectangle  $(-\frac{\Delta x}{2}, +\frac{\Delta x}{2}) \times (0, \Delta t)$ , we obtain

$$\frac{1}{\Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} U\left(\mathbf{w}_R(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R)\right) dx \leq \frac{1}{2}(U(\mathbf{u}_L) + U(\mathbf{u}_R)) - \lambda(\mathbf{F}(\mathbf{u}_R) - \mathbf{F}(\mathbf{u}_L)).$$

Accordingly, it is highly desirable that the following property of *consistency with the integral form of the entropy inequality*

$$\frac{1}{\Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} U\left(\tilde{\mathbf{w}}_R(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R)\right) dx \leq \frac{1}{2}(U(\mathbf{u}_L) + U(\mathbf{u}_R)) - \lambda(\mathbf{F}(\mathbf{u}_R) - \mathbf{F}(\mathbf{u}_L)). \quad (3.4)$$

also holds for  $\lambda$  “small enough.”

Now, using the approximate Riemann solver  $\tilde{\mathbf{w}}_R(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$ , we can approximate the solution of the Cauchy problem (2.4) by the function

$$\tilde{\mathbf{w}}(x, t) = \tilde{\mathbf{w}}_R\left(\frac{x - x_{j+1/2}}{t}; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n\right), \quad x_j < x < x_{j+1}, \quad j \in \mathbb{Z}. \quad (3.5)$$

The corresponding numerical scheme is then defined by

$$\mathbf{v}_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \tilde{\mathbf{w}}(x, \Delta t) dx$$

or equivalently by

$$\mathbf{v}_j^{n+1} = \frac{1}{\Delta x} \left\{ \int_0^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{v}_{j-1}^n, \mathbf{v}_j^n\right) dx + \int_{-\frac{\Delta x}{2}}^0 \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n\right) dx \right\}. \quad (3.6)$$

Let us state

*Theorem 3.1*

Let  $\tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right)$  be an approximate Riemann solver which satisfies the conditions (3.1) and (3.3). Then the difference scheme (3.6) can be put in conservation form and is consistent with the system of conservation laws (1.1). If in addition the condition (3.4) holds for an entropy pair  $(U, F)$ , the difference scheme is consistent with the associated entropy condition.

*Proof.* In order to define a numerical flux function, we introduce the expressions

$$\mathbf{g}_+(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) + \frac{1}{\lambda \Delta x} \int_0^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right) dx - \frac{\mathbf{v}}{2\lambda} \quad (3.7)$$

and

$$\mathbf{g}_-(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{u}) - \frac{1}{\lambda \Delta x} \int_{-\frac{\Delta x}{2}}^0 \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right) dx - \frac{\mathbf{u}}{2\lambda}. \quad (3.8)$$

Using (3.3), we have

$$\begin{aligned} \mathbf{g}_+(\mathbf{u}, \mathbf{v}) - \mathbf{g}_-(\mathbf{u}, \mathbf{v}) &= \\ &= \frac{1}{\lambda \Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right) dx - \frac{1}{2\lambda}(\mathbf{u} + \mathbf{v}) + \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) = \mathbf{0}. \end{aligned}$$

Moreover (3.1) yields

$$\mathbf{g}_+(\mathbf{u}, \mathbf{u}) = \mathbf{g}_-(\mathbf{u}, \mathbf{u}) = \mathbf{f}(\mathbf{u}).$$

Hence, setting

$$\mathbf{g} = \mathbf{g}_+ = \mathbf{g}_-, \quad (3.9)$$

we have by (3.6), (3.7), and (3.8)

$$\mathbf{v}_j^{n+1} = \lambda(\mathbf{g}(\mathbf{v}_{j-1}^n, \mathbf{v}_j^n) - \mathbf{f}(\mathbf{v}_j^n)) + \frac{\mathbf{v}_j^n}{2} - \lambda(\mathbf{g}(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n) - \mathbf{f}(\mathbf{v}_j^n)) + \frac{\mathbf{v}_j^n}{2}$$

so that the difference scheme (3.6) can be written in the conservation form

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda(\mathbf{g}(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n) - \mathbf{g}(\mathbf{v}_{j-1}^n, \mathbf{v}_j^n))$$

with a consistent numerical flux  $\mathbf{g}(\mathbf{u}, \mathbf{v})$ .

For proving the entropy consistency property, we proceed in a similar way. Define

$$G_+(\mathbf{u}, \mathbf{v}) = F(\mathbf{v}) + \frac{1}{\lambda \Delta x} \int_0^{\frac{\Delta x}{2}} U\left(\tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right)\right) dx - \frac{U(\mathbf{v})}{2\lambda}$$

and

$$G_-(\mathbf{u}, \mathbf{v}) = F(\mathbf{u}) - \frac{1}{\lambda \Delta x} \int_{-\frac{\Delta x}{2}}^0 U\left(\tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right)\right) dx - \frac{U(\mathbf{u})}{2\lambda}.$$

Using again (3.1), we have

$$G_+(\mathbf{u}, \mathbf{u}) = G_-(\mathbf{u}, \mathbf{u}) = F(\mathbf{u}).$$

On the other hand, we can only conclude from (3.4) that

$$G_+(\mathbf{u}, \mathbf{v}) \leq G_-(\mathbf{u}, \mathbf{v}).$$

Since  $U$  is a convex function, we deduce from (3.6) and Jensen's inequality

$$\begin{aligned} U(\mathbf{v}_j^{n+1}) &\leq \frac{1}{2} \left\{ U\left(\frac{2}{\Delta x} \int_0^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{v}_{j-1}^n, \mathbf{v}_j^n\right) dx\right) + \right. \\ &\quad \left. + U\left(\frac{2}{\Delta x} \int_{-\frac{\Delta x}{2}}^0 \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n\right) dx\right) \right\} \leq \\ &\leq \frac{1}{2} \left\{ \frac{2}{\Delta x} \int_0^{\frac{\Delta x}{2}} U\left(\tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{v}_{j-1}^n, \mathbf{v}_j^n\right)\right) dx + \right. \\ &\quad \left. + \frac{2}{\Delta x} \int_{-\frac{\Delta x}{2}}^0 U\left(\tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n\right)\right) dx \right\} \end{aligned}$$

and by the expressions of  $G_\pm$

$$\begin{aligned} U(\mathbf{v}_j^{n+1}) &\leq U(\mathbf{v}_j^n) - \lambda \{F(\mathbf{v}_j^n) - G_+(\mathbf{v}_{j-1}^n, \mathbf{v}_j^n)\} + \\ &\quad + \lambda \{F(\mathbf{v}_j^n) - G_-(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n)\} = \\ &= U(\mathbf{v}_j^n) - \lambda \{G_-(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n) - G_+(\mathbf{v}_{j-1}^n, \mathbf{v}_j^n)\}. \end{aligned}$$

By taking the numerical entropy flux either as  $G_+(\mathbf{u}, \mathbf{v})$  or as  $G_-(\mathbf{u}, \mathbf{v})$ , this proves the entropy consistency of the difference scheme.  $\square$

### Definition 3.1

The numerical scheme (3.6) is called a Godunov-type scheme if the approximate Riemann solver  $\tilde{\mathbf{w}}_R(\frac{x}{t}; \mathbf{u}, \mathbf{v})$  satisfies the properties (3.1) and (3.3).

Given an entropy pair  $(U, F)$ , the Godunov-type method is called entropy satisfying if the approximate Riemann solver satisfies in addition the property (3.4).

It remains to construct approximate Riemann solvers which satisfy the properties (3.1), (3.3), and (3.4). Let us then introduce a class of Riemann solvers that are widely used in practice: *the simple approximate Riemann solvers*.

### 3.1.2 Simple Approximate Riemann Solvers

A *simple* solver is a self-similar Riemann solver which consists of  $m + 1$  constant states  $\mathbf{w}_l$ ,  $1 \leq l \leq m + 1$ , separated by  $m$  waves propagating at speed  $a_l$ ,  $x = a_l t$ ,  $1 \leq l \leq m$ , i.e.,

$$\tilde{\mathbf{w}}_R\left(\frac{x}{t}; \mathbf{u}, \mathbf{v}\right) = \begin{cases} \mathbf{w}_1 = \mathbf{u}, & x < a_1 t, \\ \mathbf{w}_2, & a_1 t < x < a_2 t, \\ \dots \\ \mathbf{w}_m, & a_{m-1} t < x < a_m t, \\ \mathbf{w}_{m+1} = \mathbf{v}, & x > a_m t, \end{cases} \quad (3.10)$$

where the wave speeds  $a_l$  may depend on  $(\mathbf{u}, \mathbf{v})$ . Let us now give necessary and sufficient conditions for the simple solver to satisfy the properties (3.3) and (3.4) under the CFL condition

$$\lambda \max_{1 \leq l \leq m} |a_l| \leq \frac{1}{2}. \quad (3.11)$$

*Theorem 3.2*

Assume that the simple approximate Riemann solver (3.10) satisfies the properties (3.1) and (3.11). Then it induces a Godunov-type scheme if and only if

$$\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) = \sum_{l=1}^m a_l (\mathbf{w}_{l+1} - \mathbf{w}_l). \quad (3.12)$$

Its numerical flux is given by

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \left( \mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v}) - \sum_{l=1}^m |a_l| (\mathbf{w}_{l+1} - \mathbf{w}_l) \right). \quad (3.13)$$

Given an entropy pair  $(U, F)$ , the associated Godunov-type scheme is entropy satisfying if and only if

$$F(\mathbf{v}) - F(\mathbf{u}) \leq \sum_{l=1}^m a_l (U(\mathbf{w}_{l+1}) - U(\mathbf{w}_l)). \quad (3.14)$$

*Proof.* Under the condition (3.11), we have for any function  $\varphi = \varphi(\mathbf{u})$

$$\begin{aligned} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \varphi\left(\tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right)\right) dx &= (a_1 \Delta t + \frac{\Delta x}{2}) \varphi(\mathbf{w}_1) + \\ &+ \sum_{l=1}^m (a_{l+1} - a_l) \Delta t \varphi(\mathbf{w}_l) + (\frac{\Delta x}{2} - a_m \Delta t) \varphi(\mathbf{w}_{m+1}) \end{aligned}$$

or

$$\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \varphi\left(\tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right)\right) dx = \frac{\Delta x}{2} (\varphi(\mathbf{u}) + \varphi(\mathbf{v}) - \Delta t \sum_{l=1}^m a_l (\varphi(\mathbf{w}_{l+1}) - \varphi(\mathbf{w}_l))).$$

Hence, in the case of a simple Riemann solver, we have the properties (3.3) and (3.4), respectively, so that we may apply Theorem 3.1. It remains only to check the formula (3.13). Suppose  $a_1 < \dots < a_k \leq 0 \leq a_{k+1} < \dots < a_m$ . Then

$$\int_{-\frac{\Delta x}{2}}^0 \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right) dx = -\frac{\Delta x}{2} \mathbf{u} - \sum_{l=1}^k a_l (\mathbf{w}_{l+1} - \mathbf{w}_l)$$

and

$$\int_0^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}_R\left(\frac{x}{\Delta t}; \mathbf{u}, \mathbf{v}\right) dx = \frac{\Delta x}{2} \mathbf{v} - \sum_{l=k+1}^m a_l (\mathbf{w}_{l+1} - \mathbf{w}_l).$$

By using the expressions (3.7) and (3.8) of  $\mathbf{g}_+$  and  $\mathbf{g}_-$ , respectively, we obtain

$$\mathbf{g}_+(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - \sum_{l=k+1}^m |a_l| (\mathbf{w}_{l+1} - \mathbf{w}_l)$$

and

$$\mathbf{g}_-(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{u}) - \sum_{l=1}^k |a_l| (\mathbf{w}_{l+1} - \mathbf{w}_l).$$

Since by (3.9)

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} (\mathbf{g}_+(\mathbf{u}, \mathbf{v}) + \mathbf{g}_-(\mathbf{u}, \mathbf{v})),$$

we obtain the desired expression (3.13) of the numerical flux.  $\square$

### 3.1.3 Application to Fluid Systems

Let us now restrict ourselves to systems of conservation laws of the form

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x} (\rho u) = 0, \\ \frac{\partial}{\partial t} (\rho \Phi) + \frac{\partial}{\partial x} (\rho u \Phi + \mathbf{g}(\rho, \Phi)) = \mathbf{0} \end{cases} \quad (3.15)$$

already considered in the previous chapters (Chap. I, Sect. 2, Chap. II, Sect. 2.1) that we write

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial}{\partial x} \mathbf{F}(\mathbf{U}) = \mathbf{0} \quad (3.16)$$

with

$$\mathbf{U} = \begin{pmatrix} \rho \\ \rho \Phi \end{pmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} \rho u \\ \rho u \Phi + \mathbf{g}(\rho, \Phi) \end{pmatrix}. \quad (3.17)$$

We know that, written in Lagrangian coordinates, the system (3.15) becomes

$$\frac{\partial \mathbf{V}}{\partial t} + \frac{\partial}{\partial m} \mathbf{G}(\mathbf{V}) = \mathbf{0} \quad (3.18)$$

where

$$\mathbf{V} = \begin{pmatrix} \tau \\ \Phi \end{pmatrix}, \quad \mathbf{G}(\mathbf{V}) = \begin{pmatrix} -u \\ \mathbf{g}\left(\frac{1}{\tau}, \Phi\right) \end{pmatrix}. \quad (3.19)$$

We are given a simple approximate Riemann solver for the system (3.18)

$$\tilde{\mathbf{w}}^L\left(\frac{m}{t}; \mathbf{V}_L, \mathbf{V}_R\right) = \begin{cases} \mathbf{V}_1 = \mathbf{V}_L, & \frac{m}{t} < b_1, \\ \mathbf{V}_k, & b_{k-1} < \frac{m}{t} < b_k, \quad 2 \leq k \leq l, \\ \mathbf{V}_{l+1} = \mathbf{V}_R, & \frac{m}{t} > b_l. \end{cases} \quad (3.20)$$

We assume the following natural hypotheses concerning the intermediate states  $(\tau_k, u_k)$ ,  $2 \leq k \leq l$ :

$$\tau_k > 0, \quad (3.21)$$

$$b_k \delta \tau_k = -\delta u_k, \quad (3.22)$$

where we have noted, for all  $\{\varphi_k; 1 \leq k \leq l+1\}$ ,

$$\delta \varphi_k = \varphi_{k+1} - \varphi_k, \quad 1 \leq k \leq l,$$

which denotes the jump of a quantity  $\varphi$  across the wave with index  $k$ . Note that the conditions (3.22) are nothing but the Rankine–Hugoniot jump relations for the mass conservation equation through each wave  $m = b_k t$ .

With the Lagrangian Riemann solver (3.20), we associate a simple approximate Riemann solver for the Eulerian system (3.16), following the results in Chap. II concerning the change of frame (formula (2.22) of Theorem 2.1). Using the conditions (3.22), we define the wave speeds  $a_k$ ,  $1 \leq k \leq l$  by

$$a_k = u_k + b_k \tau_k = u_{k+1} + b_k \tau_{k+1}. \quad (3.23)$$

The Eulerian approximate Riemann solver then reads

$$\tilde{\mathbf{w}}^{\mathcal{E}}\left(\frac{x}{t}; \mathbf{U}_L, \mathbf{U}_R\right) = \begin{cases} \mathbf{U}_1 = \mathbf{U}_L = \mathbf{U}(\mathbf{V}_L), & \frac{x}{t} < a_1, \\ \mathbf{U}_k = \mathbf{U}(\mathbf{V}_k), & a_{k-1} < \frac{x}{t} < a_k, \quad 2 \leq k \leq l, \\ \mathbf{U}_{l+1} = \mathbf{U}_R = \mathbf{U}(\mathbf{V}_R), & \frac{x}{t} > a_l. \end{cases} \quad (3.24)$$

We can now state

*Theorem 3.3*

Assume the conditions (3.21) and (3.22). The Eulerian simple solver (3.24) is a Godunov-type Riemann solver (resp. an entropy satisfying Godunov-type Riemann solver) if and only if the Lagrangian simple solver (3.20) is a Godunov-type Riemann solver (resp. an entropy satisfying Godunov-type Riemann solver).

*Proof.* Suppose that the Lagrangian Riemann solver (3.20) induces a Godunov-type scheme, which means by (3.12)

$$\Delta \mathbf{G} = \mathbf{G}(\mathbf{V}_R) - \mathbf{G}(\mathbf{V}_L) = \sum_{k=1}^l b_k \delta \mathbf{V}_k,$$

i.e.,

$$\begin{cases} -\Delta u = \sum_{k=1}^l b_k \delta \tau_k, \\ \Delta \mathbf{g} = \sum_{k=1}^l b_k \delta \Phi_k. \end{cases} \quad (3.25)$$

Observe that the first relation (3.25) already follows from (3.22). Let us then check that

$$\Delta \mathbf{F} = \mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) = \sum_{k=1}^l a_k \delta \mathbf{U}_k,$$

i.e.,

$$\begin{cases} \Delta(\rho u) = \sum_{k=1}^l a_k \delta \rho_k, \\ \Delta(\rho u \Phi + \mathbf{g}) = \sum_{k=1}^l a_k \delta(\rho \Phi)_k. \end{cases} \quad (3.26)$$

If we set for any pair  $(\varphi_k, \varphi_{k+1})$  and any  $\alpha \in [0, 1]$

$$\varphi_{k,\alpha} = \alpha \varphi_k + \beta \varphi_{k+1}, \quad \beta = 1 - \alpha,$$

we begin by noticing that (3.23) yields

$$a_k = u_{k,\alpha} + b_k \tau_{k,\alpha}.$$

We may then write

$$\Delta(\rho u) = \sum_{k=1}^l \delta(\rho u)_k = \sum_{k=1}^l (u_{k,\alpha} \delta\rho_k + \rho_{k,\beta} \delta u_k)$$

and by (3.22)

$$\Delta(\rho u) = \sum_{k=1}^l (u_{k,\alpha} \delta\rho_k - b_k \rho_{k,\beta} \delta\tau_k).$$

On the other hand, since  $\tau\rho = 1$ , we have

$$\delta(\tau\rho)_k = \tau_{k,\alpha} \delta\rho_k + \rho_{k,\beta} \delta\tau_k = 0,$$

so that

$$\Delta(\rho u) = \sum_{k=1}^l (u_{k,\alpha} + b_k \tau_{k,\alpha}) \delta\rho_k = \sum_{k=1}^l a_k \delta\rho_k.$$

Next, using the second Eq. (3.25), we obtain

$$\Delta(\rho u \Phi + \mathbf{g}) = \sum_{k=1}^l (\delta(\rho u \Phi)_k + b_k \delta\Phi_k).$$

We then write

$$\begin{aligned} \delta(\rho u \Phi)_k &= u_{k,\alpha} \delta(\rho \Phi)_k + (\rho \Phi)_{k,\beta} \delta u_k, \\ \delta\Phi_k &= \delta(\tau\rho \Phi)_k = \tau_{k,\alpha} \delta(\rho \Phi)_k + (\rho \Phi)_{k,\beta} \delta\tau_k, \end{aligned}$$

and therefore

$$\begin{aligned} \Delta(\rho u \Phi + \mathbf{g}) &= \sum_{k=1}^l \{(u_{k,\alpha} + b_k \tau_{k,\alpha}) \delta(\rho \Phi)_k + (\rho \Phi)_{k,\beta} (\delta u_k + b_k \delta\tau_k)\} = \\ &= \sum_{k=1}^l a_k \delta(\rho \Phi)_k \end{aligned}$$

so that (3.26) indeed holds.

Suppose in addition that the Lagrangian Riemann solver (3.20) is entropy satisfying: given any entropy pair  $(\eta, q)$  for the system (3.18), we thus have

$$\Delta q \leq \sum_{k=1}^l b_k \delta\eta_k. \quad (3.27)$$

For the Eulerian Riemann solver to be entropy satisfying, we have to prove that

$$\Delta(\rho u \eta + q) \leq \sum_{k=1}^l a_k \delta(\rho \eta)_k. \quad (3.28)$$

Using (3.27), we have

$$\Delta(\rho u \eta + q) \leq \sum_{k=1}^l (\delta(\rho u \eta)_k + b_k \delta \eta_k)$$

Since

$$\begin{aligned}\delta(\rho u \eta)_k &= u_{k,\alpha} \delta(\rho \eta)_k + (\rho \eta)_{k,\beta} \delta u_k, \\ \delta \eta_k &= \delta(\tau \rho \eta)_k = \tau_{k,\alpha} \delta(\rho \eta)_k + (\rho \eta)_{k,\beta} \delta \tau_k,\end{aligned}$$

we find

$$\Delta(\rho u \eta + q) \leq \sum_{k=1}^l \{(u_{k,\alpha} + b_k \tau_{k,\alpha}) \delta(\rho \eta)_k + (\rho \eta)_{k,\beta} (\delta u_k + b_k \delta \tau_k)\}$$

which yields the inequality (3.28).

We have thus checked that the Eulerian Riemann solver (3.24) is a Godunov-type (entropy satisfying) solver as soon as the Lagrangian one (3.20) is a Godunov-type (entropy satisfying) solver. Clearly, the converse property is established in a fairly similar way.  $\square$

The next sections will be devoted to the construction of several Godunov-type schemes.

## 3.2 Roe's Method and Variants

We begin with Roe's method which gives one of the most popular Godunov-type schemes although it presents several drawbacks as we will see later on. However simple modifications of Roe's method are free of these drawbacks and lead to efficient numerical methods. Roe's scheme for solving system (1.1) numerically is based on the use of an approximate Riemann solver that is obtained by linearizing the system (2.4).

### 3.2.1 Roe-Type Linearization and Roe Matrix

We begin by introducing the following definition:

*Definition 3.2*

*We call  $\mathbf{A}(\mathbf{u}, \mathbf{v})$  a Roe-type linearization if  $(\mathbf{u}, \mathbf{v}) \mapsto \mathbf{A}(\mathbf{u}, \mathbf{v})$  is a mapping from  $\Omega \times \Omega$  into the set of  $p \times p$  matrices with the following properties:*

$$\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) = \mathbf{A}(\mathbf{u}, \mathbf{v})(\mathbf{v} - \mathbf{u}), \quad (3.29)$$

$$\left\{ \begin{array}{l} \text{The } p \times p \text{ matrix } \mathbf{A}(\mathbf{u}, \mathbf{v}) \text{ has real eigenvalues and} \\ \text{a corresponding set of eigenvectors that form a basis of } \mathbb{R}^p, \end{array} \right. \quad (3.30)$$

$$\mathbf{A}(\mathbf{u}, \mathbf{u}) = \mathbf{A}(\mathbf{u}). \quad (3.31)$$

Given a Roe linearization, we consider the linearized Riemann problem

$$\begin{cases} \frac{\partial \mathbf{w}}{\partial t} + \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R) \frac{\partial \mathbf{w}}{\partial x} = \mathbf{0}, & x \in \mathbb{R}, \quad t > 0, \\ \mathbf{w}(x, 0) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0, \end{cases} \end{cases} \quad (3.32)$$

whose solution is of the form

$$\mathbf{w}(x, t) = \tilde{\mathbf{w}}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) \quad (3.33)$$

and can be explicitly constructed as explained in Chap. II, Sect. 1. Denoting by  $a_k = a_k(\mathbf{u}_L, \mathbf{u}_R)$ ,  $1 \leq k \leq p$ , the *distinct* eigenvalues of the Roe matrix  $A(\mathbf{u}_L, \mathbf{u}_R)$ , we know that  $\tilde{\mathbf{w}}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right)$  is indeed of the form (3.10) with

$$\mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)(\mathbf{w}_{k+1} - \mathbf{w}_k) = a_k(\mathbf{w}_{k+1} - \mathbf{w}_k), \quad 1 \leq k \leq p, \quad (3.34)$$

(the  $\mathbf{w}_k$ 's naturally represent the intermediate states in the solution of the linearized Riemann problem). Now, we define Roe's scheme to be the Godunov-type scheme associated with the approximate Riemann solver (3.33). Clearly, condition (3.1) holds. Moreover, using (3.29) and (3.34), we have

$$\begin{aligned} \sum_{k=1}^p a_k(\mathbf{w}_{k+1} - \mathbf{w}_k) &= \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)\left(\sum_{k=1}^p (\mathbf{w}_{k+1} - \mathbf{w}_k)\right) = \\ &= \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)(\mathbf{u}_R - \mathbf{u}_L) = \mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) \end{aligned}$$

so that the condition (3.12) holds. Hence, by Theorem 3.1, Roe's method can be put in conservation form and is consistent with the nonlinear system of conservation laws (1.1). On the other hand, we will see later on that the entropy consistency condition (3.4) or equivalently the condition (3.14) may fail to be satisfied which implies that Roe's method can provide nonphysical approximate solutions.

*Remark 3.1.* Classically, the Roe linearization  $\mathbf{A}(\mathbf{u}, \mathbf{v})$  is required to satisfy the additional consistency property (3.3)

$$\mathbf{A}(\mathbf{u}, \mathbf{u}) = \mathbf{A}(\mathbf{u}).$$

In fact, although such a consistency condition seems fairly natural, it is not at all necessary for ensuring the consistency of the associated Roe's method

(see (3.35) below). In the scalar case, however, the condition comes from the identity  $a(u, v) = \frac{f(u)-f(v)}{u-v}$  if  $v \neq u$ .  $\square$

It remains to check that Roe linearizations indeed exist and to construct them explicitly in some cases of practical importance. Let us prove the existence of such a linearization when the system (1.1) admits an entropy (Harten–Lax Theorem; see Harten [590]).

*Theorem 3.4*

*Assume that the hyperbolic system (1.1) has a strictly convex entropy  $U$ . Then there exists at least one Roe-type linearization.*

*Proof.* We look for a matrix  $\mathbf{A} = \mathbf{A}(\mathbf{u}, \mathbf{v})$  of the form

$$\mathbf{A} = \mathbf{S} \cdot \mathbf{P},$$

where  $\mathbf{S}$  and  $\mathbf{P}$  are  $p \times p$  matrices and  $\mathbf{P}$  is, moreover, symmetric positive definite. Then  $\mathbf{A}$  is similar to the symmetric matrix  $\mathbf{P}^{\frac{1}{2}} \mathbf{S} \mathbf{P}^{\frac{1}{2}}$ , which already meets the condition (3.2). It remains to construct the matrices  $\mathbf{S} = \mathbf{S}(\mathbf{u}, \mathbf{v})$  and  $\mathbf{P} = \mathbf{P}(\mathbf{u}, \mathbf{v})$ . The construction follows the arguments of the proof of Theorem 5.2 in the Chap. I.

Let  $U$  be a strictly convex entropy. Then  $U''(\mathbf{u})$  is a symmetric positive definite matrix, and the mapping  $\mathbf{u} \mapsto U'(\mathbf{u})$  is one to one, so that we can define the change of variables (the entropy variables; see the Chap. I, Sect. 5, Theorem 5.2)

$$\mathbf{w}^T = U'(\mathbf{u}),$$

and we set

$$\mathbf{g}(\mathbf{w}) = \mathbf{f}(\mathbf{u}(\mathbf{w})).$$

Let us show that the matrix  $\mathbf{g}'(\mathbf{w})$  is symmetric. On the one hand, we know that the matrix  $U''(\mathbf{u})\mathbf{f}'(\mathbf{u})$  is symmetric and therefore the matrix

$$\mathbf{f}'(\mathbf{u})U''(\mathbf{u})^{-1} = U''(\mathbf{u})^{-1}(U''(\mathbf{u})\mathbf{f}'(\mathbf{u}))U''(\mathbf{u})^{-1}$$

is also symmetric. On the other hand, we have  $\mathbf{u} = \mathbf{u}(\mathbf{w})$  and

$$\mathbf{u}'(\mathbf{w})\mathbf{w}'(\mathbf{u}) = \mathbf{I}$$

so that, since  $\mathbf{w}'(\mathbf{u}) = U''(\mathbf{u})$ ,

$$\mathbf{u}'(\mathbf{w}) = U''(\mathbf{u})^{-1}.$$

Hence, the matrix

$$\mathbf{g}'(\mathbf{w}) = \mathbf{f}'(\mathbf{u}(\mathbf{w})) \cdot \mathbf{u}'(\mathbf{w}) = \mathbf{f}'(\mathbf{u}(\mathbf{w})) \cdot U''(\mathbf{u})^{-1}$$

is symmetric.

Now, given two states  $\mathbf{u}_1, \mathbf{u}_2 \in \Omega$ , we set

$$\mathbf{w}_1 = \mathbf{w}(\mathbf{u}_1), \quad \mathbf{w}_2 = \mathbf{w}(\mathbf{u}_2)$$

and write

$$\begin{aligned} \mathbf{g}(\mathbf{w}_2) - \mathbf{g}(\mathbf{w}_1) &= \int_0^1 \frac{d}{d\theta} \mathbf{g}(\theta \mathbf{w}_2 + (1 - \theta) \mathbf{w}_1) d\theta \\ &= \left\{ \int_0^1 \mathbf{g}'(\theta \mathbf{w}_2 + (1 - \theta) \mathbf{w}_1) d\theta \right\} \cdot (\mathbf{w}_2 - \mathbf{w}_1). \end{aligned}$$

Defining

$$\mathbf{S} = \mathbf{S}(\mathbf{u}_1, \mathbf{u}_2) = \int_0^1 \mathbf{g}'(\theta \mathbf{w}_2 + (1 - \theta) \mathbf{w}_1) d\theta,$$

we obtain that the matrix  $\mathbf{S}$  is symmetric and satisfies

$$\mathbf{f}(\mathbf{u}_2) - \mathbf{f}(\mathbf{u}_1) = \mathbf{g}(\mathbf{w}_2) - \mathbf{g}(\mathbf{w}_1) = \mathbf{S}(\mathbf{w}_2 - \mathbf{w}_1).$$

On the other hand, we have

$$\begin{aligned} \mathbf{w}_2 - \mathbf{w}_1 &= \int_0^1 \frac{d}{d\theta} \mathbf{w}(\theta \mathbf{u}_2 + (1 - \theta) \mathbf{u}_1) d\theta \\ &= \left\{ \int_0^1 \mathbf{w}'(\theta \mathbf{u}_2 + (1 - \theta) \mathbf{u}_1) d\theta \right\} \cdot (\mathbf{u}_2 - \mathbf{u}_1). \end{aligned}$$

We thus define

$$\mathbf{P} = \mathbf{P}(\mathbf{u}_1, \mathbf{u}_2) = \int_0^1 \mathbf{w}'(\theta \mathbf{u}_2 + (1 - \theta) \mathbf{u}_1) d\theta,$$

so that

$$\mathbf{w}_2 - \mathbf{w}_1 = \mathbf{P}(\mathbf{u}_2 - \mathbf{u}_1),$$

and it follows from the symmetry of the matrix  $\mathbf{w}'(\mathbf{u}) = U''(\mathbf{u})$  that the matrix  $\mathbf{P}$  is symmetric. Hence, we obtain

$$\mathbf{f}(\mathbf{u}_2) - \mathbf{f}(\mathbf{u}_1) = \mathbf{S}\mathbf{P}(\mathbf{u}_2 - \mathbf{u}_1),$$

so that the condition (3.1) holds.

Finally, we notice that

$$\mathbf{S}(\mathbf{u}, \mathbf{u}) = \int_0^1 \mathbf{g}'(\theta \mathbf{w} + (1 - \theta) \mathbf{w}) d\theta = \mathbf{g}'(\mathbf{w}(\mathbf{u})),$$

$$\mathbf{P}(\mathbf{u}, \mathbf{u}) = \int_0^1 \mathbf{w}'(\mathbf{u}) d\theta = \mathbf{w}'(\mathbf{u}),$$

and also

$$\mathbf{A}(\mathbf{u}, \mathbf{u}) = \mathbf{S}(\mathbf{u}, \mathbf{u})\mathbf{P}(\mathbf{u}, \mathbf{u}) = \mathbf{g}'(\mathbf{w}(\mathbf{u})) \cdot \mathbf{w}'(\mathbf{u}) = \mathbf{f}'(\mathbf{u}) = \mathbf{A}(\mathbf{u}),$$

which is the consistency condition (3.3).  $\square$

*Remark 3.2.* It is straightforward to check that

$$\mathbf{P}(\mathbf{u}_2, \mathbf{u}_1) = \mathbf{P}(\mathbf{u}_1, \mathbf{u}_2), \quad \mathbf{S}(\mathbf{u}_2, \mathbf{u}_1) = \mathbf{S}(\mathbf{u}_1, \mathbf{u}_2),$$

and therefore the Roe linearization constructed in the proof of the above theorem satisfies

$$\mathbf{A}(\mathbf{u}, \mathbf{v}) = \mathbf{A}(\mathbf{v}, \mathbf{u}).$$

This means that we can reverse the roles of  $\mathbf{u}$  and  $\mathbf{v}$ .  $\square$

### 3.2.2 The Numerical Flux of Roe's Scheme

Given a Roe-type linearization and the corresponding approximate Riemann solver (3.4),(3.5), we now detail the associated Roe's scheme. Assume that

$$\lambda \max_{1 \leq k \leq p} |a_k(\mathbf{u}_L, \mathbf{u}_R)| \leq \frac{1}{2}.$$

Then, by Theorem 3.2, the numerical flux function is given by

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \left( \mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v}) - \sum_{k=1}^p |a_k| (\mathbf{w}_{k+1} - \mathbf{w}_k) \right).$$

Since by (3.34)

$$\begin{aligned} \sum_{k=1}^p |a_k(\mathbf{u}_L, \mathbf{u}_R)| (\mathbf{w}_{k+1} - \mathbf{w}_k) &= \sum_{k=1}^p |\mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)| (\mathbf{w}_{k+1} - \mathbf{w}_k) = \\ &= |\mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)| \sum_{k=1}^p (\mathbf{w}_{k+1} - \mathbf{w}_k) = |\mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)| (\mathbf{u}_R - \mathbf{u}_L), \end{aligned}$$

we obtain

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \left( \mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v}) - |\mathbf{A}(\mathbf{u}, \mathbf{v})| (\mathbf{u} - \mathbf{v}) \right). \quad (3.35)$$

Hence, setting

$$\mathbf{A}_{j+1/2}^n = \mathbf{A}(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n), \quad (3.36)$$

we have

$$\mathbf{g}_{j+1/2}^n - \mathbf{g}_{j-1/2}^n = \frac{1}{2} \left\{ \mathbf{f}(\mathbf{v}_{j+1}^n) - \mathbf{f}(\mathbf{v}_j^n) - |\mathbf{A}_{j+1/2}^n|(\mathbf{v}_{j+1}^n - \mathbf{v}_j^n) + \right. \\ \left. + \mathbf{f}(\mathbf{v}_j^n) - \mathbf{f}(\mathbf{v}_{j-1}^n) - |\mathbf{A}_{j-1/2}^n|(\mathbf{v}_j^n - \mathbf{v}_{j-1}^n) \right\}$$

and by (3.1)

$$\mathbf{g}_{j+1/2}^n - \mathbf{g}_{j-1/2}^n = \frac{1}{2} \left\{ (\mathbf{A}_{j+1/2}^n - |\mathbf{A}_{j+1/2}^n|)(\mathbf{v}_{j+1}^n - \mathbf{v}_j^n) + \right. \\ \left. + (\mathbf{A}_{j-1/2}^n + |\mathbf{A}_{j-1/2}^n|)(\mathbf{v}_j^n - \mathbf{v}_{j-1}^n) \right\}.$$

We then obtain Roe's numerical scheme

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \left\{ (\mathbf{A}_{j+1/2}^n)^-(\mathbf{v}_{j+1}^n - \mathbf{v}_j^n) + (\mathbf{A}_{j-1/2}^n)^+(\mathbf{v}_j^n - \mathbf{v}_{j-1}^n) \right\}, \quad (3.37)$$

provided with the CFL-like condition

$$\lambda \max_j |a_k(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n)| \leq \frac{1}{2}, \quad 1 \leq k \leq p. \quad (3.38)$$

We note the similarity of (3.3) with Godunov's scheme in the linear case (2.16); the constant matrix  $\mathbf{A}$  is only replaced by Roe matrix.

Let us give a more explicit form of Roe's scheme. We denote by  $\mathbf{r}_k(\mathbf{u}, \mathbf{v})$ ,  $1 \leq k \leq p$ , the eigenvectors of the matrix  $\mathbf{A}(\mathbf{u}, \mathbf{v})$  associated with the eigenvalues  $a_k(\mathbf{u}, \mathbf{v})$ , i.e.,

$$\mathbf{A}(\mathbf{u}, \mathbf{v})\mathbf{r}_k(\mathbf{u}, \mathbf{v}) = a_k(\mathbf{u}, \mathbf{v})\mathbf{r}_k(\mathbf{u}, \mathbf{v}), \quad 1 \leq k \leq p.$$

We define the corresponding characteristic variables  $\alpha_k(\mathbf{u}, \mathbf{v})$ ,  $1 \leq k \leq p$  by  $\alpha_k(\mathbf{u}, \mathbf{v}) = \mathbf{l}_k^T(\mathbf{u}, \mathbf{v})(\mathbf{v} - \mathbf{u})$ , or

$$\mathbf{v} - \mathbf{u} = \sum_{k=1}^p \alpha_k(\mathbf{u}, \mathbf{v})\mathbf{r}_k(\mathbf{u}, \mathbf{v}). \quad (3.39)$$

Then (3.35) reads

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \left( \mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v}) - \sum_{k=1}^p (\alpha_k |a_k| \mathbf{r}_k)(\mathbf{u}, \mathbf{v}) \right). \quad (3.40)$$

Similarly, we can write (3.37) in the form

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \sum_{k=1}^p \left\{ (\alpha_k a_k^- \mathbf{r}_k)_{j+1/2}^n + (\alpha_k a_k^+ \mathbf{r}_k)_{j-1/2}^n \right\} \quad (3.41)$$

where

$$(\mathbf{r}_k)_{j+1/2}^n = \mathbf{r}_k(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n),$$

and so on.

### 3.2.3 Entropy Correction

A well-known drawback of Roe's scheme is that it may resolve nonphysical solutions (see G.R., Chapter 3, Examples 2.4 and 4.1 [539]; see also Einfeldt et al. [458]). Indeed, by (3.29) and (3.35), for a stationary discontinuity connecting two states  $\mathbf{u}$  and  $\mathbf{v}$ ,  $\mathbf{f}(\mathbf{u}) = \mathbf{f}(\mathbf{v})$  implies  $\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{u})$  whether or not the entropy condition is satisfied. Various entropy corrections have been proposed (see Harten and Hyman [593], Roe, Roe and Pike, and Huynh [643]; Yee's formula is given in Lin [802]). If we write the scheme (3.37) in viscous form (1.13), we see that the viscosity matrix is

$$\mathbf{Q}(\mathbf{u}, \mathbf{v}) = \lambda |\mathbf{A}(\mathbf{u}, \mathbf{v})|,$$

and

$$\mathbf{Q}(\mathbf{u}, \mathbf{v})(\mathbf{v} - \mathbf{u}) = \lambda \sum_{k=1}^p (\alpha_k |a_k| \mathbf{r}_k)(\mathbf{u}, \mathbf{v}). \quad (3.42)$$

In the basis of eigenvectors  $\mathbf{r}_k(\mathbf{u}, \mathbf{v})$ , the viscosity matrix is diagonal,  $\mathbf{Q} = \text{diag}(\lambda |a_k(\mathbf{u}, \mathbf{v})|)$ , and we may think that Roe's scheme is not viscous enough in the neighborhood of the sonic points where an eigenvalue vanishes. Harten [590] has proposed replacing the function  $Q(x) = \lambda|x|$  in the diagonal of  $\mathbf{Q}$  by a smooth quadratic regularization  $Q_\delta$ ,

$$Q_\delta(x) = \begin{cases} \lambda|x|, & |x| \geq \delta, \\ \lambda \left( \frac{x^2}{2\delta} + \frac{\delta}{2} \right), & |x| < \delta, \end{cases}$$

which corresponds to adding viscosity near sonic points. Unfortunately, the actual tuning of the parameter  $\delta$  is empirical and depends on the problem considered. Roe's “spreading device” (Roe 1985) does not contain an arbitrary parameter but is not so easily exhibited.

We can also extend to systems the modification of Harten and Hyman that we have presented in the scalar case (see G.R., Chapter 3, Example 4.1 [539]). Indeed (see Chap. II, Sect. 1), the initial discontinuity in the solution of the linear Riemann problem (3.4) breaks up into  $p$  discontinuity waves that propagate with speed  $a_k(\mathbf{u}_L, \mathbf{u}_R)$ ,  $1 \leq k \leq p$ ,

$$\mathbf{w}(x, t) = \sum_{k=1}^p w_k(x, t) \mathbf{r}_k(\mathbf{u}, \mathbf{v}).$$

where

$$w_k(x, t) = \begin{cases} \alpha_{kL}, & \frac{x}{t} < a_k(\mathbf{u}_L, \mathbf{u}_R), \\ \alpha_{kR}, & \frac{x}{t} > a_k(\mathbf{u}_L, \mathbf{u}_R), \end{cases}$$

where  $\alpha_{kL}$  and  $\alpha_{kR}$  are, respectively, the components of  $\mathbf{u}_L$  and  $\mathbf{u}_R$  in the basis of eigenvectors  $\mathbf{r}_k(\mathbf{u}_L, \mathbf{u}_R)$ . If the  $k$ th field is genuinely nonlinear (which corresponds to the convex scalar case) and if the corresponding  $k$ th elementary wave of the exact solution of the Riemann problem (2.1) is a rarefaction wave, we introduce an intermediate state

$$w_k(x, t) = \begin{cases} \alpha_{kL}, & \frac{x}{t} < a_k^L, \\ \alpha_k^*, & a_k^L < \frac{x}{t} < a_k^R, \\ \alpha_{kR}, & \frac{x}{t} > a_k^R. \end{cases}$$

If  $a_k^L$ ,  $a_k^R$ , and  $\delta_k$  are chosen as in the scalar case,

$$\begin{aligned} a_k^L &= a_k(\mathbf{u}_L, \mathbf{u}_R) - \delta_k, & a_k^R &= a_k(\mathbf{u}_L, \mathbf{u}_R) + \delta_k, \\ \delta_k &= \sup(0, a_k(\mathbf{u}_L, \mathbf{u}_R) - a_k(\mathbf{u}_L, \mathbf{u}), a_k(\mathbf{u}, \mathbf{u}_R) - a_k(\mathbf{u}_L, \mathbf{u}_R)), \end{aligned}$$

the supremum is taken over all  $\mathbf{u} = \theta\mathbf{u}_L + (1 - \theta)\mathbf{u}_R$  and

$$\alpha_k^* = \frac{1}{2}(\alpha_{kL} + \alpha_{kR}).$$

It amounts to modifying the viscosity componentwise, i.e., to replacing the  $k$ th diagonal element  $Q_k = \lambda|a_k|$  of  $\mathbf{Q}$  by

$$Q_k = \begin{cases} \lambda|a_k|, & |a_k| \geq \delta_k, \\ \lambda\delta_k, & |a_k| < \delta_k. \end{cases}$$

We mention another approach based on a nonlinear smooth modification of the flux function (by a Hermite polynomial of degree 3) at sonic points only, and that gives very good results (see Dubois and Mehlman [438]). Moreover, interesting considerations can be found in Lin [802]; see also [429, 985].

Finally, among the other ‘‘entropy fix,’’ we present the LLF (Shu and Osher [1059]), which we begin by detailing in the scalar case. One defines first a *local Lax–Friedrichs scheme*, whose numerical flux is

$$g_{j+\frac{1}{2}}^{LLF} = \frac{1}{2}(f(v_j) + f(v_{j+1}) - \sigma_{j+\frac{1}{2}} \Delta v_{j+\frac{1}{2}}),$$

where

$$\sigma_{j+\frac{1}{2}} = \max\{|f'(u)|, u \in [v_j, v_{j+1}]\}.$$

The classical Lax–Friedrichs scheme would correspond to  $\sigma_{j+\frac{1}{2}} = 1/\lambda$ , which is the upper bound of all the  $\sigma_{j+\frac{1}{2}}$  (under  $\text{CFL} \leq 1$ ). Assuming, for instance, that  $f$  is convex,

$$\sigma_{j+\frac{1}{2}} = \max\{|f'(v_j)|, |f'(v_{j+1})|\},$$

it is easy to check that the scheme is monotone (under  $\text{CFL} \leq 1 - \varepsilon$ ) by computing the partial derivatives  $\partial_{v_i} H^{LLF}(v_{-1}, v_0, v_1)$ ,  $i = -1, 0, 1$  (see G.R., Chapter 3, Section 3.1 [539]). Its numerical viscosity is given by

$$Q_{j+\frac{1}{2}} = \lambda \sigma_{j+\frac{1}{2}}.$$

Then one considers *Roe's scheme with LLF*: at a sonic point, one turns back from the upwind flux to the above local Lax–Friedrichs flux; thus

$$g_{j+\frac{1}{2}} = \begin{cases} f(v_j) & \text{if } f'(u) \geq 0, \quad u \text{ between } v_j, v_{j+1}, \\ f(v_{j+1}) & \text{if } f'(u) \leq 0, \quad u \text{ between } v_j, v_{j+1}, \\ g_{j+\frac{1}{2}}^{LLF} & \text{otherwise.} \end{cases}$$

For a system, we apply the procedure to each characteristic field. Using the same notations as above,  $(\alpha_k, a_k, \mathbf{r}_k, \mathbf{l}_k)_{j+\frac{1}{2}}^n$ , the  $k$ th component  $\Psi_k$  of the LLF flux  $\Psi$  on the eigenbasis  $\mathbf{r}_{k,j+\frac{1}{2}}$  is

$$\Psi_{k,j+\frac{1}{2}} = \frac{1}{2}(\mathbf{f}(\mathbf{v}_j) + (\mathbf{f}(\mathbf{v}_{j+1}))_k - \frac{1}{2}\sigma_{k,j+\frac{1}{2}}\alpha_{k,j+\frac{1}{2}}),$$

where  $(\mathbf{f}(\mathbf{v}_j))_k$  denotes the  $k$ th component of  $\mathbf{f}(\mathbf{v}_j)$  in the same basis and the coefficients  $\sigma_k$  are defined by

$$\sigma_{k,j+\frac{1}{2}} = \max\{a_k(\mathbf{v}_j), a_k(\mathbf{v}_{j+1})\}$$

(if all the characteristic fields are either genuinely nonlinear or linearly degenerate, which holds for the gas dynamics equations). The  $a_k(\mathbf{v}_j)$ 's denote the eigenvalues of the matrix  $\mathbf{A}(\mathbf{v}_j)$ , whereas the  $a_{k,j+\frac{1}{2}}$  denote the eigenvalues of the matrix  $\mathbf{A}_{j+\frac{1}{2}}$ . Then

$$\mathbf{g}(\mathbf{v}_j, \mathbf{v}_{j+1}) = \sum_k (\Phi_k \mathbf{r}_k)_{j+\frac{1}{2}},$$

where if the signs of  $a_k(\mathbf{v}_j)$  and  $a_k(\mathbf{v}_{j+1})$  are identical, we take for  $\Phi_k$  the  $k$ th component of the upwind flux, i.e.,

$$\Phi_{k,j+\frac{1}{2}} = \begin{cases} (\mathbf{f}(\mathbf{v}_j))_k, & \text{if both } a_k \geq 0, \\ (\mathbf{f}(\mathbf{v}_{j+1}))_k, & \text{if both } a_k < 0. \end{cases}$$

In the “sonic case,” for an index  $k$  such that the signs of  $a_k(\mathbf{v}_j)$  and  $a_k(\mathbf{v}_{j+1})$  are not identical, we turn to the LLF flux and take for  $\Phi_k$  the  $k$ th component  $\Psi_k$ .

*Remark 3.3.* In Roe’s scheme, the approximate Riemann solver contains a priori  $p - 1$  intermediate states between  $\mathbf{v}_j$  and  $\mathbf{v}_{j+1}$  separated by the signal velocities  $a_k(\mathbf{v}_j, \mathbf{v}_{j+1})$ . We must then mention another drawback of Roe’s scheme. When used for the approximation of the Euler equations, it may generate in some particular cases nonphysical intermediate states, for instance, with negative internal energy or density (see Einfeldt et al. [458], who describe this scheme as non *positively conservative*). Also, the eigenvalues  $a_k(\mathbf{v}_i, \mathbf{v}_{i+1}), k = 1 \text{ or } 3$ , could lie outside the range of values  $(a_k(\mathbf{v}_i), a_k(\mathbf{v}_{i+1}))$ , respectively (see Vinokur [1172] and Einfeldt et al. [458]).  $\square$

### 3.2.4 The VFRoe Scheme

In some applications, it is difficult to compute even a Roe matrix, and the VFRoe scheme introduces a simplification. The VFRoe scheme still involves a linearization of the flux by introducing the Jacobian matrix at some intermediate state, as we will soon see is the case for Roe’s scheme in many applications. However, it is no longer required to satisfy the property (3.29) in the definition of a Roe-type linearization, so that “any” averaged state is eligible, for instance, the mean value; the matrix has real eigenvalues and a basis of eigenvectors.

The associated numerical flux writes

$$g_{j+1/2} = g^{VFRoe}(\mathbf{u}_j, \mathbf{u}_{j+1}) = \mathbf{F}(\mathbf{w}_R(0; \mathbf{u}_j, \mathbf{u}_{j+1}; \bar{A}_{j+1/2}))$$

if we note  $\bar{A}_{j+1/2} = \bar{A}(\mathbf{u}_j, \mathbf{u}_{j+1})$  the linearization and  $\mathbf{w}_R(., .; \bar{A}_{j+1/2})$  the solution of the associated linear Riemann problem. As for Roe’s scheme, the formulas are easily written in terms of the eigenvectors.

Its stability properties in the scalar case are studied in [855]: the scheme is proved to be TVD and  $L^\infty$ -stable if  $\bar{a}(u_j, u_{j+1})a(u_j, u_{j+1}) \geq 0$  (under a usual CFL condition).

Some sonic entropy correction is possible, following the ideas developed for Roe’s scheme. Some drawbacks are known for this scheme, for instance, nonpositive density and pressure values in the case of the Euler system, even if in many cases it shows a good behavior.

The VFRoe $_{ncv}$  (where  $ncv$  stands for *nonconservative*) presents a systematic way to choose the average state by linearizing the system written in a chosen nonconservative set of variables. Assume that smooth solutions satisfy

$$\partial_t \mathbf{w} + B(\mathbf{w}) \partial_x \mathbf{w} = \mathbf{0}$$

where  $\mathbf{w} = \mathbf{w}(\mathbf{u})$  is an admissible set of nonconservative variables and  $B(\mathbf{w}) = \mathbf{u}'(\mathbf{w})^{-1} A(\mathbf{u}(\mathbf{w})) \mathbf{u}'(\mathbf{w})$ . Then setting  $\mathbf{u}_j = \mathbf{w}(\mathbf{u}_j)$ , the linearization of  $B$  at states  $\mathbf{u}_j, \mathbf{u}_{j+1}$  is defined as  $B(\hat{\mathbf{W}})$ , for some average state  $\hat{\mathbf{W}}(\mathbf{w}_j, \mathbf{w}_{j+1})$  function of the states  $\mathbf{w}_j, \mathbf{w}_{j+1}$ , for instance, the arithmetic mean  $\frac{1}{2}(\mathbf{w}_j + \mathbf{w}_{j+1})$ . Consider the corresponding linearized Riemann problem in terms of the variables  $\mathbf{w}$ , i.e., at the interface  $x = x_{j+1/2}$ , solve

$$\partial_t \mathbf{w} + B(\hat{\mathbf{W}}_{j+1/2}) \partial_x \mathbf{w} = \mathbf{0}$$

with data  $\mathbf{w}_j, \mathbf{w}_{j+1}$ . Denoting by  $\hat{\mathbf{W}}_{j+1/2}^*$  the solution on the interface and setting  $\mathbf{U}_{j+1/2}^* = \mathbf{u}(\hat{\mathbf{W}}_{j+1/2}^*)$ , we take as  $\mathbf{F}(\mathbf{U}_{j+1/2}^*)$  the numerical flux of the VFRoe scheme.

For example, in the application to the Euler system, one can choose the primitive set of variables  $(\tau, u, p)$ . We refer to [215, 855], and [501] and [500] for more general equations of state and also [502] and for application to the shallow water system [503] and also [141].

### 3.3 The H.L.L. Method

The method is called after Ami Harten, Peter Lax, and Bram van Leer who popularized the notion of Godunov-type scheme in a well-known paper (1983, [595]). In fact, the simplest Godunov-type method consists in using a simple Riemann solver with a single intermediary state, i.e., of the form

$$\tilde{\mathbf{w}}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) = \begin{cases} \mathbf{u}_L, & \frac{x}{t} < a_L, \\ \mathbf{u}_*, & a_L < \frac{x}{t} < a_R, \\ \mathbf{u}_R, & \frac{x}{t} > a_R. \end{cases} \quad (3.43)$$

The associated Godunov-type method is called the H.L.L. method (for Harten, Lax, and van Leer). The property (3.3) (of consistency with the integral form of the system of conservation laws) reads here

$$\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) = a_L(\mathbf{u}_* - \mathbf{u}_L) - a_L(\mathbf{u}_R - \mathbf{u}_*).$$

Hence, given the wave velocities  $a_L$  and  $a_R$ , the intermediate state  $\mathbf{u}_*$  is uniquely determined by

$$\mathbf{u}_* = \frac{1}{a_R - a_L} \{a_R \mathbf{u}_R - a_L \mathbf{u}_L - (\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L))\}. \quad (3.44)$$

*Proposition 3.1*

The numerical flux of the H.L.L. method is given by

$$\mathbf{g}(\mathbf{u}_L, \mathbf{u}_R) = \frac{1}{a_R^+ - a_L^-} \{ a_R^+ \mathbf{f}(\mathbf{u}_L) - a_L^- \mathbf{f}(\mathbf{u}_R) + a_R^+ a_L^- (\mathbf{u}_R - \mathbf{u}_L) \}. \quad (3.45)$$

In (3.45), we have set as usual  $a^+ = \max(a, 0)$ ,  $a^- = \min(a, 0)$ .

*Proof.* Using (3.44), we have

$$\begin{aligned} \mathbf{u}_* - \mathbf{u}_L &= \frac{1}{a_R - a_L} \{ a_R (\mathbf{u}_R - \mathbf{u}_L) - (\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)) \}, \\ \mathbf{u}_R - \mathbf{u}_* &= \frac{1}{a_R - a_L} \{ a_L (\mathbf{u}_R - \mathbf{u}_L) + (\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)) \}. \end{aligned}$$

Therefore (3.13) gives here

$$\begin{aligned} \mathbf{g}(\mathbf{u}_L, \mathbf{u}_R) &= \frac{1}{2} \{ \mathbf{f}(\mathbf{u}_L) + \mathbf{f}(\mathbf{u}_R) - |a_L|(\mathbf{u}_* - \mathbf{u}_L) - |a_R|(\mathbf{u}_R - \mathbf{u}_*) \} = \\ &= \frac{1}{2(a_R - a_L)} \{ ((a_R + |a_R|) - (a_L + |a_L|))\mathbf{f}(\mathbf{u}_L) + \\ &\quad + ((a_R - |a_R|) - (a_L - |a_L|))\mathbf{f}(\mathbf{u}_R) + +(|a_R|a_L - |a_L|a_R)(\mathbf{u}_R - \mathbf{u}_L) \} \end{aligned}$$

or

$$\mathbf{g}(\mathbf{u}_L, \mathbf{u}_R) = \begin{cases} \mathbf{f}(\mathbf{u}_R) & \text{if } a_R \leq 0, \\ \frac{1}{a_R - a_L} (a_R \mathbf{f}(\mathbf{u}_L) - a_L \mathbf{f}(\mathbf{u}_R) + a_L a_R (\mathbf{u}_R - \mathbf{u}_L)) & \text{if } a_L \leq 0 \leq a_R, \\ \mathbf{f}(\mathbf{u}_L) & \text{if } a_L \geq 0 \end{cases}$$

or equivalently the formula (3.45).  $\square$

Several remarks are now in order concerning the choice of the numerical velocities  $a_\alpha(\mathbf{u}_L, \mathbf{u}_R)$ ,  $\alpha = L, R$ . For stability reasons, a bound on these velocities is needed, which we will study below. Let us first consider some specific situations.

First, if we choose  $a_R = -a_L = \frac{1}{\lambda}$ , we obtain

$$\mathbf{g}(\mathbf{u}_L, \mathbf{u}_R) = \frac{1}{2} \left( \mathbf{f}(\mathbf{u}_L) + \mathbf{f}(\mathbf{u}_R) - \frac{1}{\lambda} (\mathbf{u}_R - \mathbf{u}_L) \right)$$

which is in fact the numerical flux of the Lax–Friedrichs method. On the other hand, if we consider a discontinuity wave

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_L, & x < \sigma t, \\ \mathbf{u}_R, & x > \sigma t \end{cases}$$

with the Rankine–Hugoniot jump conditions

$$\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) = \sigma(\mathbf{u}_R - \mathbf{u}_L),$$

we obtain from (3.44)

$$\mathbf{u}_* = \frac{1}{a_R - a_L}((\sigma - a_L)\mathbf{u}_L + (a_R - \sigma)\mathbf{u}_R)$$

so that the approximate Riemann solver  $\tilde{\mathbf{w}}_R(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$  coincides with  $\mathbf{u}(x, t) = \mathbf{w}_R(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$  if and only if  $\sigma = a_L$  or  $\sigma = a_R$ .

In the general case, it remains to choose the wave velocities  $a_L$  and  $a_R$  in such a way that the H.L.L. scheme is entropy satisfying. Let us first state a general result.

*Theorem 3.5*

Assume that  $a_L$  (resp.  $a_R$ ) is a lower bound (resp. an upper bound) for the wave velocities of the exact solution  $\mathbf{w}(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$  of the Riemann problem. Then the corresponding H.L.L. method is entropy satisfying.

*Proof.* Assuming as usual

$$\lambda \max(|a_L|, |a_R|) \leq \frac{1}{2},$$

we first check that

$$\mathbf{u}_* = \frac{1}{(a_R - a_L)\Delta t} \int_{a_L \Delta t}^{a_R \Delta t} \mathbf{w}\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) dx. \quad (3.46)$$

Indeed, the consistency property (with the integral form of the system of conservation laws) means that

$$\int_{-\Delta x/2}^{\Delta x/2} \tilde{\mathbf{w}}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) dx = \int_{-\Delta x/2}^{\Delta x/2} \mathbf{w}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) dx.$$

Now writing

$$\int_{-\Delta x/2}^{\Delta x/2} = \int_{-\Delta x/2}^{a_L \Delta t} + \int_{a_L \Delta t}^{a_R \Delta t} + \int_{a_R \Delta t}^{\Delta x/2}$$

and observing that by hypothesis

$$\mathbf{w}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) = \begin{cases} \mathbf{u}_L, & x < a_L \Delta t \\ \mathbf{u}_R, & x > a_R \Delta t \end{cases},$$

we obtain

$$(a_L \Delta t + \frac{\Delta x}{2}) \mathbf{u}_L + (a_R - a_L) \Delta t \mathbf{u}_* + (\frac{\Delta x}{2} - a_R \Delta t) \mathbf{u}_R = \\ = (a_L \Delta t + \frac{\Delta x}{2}) \mathbf{u}_L + \int_{a_L \Delta t}^{a_R \Delta t} \mathbf{w}_R(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R) dx + (\frac{\Delta x}{2} - a_R \Delta t) \mathbf{u}_R$$

which yields (3.46).

As a consequence, we find that  $\mathbf{u}_*$  belongs to the state space  $\Omega$ . Next, using Jensen's inequality, we obtain for any convex function  $U : \Omega \rightarrow \mathbb{R}$

$$U(\mathbf{u}_*) = U\left(\frac{1}{(a_R - a_L)\Delta t} \int_{a_L \Delta t}^{a_R \Delta t} \mathbf{w}\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) dx\right) \leq \\ \leq \left(\frac{1}{(a_R - a_L)\Delta t} \int_{a_L \Delta t}^{a_R \Delta t} U\left(\mathbf{w}\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right)\right) dx\right)$$

and therefore

$$\int_{-\Delta x/2}^{\Delta x/2} U\left(\tilde{\mathbf{w}}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right)\right) dx \leq \int_{-\Delta x/2}^{\Delta x/2} U\left(\mathbf{w}_R\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right)\right) dx$$

which implies the consistency property (3.4) with the integral form of the entropy inequality. The result then follows from Theorem 3.1.  $\square$

We are left with the problem of finding explicit bounds for the wave velocities of  $\mathbf{w}_R(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$ . This is an easy problem when the extreme waves are rarefaction waves or contact discontinuities since the wave velocities are then bounded below by the smallest eigenvalue of  $\mathbf{A}(\mathbf{u}_L)$  and bounded above by the largest eigenvalue of  $\mathbf{A}(\mathbf{u}_R)$ . This is more complicated when these extreme waves are shocks. Indeed, as we have already noticed it, if we choose  $a_L$  and  $a_R$  to be the shock speeds, we represent exactly these waves. However, determining these speeds amounts to solve exactly the Riemann problem that we want to avoid. Hence, we need to find "reasonable" approximations of these velocities at a low cost. If we are able to do so, the H.L.L. method will provide good approximations of 1- and  $p$ -shocks.

In order to make this point more precise and to give a practical way of determining  $a_L$  and  $a_R$ , we study the dissipation properties of the H.L.L. scheme. We introduce the numerical viscosity matrix  $\mathbf{Q}(\mathbf{u}_L, \mathbf{u}_R)$  of the method defined by

$$\mathbf{g}(\mathbf{u}_L, \mathbf{u}_R) = \frac{1}{2} \{ \mathbf{f}(\mathbf{u}_L) + \mathbf{f}(\mathbf{u}_R) \} - \frac{1}{2\lambda} \mathbf{Q}(\mathbf{u}_L, \mathbf{u}_R) (\mathbf{u}_R - \mathbf{u}_L). \quad (3.47)$$

Using (3.45), we obtain

$$\frac{1}{\lambda} \mathbf{Q}(\mathbf{u}_L, \mathbf{u}_R)(\mathbf{u}_R - \mathbf{u}_L) = \begin{cases} -(\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)) & \text{for } a_R \leq 0, \\ \frac{1}{a_R - a_L} \{(a_L + a_R)(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)\} - \\ \quad -2a_L a_R (\mathbf{u}_R - \mathbf{u}_L) \} & \text{for } a_L \leq 0 \leq a_R, \\ \mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) & \text{for } a_L \geq 0. \end{cases}$$

Let  $\mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)$  be a Roe linearization of the flux function  $\mathbf{f}$  so that

$$\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) = \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)(\mathbf{u}_R - \mathbf{u}_L).$$

Then the numerical viscosity matrix is given by

$$\mathbf{Q}(\mathbf{u}_L, \mathbf{u}_R) = \begin{cases} -\lambda \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R) & \text{for } a_R \leq 0, \\ \frac{\lambda}{a_R - a_L} \{(a_L + a_R) \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R) - \\ \quad -2a_L a_R \mathbf{I}\} & \text{for } a_L \leq 0 \leq a_R, \\ \lambda \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R) & \text{for } a_L \geq 0, \end{cases}$$

or in condensed form

$$\mathbf{Q}(\mathbf{u}_L, \mathbf{u}_R) = \frac{\lambda}{a_R^+ - a_L^-} \{(a_L^- + a_R^+) \mathbf{A}(\mathbf{u}_L, \mathbf{u}_R) - 2a_L^- a_R^+ \mathbf{I}\}. \quad (3.48)$$

The eigenvalues  $\mu(\mathbf{u}_L, \mathbf{u}_R)$  of the viscosity matrix  $\mathbf{Q}(\mathbf{u}_L, \mathbf{u}_R)$  are therefore related to the eigenvalues  $a^{Roe}(\mathbf{u}_L, \mathbf{u}_R)$  of the Roe matrix  $\mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)$  by

$$\mu = \frac{\lambda}{a_R^+ - a_L^-} \{(a_L^- + a_R^+) a^{Roe} - 2a_L^- a_R^+\}.$$

According to Kreiss' theory, a necessary condition for the scheme to be  $L^2$ -stable is that the eigenvalues of the viscosity matrix  $\mathbf{Q}$  are nonnegative. In fact, in analogy with the scalar case, we require the following stronger condition:

$$|\mu(\mathbf{u}_L, \mathbf{u}_R)| \geq \lambda |a^{Roe}(\mathbf{u}_L, \mathbf{u}_R)|. \quad (3.49)$$

This condition implies the TVD property of the scheme in the scalar case and should be viewed as a heuristic extension in the case of systems.

*Remark 3.4.* Note that, in the case of Roe's method, the numerical viscosity matrix is indeed  $\lambda |\mathbf{A}(\mathbf{u}_L, \mathbf{u}_R)|$  and the inequality (3.49) becomes an equality. Since we know that Roe's method fails to be entropy satisfying, the condition (3.49) does not imply that the scheme is entropy satisfying.  $\square$

### Proposition 3.2

*The TVD-type stability criterion (3.49) holds if and only if  $a_L$  and  $a_R$  are chosen in such a way that we have for any eigenvalue  $a^{Roe}(\mathbf{u}_L, \mathbf{u}_R)$  of the Roe matrix*

$$a_L^- \leq a^{Roe}(\mathbf{u}_L, \mathbf{u}_R) \leq a_R^+. \quad (3.50)$$

*Proof.* Since  $a_R^+ - a_L^-$  is  $> 0$ , we easily check that condition (3.49) reads

$$(a_L^- + a_R^+)a^{Roe} - 2a_L^- a_R^+ \geq (a_R^+ - a_L^-)|a^{Roe}|$$

or equivalently

$$a_R^+ a^{Roe-} - a_L^- a^{Roe+} \geq a_L^- a_R^+.$$

This latter inequality indeed coincides with (3.50).  $\square$

On the other hand, given a discontinuity wave whose speed  $\sigma$  is not close to  $a_L$  or  $a_R$ , we expect that the H.L.L. method will not represent accurately such a wave and will smear it strongly. For instance, in the case of the gas dynamics system, the H.L.L. method is expected to smear contact discontinuities, while it can approximate in a satisfactory way both 1- and 3-shock waves. Hence, one can also think of a scheme with two intermediate states: a third approximate velocity is then needed with the role of the contact discontinuity, and the two intermediate states are still computed so as to get consistency with the integral form of the conservation laws, as required for a Godunov-type method.

Besides its simplicity, the method applied to gas dynamics, which we will describe in Sect. 4.5.2, does not predict nonphysical states since the averaging procedure involved keeps the approximate solution in the set of physical states  $\Omega$ , which is convex. The HLLEM version introduces some antidi diffusion with the aim of improving the resolution (see Harten et al. [595], Einfeldt [457], Einfeldt et al. [458], Davis [393], and Charrier et al. [281]; see also [91, 440]).

Concerning H.L.L.-type schemes for a fluid system, another approach for the choice of the wave speeds is given by the *relaxation approximation*. The velocities are the exact velocities of the Riemann solver of a relaxation system approximating the fluid system. We will detail this approach in Sect. 8.

### 3.4 Osher's Scheme

#### 3.4.1 The Scalar Case

The Engquist–Osher's scheme in the scalar case ([461]; see G.R., Chapter 3, Example 2.5 [539]) is given by the formula

$$v_j^{n+1} = v_j - \frac{\lambda}{2}(f(v_{j+1}) - f(v_{j-1})) + \frac{\lambda}{2} \left( \int_{v_j}^{v_{j+1}} |a(\xi)| d\xi - \int_{v_{j-1}}^{v_j} |a(\xi)| d\xi \right)$$

(the index  $n$  has been omitted on the right-hand side), with  $a = f'$ .

In the *strictly convex* case ( $f'' > 0$ ), we can give a simple geometric interpretation. Define

$$\begin{aligned} f^+(u) &= f(\max(u, \bar{u})), \\ f^-(u) &= f(\min(u, \bar{u})), \end{aligned}$$

where  $\bar{u}$  is the only stagnation (or sonic) point, i.e.,  $a(\bar{u}) = f'(\bar{u}) = 0$ . Then, we have

$$\begin{aligned} f(v) &= f^+(v) + f^-(v) - f(\bar{u}), \\ |f'(v)| &= f'^+(v) + f'^-(v). \end{aligned}$$

Hence, an easy computation shows that  $g^{E.O.}$  is given by

$$g^{E.O.}(u, v) = f^+(u) + f^-(v) - f(\bar{u})$$

or

$$g^{E.O.}(u, v) = f^+(u) + f^-(v)$$

since  $g$  is defined up to an additive constant. In domains where the sign of  $f'$  is constant, it reduces to the standard first-order upwind scheme.

Let us see that Osher's scheme can be interpreted as a Godunov-type scheme (see G.R., Chapter 3, Section 4), i.e.,

$$g^{E.O.}(u_L, u_R) = f(w(0; u_L, u_R)), \quad (3.51)$$

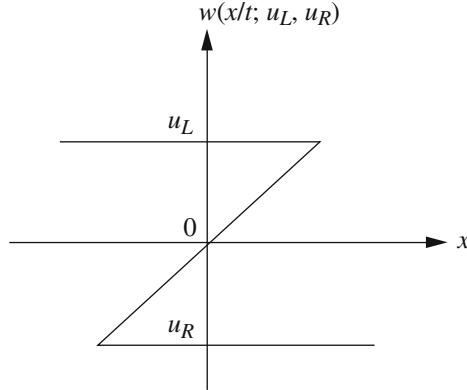
where the approximate Riemann solver  $w$  is defined as follows (we still assume  $f'' > 0$ ), setting  $\xi = \frac{x}{t}$ :

$$w(\xi; u_L, u_R) = \begin{cases} u_L, & \xi \leq a(u_L), \\ a^{-1}(\xi), & a(\min(u_L, u_R)) \leq \xi \leq a(\max(u_L, u_R)), \\ u_R, & \xi \geq a(u_R). \end{cases} \quad (3.52)$$

For  $u_L > u_R$ , the (approximate) Riemann solver replaces the shock wave solution of the Riemann problem by a multivalued function that may be viewed as a nonadmissible rarefaction wave (a “compression” wave). We have thus three branches  $w_i(\xi) = w_i(\xi; u_L, u_R)$ :

- The constant state  $w_1(\xi) = u_L$ .
- The rarefaction wave  $w_2(\xi) = a^{-1}(\xi)$ .
- The constant state  $w_3(\xi) = u_R$ .

*Example 3.1.* In the case of Burgers' equation  $a^{-1}(\xi) = \xi$ , we get the function drawn in Fig. 3.1 (see van Leer (1984)).  $\square$



**Fig. 3.1** Compression wave for Burgers' equation

Some care must be taken in defining the associated numerical flux. Let us show that we can still define  $g^{E.O.}(u_L, u_R)$  by the formula (3.51) provided that we set

$$g^{E.O.}(u_L, u_R) = \sum_i (-1)^{i-1} f(w_i(0; u_L, u_R)), \quad (3.53)$$

where the sum is taken over all branches present at 0.

First, if  $a(u_L) \geq 0 \geq a(u_R)$ , (3.52) is multivalued at point 0, and (3.53) gives

$$g^{E.O.}(u_L, u_R) = f(w_1(0)) - f(w_2(0)) + f(w_3(0)).$$

Since  $w_2(0) = \bar{u}$ , we find in that case

$$g^{E.O.}(u, v) = f(u) - f(\bar{u}) + f(v).$$

Now, if  $a(u_L) \leq 0 \leq a(u_R)$ ,

$$g^{E.O.}(u_L, u_R) = f(w(0; u_L, u_R)) = f(w_2(0)) = f(\bar{u}).$$

If 0 does not belong to the interval  $(a(u_L), a(u_R))$ , we find

$$g^{E.O.}(u_L, u_R) = f(w(0; u_L, u_R)) = \begin{cases} f(w_1(0)) & \text{if } a(u_L) \text{ and } a(u_R) > 0, \\ f(w_3(0)) & \text{if } a(u_L) \text{ and } a(u_R) < 0. \end{cases}$$

Since the flux is defined up to an additive constant, the formula is valid in all cases.

### 3.4.2 Osher's Scheme for a System

For a system, the numerical flux is defined by

$$\mathbf{g}^{E.O.}(\mathbf{u}_L, \mathbf{u}_R) = \frac{1}{2}(\mathbf{f}(\mathbf{u}_L) + \mathbf{f}(\mathbf{u}_R)) - \frac{1}{2} \int_{\Gamma} |\mathbf{A}(\mathbf{w})| d\mathbf{w}$$

or, equivalently,

$$\mathbf{g}^{E.O.}(\mathbf{u}_L, \mathbf{u}_R) = \mathbf{f}(\mathbf{u}_L) + \int_{\Gamma} \mathbf{A}^-(\mathbf{w}) d\mathbf{w} = \mathbf{f}(\mathbf{u}_R) - \int_{\Gamma} \mathbf{A}^+(\mathbf{w}) d\mathbf{w}, \quad (3.54)$$

for a suitable path of integration  $\Gamma$  connecting  $\mathbf{u}_L$  and  $\mathbf{u}_R$  (in the state space). The idea consists of choosing the path  $\Gamma$  in a “natural” way using the integral curves of the (right) eigenvectors of the matrix  $\mathbf{A}$ . This means that we connect  $\mathbf{u}_L$  to  $\mathbf{u}_R$  by a sequence of intermediate states  $\mathbf{u}_k$  as in the solution of the Riemann problem (Chap. II, Theorem 6.1), except that we replace the  $k$ -shock waves by multivalued  $k$ -rarefaction waves.

More precisely, let us recall (see Chap. II, Sect. 3) that, on the one hand, when the  $k$ th field is *genuinely nonlinear* (or equivalently in the scalar case if  $f$  is strictly convex) if  $a_k(\mathbf{u}_L) < a_k(\mathbf{u}_R)$ , and if  $\mathbf{u}_R$  belongs to the integral curve of the vector field  $\mathbf{r}_k$ , then  $\mathbf{u}_L$  and  $\mathbf{u}_R$  can be connected by a  $k$ -rarefaction wave

$$\begin{cases} \mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi)), \\ \mathbf{v}(\lambda_k(\mathbf{u}_L)) = \mathbf{u}_L, \end{cases} \quad (3.55)$$

( $a_k(\mathbf{u}_L) < a_k(\mathbf{u}_R)$  since  $a_k$  is increasing along the curve). If  $a_k(\mathbf{u}_R) < a_k(\mathbf{u}_L)$  and if  $\mathbf{u}_R$  belongs to the  $k$ -shock curve, then the states  $\mathbf{u}_L$  and  $\mathbf{u}_R$  are connected by a  $k$ -shock.

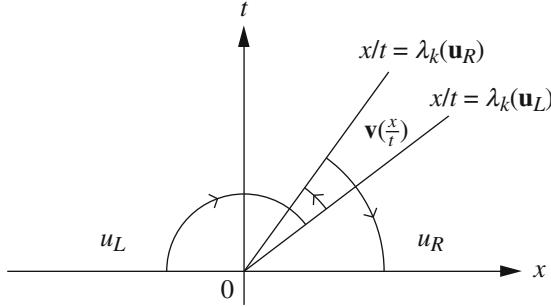
Instead of the admissible part of the  $k$ -shock curve, we shall use the “non-admissible” part of the integral curve of the vector  $\mathbf{r}_k$ , which is osculatory (it starts from  $\mathbf{u}_L$ , in the opposite direction to the rarefaction curve). This amounts to connecting  $\mathbf{u}_L$  and  $\mathbf{u}_R$  by a multivalued rarefaction (Fig. 3.2). So we define for  $a_k(\mathbf{u}_R) < a_k(\mathbf{u}_L)$  a “compression” wave that is a multivalued solution by

$$\mathbf{w}(\xi; u, v) = \begin{cases} \mathbf{u}_L, & \xi \leq a_k(\mathbf{u}_L), \\ \mathbf{v}(\xi), & a_k(\mathbf{u}_R) \leq \xi \leq a_k(\mathbf{u}_L), \\ \mathbf{u}_R, & \xi \geq a_k(\mathbf{u}_R), \end{cases}$$

where  $\mathbf{v}$  is the solution of (3.55) and  $\xi = x/t$ .

Again, for  $a_k(\mathbf{u}_R) \leq \xi \leq a_k(\mathbf{u}_L)$ , we have three states:

- The constant state  $\mathbf{w}^1(\xi) = \mathbf{u}_L$ .
- The rarefaction wave  $\mathbf{w}^2(\xi) = \mathbf{v}(\xi)$ .
- The constant state  $\mathbf{w}^3(\xi) = \mathbf{u}_R$ .



**Fig. 3.2** Compression wave in the  $(x, t)$ -plane

On the other hand, if the  $k$ th field is *linearly degenerate*, we have seen that the set of states  $\mathbf{v}$  that can be connected to a given state  $\mathbf{u}$  by a  $k$ -contact discontinuity is also an integral curve of the vector field  $\mathbf{r}_k$ .

We want now to define an approximate Riemann solver. We shall use the following family of  $k$ -waves: rarefactions (multivalued if necessary) if the  $k$ th field is genuinely nonlinear and contact discontinuities if the  $k$ th field is linearly degenerate. By an analog of Theorem 6.1 in Chap. II, we know that for sufficiently close initial states  $\mathbf{u}_L, \mathbf{u}_R$ , the Riemann problem has a multivalued solution consisting of at most  $(p + 1)$  constant states  $\mathbf{u}_k$ , such that  $\mathbf{u}_{k+1}$  is connected to  $\mathbf{u}_k$  by a  $k$ -wave of the above family, and such a solution is unique. The proof follows exactly that of Theorem 6.1 (see Dubois [431]). This Riemann solver is exact when the solution of the Riemann problem has no shocks.

If we denote by  $\theta_k$  the strength of the  $k$ th wave (corresponding to  $\varepsilon_k$  in Chap. II, Theorem 6.1, but now  $\theta_k$  can be  $\geq 0$  or  $\leq 0$ ), we can compute easily the integral in (3.54). The path  $\Gamma$  is the union of  $\Gamma_k : \Gamma = \bigcup \Gamma_k$ , each  $\Gamma_k$  being a portion of an integral curve of  $\mathbf{r}_k$ . We have thus

$$\int_{\Gamma} \mathbf{A}^-(\mathbf{w}) d\mathbf{w} = \sum_k \int_0^{\theta_k} a_k^-(\mathbf{w}_k(\xi)) \mathbf{r}_k(\mathbf{w}_k(\xi)) d\xi,$$

$$\int_{\Gamma} \mathbf{A}^+(\mathbf{w}) d\mathbf{w} = \sum_k \int_0^{\theta_k} a_k^+(\mathbf{w}_k(\xi)) \mathbf{r}_k(\mathbf{w}_k(\xi)) d\xi,$$

where  $\mathbf{w}_k$  is solution of (3.55)

$$\mathbf{w}'_k(\xi) = \mathbf{r}_k(\mathbf{w}_k(\xi)), \quad \text{for } 0 \leq \xi \leq \theta_k \text{ or } 0 \geq \xi \geq \theta_k, \quad k = 1, \dots, p,$$

and

$$\begin{cases} \mathbf{w}_{k+1}(0) = \mathbf{w}_k(\theta_k) = \mathbf{u}_k, & k = 1, \dots, p-1, \\ \mathbf{w}_1(0) = \mathbf{u}_L, \\ \mathbf{w}_p(\theta_p) = \mathbf{u}_R. \end{cases}$$

On the one hand, if the  $k$ th field is linearly degenerate, we know that  $a_k$  is a  $k$ -Riemann invariant (see also Chap. II, Theorem 4.2) and is constant along the integral curve of  $\mathbf{r}_k$ . Hence

$$\int_0^{\theta_k} a_k^+(\mathbf{w}_k(\xi)) \mathbf{r}_k(\mathbf{w}_k(\xi)) d\xi = \begin{cases} \mathbf{f}(\mathbf{w}_{k+1}(0)) - \mathbf{f}(\mathbf{w}_k(0)) & \text{if } a_k(\mathbf{w}_k) > 0, \\ 0 & \text{if } a_k(\mathbf{w}_k) \leq 0. \end{cases}$$

On the other hand, if the  $k$ th field is genuinely nonlinear,  $\xi \rightarrow a_k(\mathbf{w}_k(\xi))$  is strictly monotone. Let us denote by  $\bar{s}_k$  the “sonic” point (if it exists) such that

$$a_k(\mathbf{w}_k(\bar{s}_k)) = 0,$$

and the corresponding “sonic” state by

$$\bar{\mathbf{u}}_k = \mathbf{w}_k(\bar{s}_k).$$

Then, according to the respective signs of  $a_k(\mathbf{u}_k)$ , we obtain

$$\int_0^{\theta_k} a_k^+(\mathbf{w}_k(\xi)) \mathbf{r}_k(\mathbf{w}_k(\xi)) d\xi = \begin{cases} \mathbf{f}(\mathbf{w}_{k+1}(0)) - \mathbf{f}(\mathbf{w}_k(0)) & \text{if } a_k(\mathbf{w}_{k+1}(0)) > 0, \ a_k(\mathbf{w}_k(0)) > 0, \\ \mathbf{f}(\mathbf{w}_{k+1}(0)) - \mathbf{f}(\bar{\mathbf{u}}_k) & \text{if } a_k(\mathbf{w}_{k+1}(0)) > 0, \ a_k(\mathbf{w}_k(0)) \leq 0, \\ \mathbf{f}(\bar{\mathbf{u}}_k) - \mathbf{f}(\mathbf{w}_k(0)) & \text{if } a_k(\mathbf{w}_{k+1}(0)) \leq 0, \ a_k(\mathbf{w}_k(0)) > 0, \\ \mathbf{0} & \text{if } a_k(\mathbf{w}_{k+1}(0)) \leq 0, \ a_k(\mathbf{w}_k(0)) \leq 0. \end{cases}$$

It remains to derive the numerical flux. As in the scalar case, we sum up the different possibilities with the following formula, which can indeed be interpreted as a Godunov-type flux (see Dubois [431]):

$$\mathbf{g}^{E.O.}(\mathbf{u}_L, \mathbf{u}_R) = \sum_{k=0}^p \varepsilon_k \mathbf{f}(\mathbf{u}_k) + \sum_{k=1}^p \bar{\varepsilon}_k \mathbf{f}(\bar{\mathbf{u}}_k), \quad (3.56)$$

where the  $\varepsilon_k$  and  $\bar{\varepsilon}_k$  are defined by

$$\varepsilon_k = \begin{cases} 1 & \text{if } a_k(\mathbf{u}_k) \leq 0 < a_{k+1}(\mathbf{u}_k), \\ 0 & \text{elsewhere,} \end{cases}$$

$$\bar{\varepsilon}_k = \begin{cases} 1 & \text{if } a_k(\mathbf{u}_{k-1}) \leq 0 < a_k(\mathbf{u}_k), \\ -1 & \text{if } a_k(\mathbf{u}_k) \leq 0 < a_k(\mathbf{u}_{k-1}), \\ 0 & \text{elsewhere,} \end{cases}$$

(with the convention  $a_0 = -\infty, a_{n+1} = +\infty$ ). Thus, a rarefaction wave has a positive sign, whereas a “compression” wave has a negative sign. Note that this method can be related to that developed in Sect. 4.3 where only shock waves are considered, whereas here only simple waves and contact discontinuities are involved.

*Remark 3.5.* We might also have used the reverse order ( $\mathbf{u}_{k+1}$  is connected to  $\mathbf{u}_k$  by a  $(p-k)$ -wave). This has been proposed by Osher and Solomon [922]. The ordering of the path may be significant (see Hoff [620] for the problem of invariance or Roberts [977] and [681] for that of slowly moving shocks).  $\square$

*Remark 3.6.* Osher’s scheme and Godunov’s and Roe’s schemes as well are *flux difference splitting* methods, i.e., their flux can be written as

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v})) - \frac{1}{2}(\Delta\mathbf{f}^+(\mathbf{u}, \mathbf{v}) - \Delta\mathbf{f}^-(\mathbf{u}, \mathbf{v})),$$

where

$$\Delta\mathbf{f}(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}) = \Delta\mathbf{f}^+(\mathbf{u}, \mathbf{v}) + \Delta\mathbf{f}^-(\mathbf{u}, \mathbf{v}),$$

and  $\Delta\mathbf{f}^+$  (resp.  $\Delta\mathbf{f}^-$ ) corresponds more or less to right (resp. left) running waves, i.e., positive (resp. negative) signal speeds. For Osher’s scheme,

$$\Delta\mathbf{f}^\pm(\mathbf{u}, \mathbf{v}) = \int_\Gamma \mathbf{A}^\pm(\mathbf{w}) d\mathbf{w} = \sum_k \int_0^{\theta_k} a_k^\pm(\mathbf{w}_k(\xi)) \mathbf{r}_k(\mathbf{w}(\xi)) d\xi,$$

while for Godunov’s flux (2.9),

$$\begin{aligned} \Delta\mathbf{f}^+(\mathbf{u}, \mathbf{v}) &= \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{w}_R(0, \mathbf{u}, \mathbf{v})), \\ \Delta\mathbf{f}^-(\mathbf{u}, \mathbf{v}) &= -(\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{w}_R(0, \mathbf{u}, \mathbf{v}))), \end{aligned}$$

and for Roe’s flux (3.40),

$$\Delta\mathbf{f}^\pm(\mathbf{u}, \mathbf{v}) = \mathbf{A}^\pm(\mathbf{u}, \mathbf{v})(\mathbf{v} - \mathbf{u}) = \sum_k \alpha_k a_k^\pm \mathbf{r}_k(\mathbf{u}, \mathbf{v}).$$

Note that in Osher’s or Roe’s scheme, the upwinding is achieved via a field-by-field decomposition.  $\square$

Note that the Osher solver can be extended to nonconservative systems [449].

We shall now focus on the gas dynamics equations (this application was first treated in Osher and Solomon [922]).

### 3.4.3 Application to the Gas Dynamics System

Osher's scheme applied to the Euler equations gives the result that  $\mathbf{U}_L$  is connected to  $\mathbf{U}_R$  via two intermediate states as follows (see Chap. III, Sect. 3):

- $\mathbf{U}_L$  is connected to  $\mathbf{U}_1$  by a 1 (multivalued)-rarefaction wave.
- $\mathbf{U}_1$  to  $\mathbf{U}_2$  by a 2-contact discontinuity.
- $\mathbf{U}_2$  to  $\mathbf{U}_R$  by a 3 (multivalued)-rarefaction wave.

As in the classical Riemann problem, we begin by determining the states  $\mathbf{U}_1$  and  $\mathbf{U}_2$  on each side of the contact discontinuity. We use the  $k$ -Riemann invariants (see Chap. II, Example 3.3), which are constant on the  $k$ -wave since it is an integral curve of  $\mathbf{r}_k$ . We deduce first that  $\mathbf{U}_1$  and  $\mathbf{U}_2$  have the same velocity  $u_1 = u_2 = u^*$  and pressure  $p_1 = p_2 = p^*$ . We find  $u^*$  and  $p^*$  by using the 1- and 3-Riemann invariants  $u \pm \ell$  and  $s$ : since  $s_1^* = s_L$  and  $s_2^* - s_R$ , we obtain

$$\begin{aligned} u^* + \ell(\rho(p^*, s_L), s_L) &= u_L + \ell(\rho_L, s_L), \\ u^* - \ell(\rho(p^*, s_R), s_R) &= u_R - \ell(\rho_R, s_R). \end{aligned}$$

The densities  $\rho_1$  and  $\rho_2$  are then computed by using the equation  $\rho = \rho(p, s)$ . Once  $\mathbf{U}_1$  and  $\mathbf{U}_2$  are computed, we need to specify the states that can be present in the 1- and 3-rarefaction waves at  $\frac{x}{t} = 0$ , i.e., the “sonic” states  $\bar{\mathbf{U}}_1 = (\bar{\rho}_1, \bar{u}_1, s_L)$  and  $\bar{\mathbf{U}}_3 = (\bar{\rho}_3, \bar{u}_3, s_R)$ . These satisfy

$$a_1(\bar{\mathbf{U}}_1) = \bar{u}_1 - c(\bar{u}_1, s_L) = 0,$$

$$\bar{u}_1 + \ell(\bar{\rho}_1, s_L) = u_L + \ell(\rho_L, s_L),$$

$$a_3(\bar{\mathbf{U}}_3) = \bar{u}_3 + c(\bar{u}_3, s_R) = 0,$$

$$\bar{u}_1 - \ell(\bar{\rho}_1, s_R) = u_R - \ell(\rho_R, s_R).$$

Formula (3.56) becomes

$$\begin{aligned} \mathbf{g}^{E.O.}(\mathbf{u}_L, \mathbf{u}_R) &= \varepsilon_L \mathbf{f}(\mathbf{U}_L) + \bar{\varepsilon}_1 \mathbf{f}(\bar{\mathbf{U}}_1) + \varepsilon_1 \mathbf{f}(\mathbf{U}_1) \\ &\quad + \varepsilon_2 \mathbf{f}(\mathbf{U}_2) + \bar{\varepsilon}_3 \mathbf{f}(\bar{\mathbf{U}}_3) + \varepsilon_R \mathbf{f}(\mathbf{U}_R), \end{aligned}$$

where

$$\begin{aligned}\varepsilon_L &= \begin{cases} 1 & \text{if } a_1(\mathbf{U}_L) = u_L - c_L > 0, \\ 0 & \text{if } a_1(\mathbf{U}_L) < 0, \end{cases} \\ \bar{\varepsilon}_1 &= \begin{cases} 1 & \text{if } a_1(\mathbf{U}_L) \leq 0 < a_1(\mathbf{U}_1), \\ -1 & \text{if } a_1(\mathbf{U}_1) \leq 0 < a_1(\mathbf{U}_L), \\ 0 & \text{elsewhere,} \end{cases} \\ \varepsilon_1 &= \begin{cases} 1 & \text{if } a_1(\mathbf{U}_1) < 0 < u^*, \\ 0 & \text{if } a_1(\mathbf{U}_1) \geq 0 \text{ or } u^* \leq 0, \end{cases} \\ \varepsilon_2 &= \begin{cases} 1 & \text{if } u^* \leq 0 < a_3(\mathbf{U}_2) = u^* + c_2, \\ 0 & \text{if } a_3(\mathbf{U}_2) \leq 0 \text{ or } u^* > 0, \end{cases} \\ \bar{\varepsilon}_3 &= \begin{cases} 1 & \text{if } a_3(\mathbf{U}_2) \leq 0 < a_3(\mathbf{U}_R), \\ -1 & \text{if } a_3(\mathbf{U}_R) \leq 0 < a_3(\mathbf{U}_2), \\ 0 & \text{elsewhere,} \end{cases} \\ \varepsilon_R &= \begin{cases} 1 & \text{if } a_3(\mathbf{U}_R) \leq 0, \\ 0 & \text{if } a_3(\mathbf{U}_R) > 0. \end{cases}\end{aligned}$$

For a polytropic perfect gas (see Chap. III, Sect. 1.2), we know that

$$c^2 = \gamma \frac{p}{\rho}, \quad \ell = \frac{2c}{(\gamma - 1)}.$$

Moreover, the pressure as the function  $p = p(c, s)$  satisfies

$$p^{\frac{(\gamma-1)}{2\gamma}} = \sqrt{\gamma} \frac{c}{s},$$

which implies

$$\frac{c_1}{s_L} = \frac{c_2}{s_R}.$$

Hence the equations for  $\mathbf{U}_1$  and  $\mathbf{U}_2$  are easily solved:

$$\begin{aligned}c_1 &= \left(1 + \frac{s_R}{s_L}\right)^{-1} \left\{ c_L + c_R + \frac{(\gamma - 1)}{2}(u_L - u_R) \right\}, \\ c_2 &= \left(1 + \frac{s_L}{s_R}\right)^{-1} \left\{ c_L + c_R + \frac{(\gamma - 1)}{2}(u_L - u_R) \right\}, \\ u^* &= \left(1 + \frac{s_L}{s_R}\right)^{-1} \left\{ u_L + \frac{2c_L}{(\gamma - 1)} \right\} + \left(1 + \frac{s_R}{s_L}\right)^{-1} \left\{ u_R - \frac{2c_R}{(\gamma - 1)} \right\}.\end{aligned}$$

Also, one finds

$$\bar{u}_1 = \frac{(\gamma - 1)}{(\gamma + 1)} \left\{ u_L + \frac{2c_L}{(\gamma - 1)} \right\}, \quad \bar{u}_3 = \frac{(\gamma - 1)}{(\gamma + 1)} \left\{ u_R - \frac{2c_R}{(\gamma - 1)} \right\}.$$

Besides the smoothness of the flux, Osher's scheme does not need entropy correction and gives a good resolution of contact discontinuities (see Coquel and Liou [359]), which makes it a good candidate for the extension to Navier–Stokes equations in spite of its relative complexity. (It provides moreover a consistent treatment of boundary conditions (Koren [705])). It is also superior on the slowly moving shock problem and the carbuncle phenomenon (Lin [802], Pandolfi and D'Ambrosio [925], and references therein). For the extension to chemical and vibrational equilibrium or nonequilibrium gas flows, we refer to Dubois [431], Abgrall and Montagné [14], and Abgrall et al. [9]. For the extension of Engquist–Osher's scheme to systems that are not strictly hyperbolic, see Bell et al. [99].

## 4 Roe-Type Methods for the Gas Dynamics System

We begin by presenting the original Roe's method for the gas dynamics system in Eulerian coordinates. We will present below another derivation, starting from the Lagrangian frame.

### 4.1 Roe's Method for the Gas Dynamics Equations: (I) The Ideal Gas Case

#### 4.1.1 Computation of Roe's Matrix via Parameter Vectors

In order to apply Roe's method to the gas dynamics equations (2.19), it remains to construct a Roe-type linearization. We present the derivation in the original way (Roe [979]), only considering the case of an equation of state of the form

$$p = (\gamma - 1)\rho\varepsilon + c_{\text{ref}}^2(\rho - \rho_{\text{ref}}), \quad (4.1)$$

which generalizes the usual ideal gas law and is known as a stiffened equation of state of Grüneisen type (see Chap. III, Example 1.2). It is convenient to set  $p_\infty = c_{\text{ref}}^2\rho_{\text{ref}}/\gamma$ . However, we shall try to carry on the computations in a more general case as long as possible. In this case, we are able to find the Roe matrix in the form

$$\mathbf{A}(\mathbf{U}, \mathbf{V}) = \mathbf{A}(M(\mathbf{U}, \mathbf{V})),$$

where  $\mathbf{A}$  is the Jacobian matrix of  $\mathbf{F}$ ,  $\mathbf{A} = \mathbf{F}'$  and  $M(\mathbf{U}, \mathbf{V})$  is some average of the two states  $\mathbf{U}, \mathbf{V}$  that satisfies

$$\begin{aligned} M(\mathbf{U}, \mathbf{V}) &= M(\mathbf{V}, \mathbf{U}), \\ M(\mathbf{U}, \mathbf{U}) &= \mathbf{U}. \end{aligned}$$

The operator  $M$  will be constructed via a change of variables  $\mathbf{W} \mapsto \mathbf{U}(\mathbf{W})$ , where  $\mathbf{W}$  is the *parameter vector*, such that if we set

$$\mathbf{G}(\mathbf{W}) = \mathbf{F}(\mathbf{U}(\mathbf{W})), \quad (4.2)$$

then

$$\mathbf{U}(\mathbf{W}) \text{ and } \mathbf{G}(\mathbf{W}) \text{ are homogeneous quadratic functions of } \mathbf{W} \quad (4.3)$$

(up to an additive constant vector).

*Lemma 4.1*

Assume that there exists a change of variables  $\mathbf{W} \mapsto \mathbf{U}(\mathbf{W})$  such that (4.3) holds. The expressions

$$\mathbf{A}(\mathbf{U}_R, \mathbf{U}_L) = \mathbf{A}(\bar{\mathbf{U}}), \quad \bar{\mathbf{U}} = \mathbf{U}(\mathbf{W}^*), \quad \mathbf{W}^* = \frac{1}{2}(\mathbf{W}_R + \mathbf{W}_L) \quad (4.4)$$

define a Roe-type linearization.

*Proof.* Assume that (4.3) holds. This yields

$$\mathbf{U}_R - \mathbf{U}_L = \mathbf{U}(\mathbf{W}_R) - \mathbf{U}(\mathbf{W}_L) = \mathbf{U}'(\mathbf{W}^*)(\mathbf{W}_R - \mathbf{W}_L)$$

where

$$\mathbf{W}^* = \frac{1}{2}(\mathbf{W}_R + \mathbf{W}_L)$$

and

$$\mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) = \mathbf{G}(\mathbf{W}_R) - \mathbf{G}(\mathbf{W}_L) - \mathbf{G}'(\mathbf{W}^*)(\mathbf{W}_R - \mathbf{W}_L).$$

Thus, we get

$$\begin{aligned} \mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) &= \mathbf{F}'(\mathbf{U}(\mathbf{W}^*))\mathbf{U}'(\mathbf{W}^*)(\mathbf{W}_R - \mathbf{W}_L) \\ &= \mathbf{A}(\mathbf{U}(\mathbf{W}^*))( \mathbf{U}_R - \mathbf{U}_L ). \end{aligned}$$

Setting

$$\bar{\mathbf{U}} = \mathbf{U}(\mathbf{W}^*), \quad \mathbf{W}^* = \frac{1}{2}(\mathbf{W}_R + \mathbf{W}_L),$$

we have

$$\mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) = \mathbf{A}(\bar{\mathbf{U}})(\mathbf{U}_R - \mathbf{U}_L),$$

and

$$M(\mathbf{U}, \mathbf{V}) = \bar{\mathbf{U}} \quad (4.5)$$

is the desired “Roe averaged state.”  $\square$

In this case, Roe’s scheme can be written as

$$\begin{cases} \mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \lambda \sum_{k=1}^3 \alpha_{k,j+\frac{1}{2}}^n a_k^- (\bar{\mathbf{v}}_{j+\frac{1}{2}}^n) \mathbf{r}_k(\bar{\mathbf{v}}_{j+\frac{1}{2}}^n) \\ \quad + \alpha_{k,j-\frac{1}{2}}^n a_k^+ (\bar{\mathbf{v}}_{j-\frac{1}{2}}^n) \mathbf{r}_k(\bar{\mathbf{v}}_{j-\frac{1}{2}}^n), \end{cases} \quad (4.6)$$

where again  $a_1(\bar{\mathbf{U}}) = \bar{u} - \bar{c}$ ,  $a_2(\bar{\mathbf{U}}) = \bar{u}$ ,  $a_3(\bar{\mathbf{U}}) = \bar{u} + \bar{c}$ , and by (4.5) and (4.6)

$$\bar{\mathbf{v}}_{j+\frac{1}{2}}^n = M(\mathbf{v}_j^n, \mathbf{v}_{j+1}^n);$$

moreover, the coefficients  $\alpha_k$  are defined by (3.39), i.e.,

$$\mathbf{v}_{j+1}^n - \mathbf{v}_j^n \doteq \Delta \mathbf{v}_{j+\frac{1}{2}}^n = \sum_{k=1}^3 \alpha_{k,j+\frac{1}{2}}^n \mathbf{r}_k(\bar{\mathbf{v}}_{j+\frac{1}{2}}^n). \quad (4.7)$$

Let us check the property (4.3) in the case of an equation of state (4.1). We write (2.19) as

$$\mathbf{U} = \begin{pmatrix} \rho \\ q \\ E \end{pmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} q \\ q^2/\rho + p \\ (E+p)q/\rho \end{pmatrix},$$

where

$$q = \rho u, \quad E = \rho e = \rho \left( \varepsilon + \frac{1}{2} u^2 \right), \quad p = p(\rho, \varepsilon) = p \left( \rho, \frac{E}{\rho} - \frac{q^2}{2\rho^2} \right).$$

We introduce the total specific enthalpy

$$H = e + \frac{p}{\rho}$$

and set

$$\mathbf{W} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} \rho^{\frac{1}{2}} \\ \rho^{\frac{1}{2}} u \\ \rho^{\frac{1}{2}} H \end{pmatrix}. \quad (4.8)$$

*Lemma 4.2*

Assume an equation of state of the form (4.1). Defining the parameter vector  $\mathbf{W}$  by (4.8), the property (4.3) is satisfied. More precisely, we have

$$\begin{cases} \mathbf{U}(\mathbf{W}) = (w_1^2, w_1 w_2, w_1 w_3 - p)^T, \\ \mathbf{G}(\mathbf{W}) = (w_1 w_2, w_2^2 + p, w_2 w_3)^T, \end{cases} \quad (4.9)$$

where

$$p + p_\infty = \frac{c_{\text{ref}}^2}{\gamma} w_1^2 - \frac{\gamma - 1}{2\gamma} w_2^2 + \frac{\gamma - 1}{\gamma} w_1 w_3.$$

*Proof.* It follows immediately from (4.8) that

$$\begin{cases} \rho = w_1^2, \\ q = \rho^{\frac{1}{2}} (\rho^{\frac{1}{2}} u) = w_1 w_2, \\ E = \rho H - p = w_1 w_3 - p. \end{cases} \quad (4.10)$$

Let us now express the pressure in terms of  $\mathbf{W}$ . On the one hand,

$$\rho H = \rho^{\frac{1}{2}}(\rho^{\frac{1}{2}}H) = w_1 w_3;$$

on the other hand,

$$\begin{aligned} \rho H &= \rho\left(e + \frac{p}{\rho}\right) = E + p = \rho\left(\varepsilon + \frac{u^2}{2}\right) + p \\ &= \gamma(\gamma - 1)^{-1}p - (\gamma - 1)^{-1}c_{\text{ref}}^2(\rho - \rho_{\text{ref}}) + \frac{\rho u^2}{2}. \end{aligned}$$

With the notation  $p_\infty$ , we get

$$\begin{aligned} \rho H &= \gamma(\gamma - 1)^{-1}(p + p_\infty) - (\gamma - 1)^{-1}c_{\text{ref}}^2 \rho + \frac{\rho u^2}{2} \\ &= \gamma(\gamma - 1)^{-1}(p + p_\infty) - (\gamma - 1)^{-1}c_{\text{ref}}^2 w_1^2 + \frac{w_2^2}{2}. \end{aligned}$$

Equating the two expressions for  $\rho H$ , we obtain

$$p + p_\infty = \frac{c_{\text{ref}}^2}{\gamma}w_1^2 - \frac{\gamma - 1}{2\gamma}w_2^2 + \frac{\gamma - 1}{\gamma}w_1 w_3. \quad (4.11)$$

Thus, we have

$$E = \frac{w_1 w_3}{\gamma} + \frac{\gamma - 1}{2\gamma}w_2^2 - \frac{c_{\text{ref}}^2}{\gamma}w_1^2 + p_\infty.$$

Note that for a polytropic ideal gas,  $p_\infty = 0 = c_{\text{ref}}$ , and the expression for  $E$  simplifies into

$$E = \frac{1}{\gamma}w_1 w_3 + \frac{\gamma - 1}{2\gamma}w_2^2.$$

Consider next  $\mathbf{F}(\mathbf{U})$ ; we have already

$$q = w_1 w_2;$$

then

$$\begin{aligned} \frac{q^2}{\rho} + p &= \rho u^2 + p \\ &= \frac{\gamma - 1}{\gamma}w_1 w_3 + \frac{\gamma + 1}{2\gamma}w_2^2 + \frac{c_{\text{ref}}^2}{\gamma}w_1^2 - p_\infty \end{aligned}$$

and

$$(E + p)\frac{q}{\rho} = (\rho e + p)u = \rho H u = w_2 w_3,$$

which ends the proof.  $\square$

Expressions (4.9) show that  $\mathbf{U}$  is indeed a homogeneous quadratic function of  $\mathbf{W}$  up to an additive constant vector, so that we may apply Lemma 4.1. It remains to compute the matrix  $\mathbf{A}(\bar{\mathbf{U}})$ . We give first in the following lemma the expression for the sound speed; this indeed can be done for a more general equation of state (see Chap. III, Sect. 1.2, (1.20)).

*Lemma 4.3*

Assume an equation of state of the form

$$p = p(\rho, \tilde{\varepsilon}), \quad (4.12)$$

where

$$\tilde{\varepsilon} = \rho\varepsilon = E - \frac{\rho u^2}{2}.$$

Defining the specific enthalpy  $h$  by

$$h = \varepsilon + \frac{p}{\rho}$$

and setting

$$\kappa = p_{\tilde{\varepsilon}} = \frac{\partial p(\rho, \tilde{\varepsilon})}{\partial \tilde{\varepsilon}}, \quad \chi = p_{\rho} = \frac{\partial p(\rho, \tilde{\varepsilon})}{\partial \rho}, \quad (4.13)$$

the sound velocity  $c$  satisfies the identity

$$c^2 = \kappa h + \chi. \quad (4.14)$$

*Proof.* The sound speed  $c$  is defined (see Chap. III, (1.4)) by

$$c^2 = \tau^2 \left( -\frac{\partial p(\tau, s)}{\partial \tau} \right) = \frac{\partial p(\rho, s)}{\partial \rho},$$

and hence by (4.12)

$$c^2 = \frac{\partial p(\rho, s)}{\partial \rho} = \frac{\partial p(\rho, \tilde{\varepsilon})}{\partial \tilde{\varepsilon}} \frac{\partial \tilde{\varepsilon}(\rho, s)}{\partial \rho} + \frac{\partial p(\rho, \tilde{\varepsilon})}{\partial \rho}.$$

Now, the second law of thermodynamics,

$$d\varepsilon = T ds - pd\tau,$$

implies

$$d\tilde{\varepsilon} = \rho d\varepsilon + \varepsilon d\rho = \rho T ds + \left( \varepsilon + \frac{p}{\rho} \right) d\rho.$$

Introducing the specific enthalpy  $h$ , we have

$$\frac{\partial \tilde{\varepsilon}(\rho, s)}{\partial \rho} = h,$$

and with shorthand notations, we get

$$c^2 = p_{\bar{\varepsilon}} h + p_{\rho} = \kappa h + \chi,$$

which is (4.14).  $\square$

Note that in the particular case of an equation of state of the form (4.1), we have

$$\begin{cases} \kappa = \gamma - 1, \\ \chi = c_{\text{ref}}^2, \end{cases} \quad (4.15)$$

and the expression (4.13) for  $c^2$  gives

$$c^2 = (\gamma - 1) \left( H - \frac{u^2}{2} \right) + c_{\text{ref}}^2,$$

since  $h$  is related to the total specific enthalpy  $H$  by

$$H = h + \frac{u^2}{2}.$$

*Lemma 4.4*

Assume the hypothesis of Lemma 4.3. The Jacobian matrix  $\mathbf{A}$  of system (2.19) is given by

$$\mathbf{A}(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 \\ K - u^2 & (2 - \kappa)u & \kappa \\ u(K - H) & H - \kappa u^2 & (1 + \kappa)u \end{pmatrix}, \quad (4.16)$$

where we have set

$$K = p_{\rho} + p_{\bar{\varepsilon}} \frac{u^2}{2} = \chi + \kappa \frac{u^2}{2}, \quad (4.17)$$

and the eigenvectors may be chosen as

$$\begin{cases} \mathbf{r}_1(\mathbf{U}) = (1, u - c, H - uc)^T, \\ \mathbf{r}_2(\mathbf{U}) = \left( 1, u, H - \frac{c^2}{\kappa} \right)^T, \\ \mathbf{r}_3(\mathbf{U}) = (1, u + c, H + uc)^T. \end{cases} \quad (4.18)$$

*Proof.* We have already observed (Chap. II, Sect. 2) that we can work with a nonconservative form of the system. Here, it is convenient to use

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0, \\ \frac{\partial q}{\partial t} + \frac{\partial(\rho u^2)}{\partial x} + \frac{\partial p}{\partial x} = 0, \\ \frac{\partial E}{\partial t} + u \frac{\partial E}{\partial x} + u \frac{\partial p}{\partial x} + (E + p) \frac{\partial u}{\partial x} = 0. \end{cases} \quad (4.19)$$

We have

$$\frac{\partial p(\rho, \tilde{\varepsilon})}{\partial x} = p_\rho \frac{\partial \rho}{\partial x} + p_{\tilde{\varepsilon}} \frac{\partial \tilde{\varepsilon}}{\partial x}.$$

First

$$\frac{\partial \tilde{\varepsilon}}{\partial x} = \frac{\partial E}{\partial x} - \frac{1}{2} \frac{\partial}{\partial x} (\rho u^2),$$

and an easy computation gives

$$\frac{\partial}{\partial x} \left( \frac{\rho u^2}{2} \right) = \rho \frac{\partial}{\partial x} \left( \frac{u^2}{2} \right) + \frac{u^2}{2} \frac{\partial \rho}{\partial x} = u \frac{\partial q}{\partial x} - \frac{u^2}{2} \frac{\partial \rho}{\partial x}.$$

Thus,

$$\frac{\partial p}{\partial x} = p_{\tilde{\varepsilon}} \frac{\partial E}{\partial x} - p_{\tilde{\varepsilon}} u \frac{\partial q}{\partial x} + \left( p_\rho + \frac{p_{\tilde{\varepsilon}} u^2}{2} \right) \frac{\partial \rho}{\partial x}$$

and

$$\frac{\partial(\rho u^2)}{\partial x} + \frac{\partial p}{\partial x} = \left( p_\rho + \frac{p_{\tilde{\varepsilon}} u^2}{2} - u^2 \right) \frac{\partial \rho}{\partial x} + (2 - p_{\tilde{\varepsilon}}) u \frac{\partial q}{\partial x} + p_{\tilde{\varepsilon}} \frac{\partial E}{\partial x}.$$

Now,

$$(E + p) \frac{\partial u}{\partial x} = \rho H \frac{\partial u}{\partial x} = H \left( \frac{\partial q}{\partial x} - u \frac{\partial \rho}{\partial x} \right),$$

so that the last equation gives

$$\frac{\partial}{\partial x} ((E + p) u) = u \left( -H + p_\rho + \frac{p_{\tilde{\varepsilon}} u^2}{2} \right) \frac{\partial \rho}{\partial x} + (H - p_{\tilde{\varepsilon}} u^2) \frac{\partial q}{\partial x} + (p_{\tilde{\varepsilon}} + 1) u \frac{\partial E}{\partial x}.$$

Defining  $K$  by (4.17) or, equivalently,

$$K = c^2 - \kappa \left( h - \frac{u^2}{2} \right) = c^2 - \kappa(H - u^2),$$

we get the desired expression for  $\mathbf{A}$ . Let us note that

$$K = \frac{\partial p(\rho, \tilde{\varepsilon})}{\partial \rho} + \frac{\partial p(\rho, \tilde{\varepsilon})}{\partial \tilde{\varepsilon}} \left( \frac{\partial \tilde{\varepsilon}}{\partial \rho} \right)_{|q,E} = \left( \frac{\partial p}{\partial \rho} \right)_{|q,E}$$

since  $\tilde{\varepsilon} = E - \frac{q^2}{2\rho}$  implies  $\left( \frac{\partial \tilde{\varepsilon}}{\partial \rho} \right)_{|q,E} = \frac{u^2}{2}$ .

Using (4.14), it is then easy to check that the eigenvectors may be chosen as in (4.18).  $\square$

In the case of Eq. (4.1), the Jacobian matrix (4.16) gives

$$\mathbf{A}(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 \\ (\gamma - 3)\frac{u^2}{2} + c_{\text{ref}}^2 & -(\gamma - 3)u & (\gamma - 1) \\ u(-H + c_{\text{ref}}^2 + (\gamma - 1)\frac{u^2}{2}) & H - (\gamma - 1)u^2 & \gamma u \end{pmatrix}. \quad (4.20)$$

and in the particular case of a polytropic ideal gas, we set  $c_{\text{ref}}^2 = 0$ . From (4.18) and (4.20), we notice that for the equation of state (4.1), we need only to compute the values  $\bar{u}, \bar{H}, \bar{c}$  in order to determine  $\mathbf{A}(\bar{\mathbf{U}})$  and the eigenvectors  $\mathbf{r}_k(\bar{\mathbf{U}})$ , since  $\chi$  and  $\kappa$  are constant. In particular, for a polytropic ideal gas,  $p = (\gamma - 1)\rho\varepsilon$  implies a simplification of the last component of  $\mathbf{r}_2$ , which can be written as  $\mathbf{r}_2 = (1, u, \frac{1}{2}|u|^2)^T$ .

*Lemma 4.5*

Assume an equation of state of the form (4.1). The velocity, total specific enthalpy, and sound speed at Roe averaged state  $\bar{\mathbf{U}}$  are given (using the notations (4.15)) by

$$\begin{cases} \bar{u} = \frac{\sqrt{\rho_L}u_L + \sqrt{\rho_R}u_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \\ \bar{H} = \frac{\sqrt{\rho_L}H_L + \sqrt{\rho_R}H_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \\ \bar{c}^2 = \kappa \left( \bar{H} - \frac{\bar{u}^2}{2} \right) + \chi. \end{cases} \quad (4.21)$$

*Proof.* First, since the function  $\mathbf{U}(\mathbf{W})$  satisfies

$$u = \frac{q}{\rho} = \frac{w_2}{w_1}, \quad H = \frac{w_3}{w_1},$$

we have

$$\bar{u} = \frac{w_2^*}{w_1^*} \quad \text{and} \quad \bar{H} = \frac{w_3^*}{w_1^*},$$

where  $\mathbf{W}^* = (w_1^*, w_2^*, w_3^*)^T$ . By (4.8) and the definition of  $\mathbf{W}^*$  (see (4.4)), we get the two first formulas in (4.21). Using (4.14) once more, we finally obtain  $\bar{c}$  in terms of the velocity  $\bar{u}$  and the enthalpy  $\bar{H}$ .  $\square$

*Remark 4.1.* The velocity and enthalpy of the Roe averaged state appear as a convex combination,

$$\bar{u} = \theta u_L + (1 - \theta)u_R, \quad \bar{H} = \theta H_L + (1 - \theta)H_R,$$

where  $\theta = \frac{\sqrt{\rho_L}}{\sqrt{\rho_L} + \sqrt{\rho_R}}$ .

The computations can also be carried out by taking a priori the velocity to be a linear combination of  $u_L$  and  $u_R$  (see Vinokur [1172]).  $\square$

### 4.1.2 Determination of the Coefficients in Roe's Scheme

In order to complete the construction of Roe's scheme, it remains to compute according to formulas (4.6) and (4.7) the coefficients  $\alpha_k$  of  $\Delta\mathbf{U}$  in the basis  $(\mathbf{r}_k(\bar{\mathbf{U}}))$  of eigenvectors of  $\mathbf{A}(\bar{\mathbf{U}})$ ,

$$\mathbf{U}_R - \mathbf{U}_L = \Delta\mathbf{U} = \sum_{k=1}^3 \alpha_k \mathbf{r}_k(\bar{\mathbf{U}}), \quad (4.22)$$

where  $\bar{\mathbf{U}}$  is Roe's average state defined in (4.4).

It will be convenient to prove first the following simple algebraic results, which are obtained in a straightforward way. To make the formulas simpler, we use the shorthand notation  $a$  for a pair  $(a_L, a_R)$ .

*Lemma 4.6*

Define for a given pair  $\rho = (\rho_L, \rho_R)$  the following averaging operators by

$$m_\rho(a) = \frac{\sqrt{\rho_L}a_L + \sqrt{\rho_R}a_R}{\sqrt{\rho_L} + \sqrt{\rho_R}} \quad (4.23)$$

and

$$\hat{m}_\rho(a) = \frac{\sqrt{\rho_R}a_L + \sqrt{\rho_L}a_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}. \quad (4.24)$$

We have

$$\hat{m}_\rho(\rho) = (\rho_L\rho_R)^{\frac{1}{2}}, \quad (4.25)$$

and for any  $a = (a_L, a_R), b = (b_L, b_R)$ , we have

$$\Delta(ab) = \hat{m}_\rho(a)\Delta b + m_\rho(b)\Delta a \quad (4.26)$$

$$\hat{m}_\rho(\rho a) = \hat{m}_\rho(\rho)m_\rho(a). \quad (4.27)$$

Here, we use the obvious notation

$$\Delta a = a_R - a_L.$$

In the following, we drop the dependence of  $m_\rho$  on  $\rho$  since no ambiguity arises.

Let us note that, by Lemma 4.5, Roe averaged state satisfies

$$\bar{u} = m(u), \quad \bar{H} = m(H). \quad (4.28)$$

*Lemma 4.7*

Assume an equation of state of the form (4.1). The coefficients  $\alpha_k$  in formula (4.21) are given (with the notations (4.24)) by

$$\begin{cases} \alpha_1 = \frac{\Delta p - \bar{c}\hat{m}(\rho)\Delta u}{2\bar{c}^2}, \\ \alpha_2 = \Delta\rho - \frac{\Delta p}{\bar{c}^2}, \\ \alpha_3 = \frac{\Delta p + \bar{c}\hat{m}(\rho)\Delta u}{2\bar{c}^2}. \end{cases} \quad (4.29)$$

*Proof.* Using the expressions (4.18) for the eigenvectors, we obtain a system of three equations:

$$\begin{cases} \Delta\rho = \alpha_1 + \alpha_2 + \alpha_3, \\ \Delta q = \alpha_1(\bar{u} - \bar{c}) + \alpha_2\bar{u} + \alpha_3(\bar{u} + \bar{c}), \\ \Delta E = \alpha_1(\bar{H} - \bar{u}\bar{c}) + \alpha_2\left(\bar{H} - \frac{\bar{c}^2}{\kappa}\right) + \alpha_3(\bar{H} + \bar{u}\bar{c}). \end{cases} \quad (4.30)$$

By (4.26),

$$\Delta q - \bar{u}\Delta\rho = \hat{m}(\rho)\Delta u,$$

and hence the system (4.30) is equivalently written as

$$\begin{aligned} \alpha_1 + \alpha_3 &= \Delta\rho - \alpha_2, \\ \alpha_3 - \alpha_1 &= \hat{m}(\rho)\frac{\Delta u}{\bar{c}}, \\ \frac{\bar{c}^2}{\kappa}\alpha_2 &= -\Delta E + \bar{H}\Delta\rho + \bar{u}\hat{m}(\rho)\Delta u. \end{aligned}$$

The third equation gives  $\alpha_2$ , from which we deduce  $\alpha_1$  and  $\alpha_3$ . Lemma 4.6 enables us to derive a simple expression for  $\alpha_2$ . Indeed, we have

$$\begin{aligned} -\Delta E + \bar{H}\Delta\rho + \bar{u}(\Delta q - \bar{u}\Delta\rho) &= -\Delta\tilde{\varepsilon} - \Delta\left(\frac{\rho u^2}{2}\right) + \bar{H}\Delta\rho \\ &\quad + \bar{u}(\Delta(\rho u) - \bar{u}\Delta\rho) = -\Delta\tilde{\varepsilon} + \bar{H}\Delta\rho - \frac{1}{2}\bar{u}^2\Delta\rho = -\Delta\tilde{\varepsilon} + \bar{h}\Delta\rho, \end{aligned}$$

and hence we obtain by (4.11)

$$\bar{c}^2\alpha_2 = \kappa(-\Delta\tilde{\varepsilon} + \bar{h}\Delta\rho) = -\Delta p + (\chi + \kappa\bar{h})\Delta\rho$$

and, by (4.21),

$$\alpha_2 = \Delta\rho - \frac{\Delta p}{\bar{c}^2}.$$

Then, we get

$$\alpha_1 = \frac{\Delta p - \bar{c}\hat{m}(\rho)\Delta u}{2\bar{c}^2},$$

$$\alpha_3 = \frac{\Delta p + \bar{c}\hat{m}(\rho)\Delta u}{2\bar{c}^2}.$$

Note that in the above proof, we only use the “incremental” form of (4.11),

$$\Delta p = \chi\Delta\rho + \kappa\Delta\varepsilon.$$

This will be used later (see formula (4.36)).  $\square$

Gathering the previous lemmas together, we have thus proved the following theorem.

#### Theorem 4.1

Assume an equation of state of the form (4.12). There exists a Roe-type linearization of the form  $\mathbf{A}(\mathbf{U}_R, \mathbf{U}_L) = \mathbf{A}(\bar{\mathbf{U}})$ , where the velocity, total specific enthalpy, and sound velocity at Roe averaged state  $\bar{\mathbf{U}}$  are given by formulas (4.21). Moreover, this is the unique linearization that is the Jacobian matrix evaluated at some average state  $\bar{\mathbf{U}}$ . Finally, Roe’s scheme is defined by (4.6), where the coefficients  $\alpha_k$  are given by (4.29).

Uniqueness will follow from the following remark together with the computations in the next section, where we consider the case of a general equation of state.

*Remark 4.2.* The expressions (4.29) can be interpreted in terms of small fluctuations (Roe and Pike, [990], Roe, [981]), which gives another algebraic way to derive the Roe averaged state. Indeed, when one computes the coefficients of  $\mathbf{U}_R - \mathbf{U}_L = \Delta\mathbf{U}$ , where  $\mathbf{U}_L$  and  $\mathbf{U}_R$  are close to some given reference state  $\bar{\mathbf{U}}$ , on the basis of eigenvectors  $\mathbf{r}_k(\bar{\mathbf{U}})$ ,

$$\mathbf{U}_R - \mathbf{U}_L = \Delta\mathbf{U} = \sum_k \alpha_k \mathbf{r}_k(\bar{\mathbf{U}}),$$

one finds within  $\mathcal{O}(\Delta^2)$  the analog of (4.29)

$$\begin{cases} \alpha_1 = \frac{\Delta p - \bar{c}\Delta u}{2\bar{c}^2}, \\ \alpha_2 = \Delta\rho - \frac{\Delta p}{\bar{c}^2}, \\ \alpha_3 = \frac{\Delta p + \bar{c}\Delta u}{2\bar{c}^2}, \end{cases}$$

with the same computations as in Lemma 4.7. One also checks that with the same coefficients  $\alpha_k$ , one has

$$\Delta\mathbf{F}(\mathbf{U}) = \mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) = \sum_k a_k(\bar{\mathbf{U}}) \alpha_k \mathbf{r}_k(\bar{\mathbf{U}}).$$

The average values  $(\bar{\rho}, \bar{u}, \bar{c})$  are then determined in order that the two jump relations

$$\Delta \mathbf{U} = \sum_k \alpha_k \mathbf{r}_k(\bar{\mathbf{U}}), \quad \Delta \mathbf{F}(\mathbf{U}) = \sum_k a_k(\bar{\mathbf{U}}) \alpha_k \mathbf{r}_k(\bar{\mathbf{U}})$$

are identically satisfied whether or not the states  $\mathbf{U}_L$  and  $\mathbf{U}_R$  are close. In fact, we shall detail the computations in the next section. Let us just say that, though the system seems overdetermined (six equations for three unknowns), it turns out that one equation is trivially satisfied, two others coincide, and one is implied by the others, so that we are left with the only three equations corresponding to  $\Delta \mathbf{F}$ . We obtain in particular for  $\bar{u}$  a quadratic equation; one of the roots coincides as expected with (4.21), and the other is

$$u = \frac{\sqrt{\rho_L} u_L - \sqrt{\rho_R} u_R}{\sqrt{\rho_L} - \sqrt{\rho_R}},$$

which is not satisfactory. Then  $\bar{\rho} = (\rho_L \rho_R)^{\frac{1}{2}}$ , and  $\bar{c}$  or  $\bar{H}$  is given as above. It must be emphasized that  $\bar{\rho}$  does not correspond to the density of the Roe averaged state (see Remark 4.4 below)

Note lastly that if one imposes only the jump relation

$$\Delta \mathbf{F}(\mathbf{U}) = \mathbf{A}(\bar{\mathbf{U}}) \Delta \mathbf{U},$$

the first equation is trivially satisfied whatever the value of  $\bar{\mathbf{U}}$ , and one finds again the same expressions for  $\bar{u}$  and  $\bar{c}$ .  $\square$

*Remark 4.3.* Roe's scheme applied to gas dynamics in Lagrangian coordinates is studied in Munz [883]. The relation between Roe matrices in Lagrangian and Eulerian coordinates is studied in Gallice [496]. It results from the general correspondence established for the change of frame between the Eulerian and Lagrangian description.  $\square$

## 4.2 Roe's Method for the Gas Dynamics Equations: (II) The “Real Gas” Case

For an arbitrary gas, it is usual to consider the pressure as a function of density and specific internal energy  $\varepsilon$ ,  $p = p(\rho, \varepsilon)$ . However, in the extension of Roe's scheme to an equilibrium “real gas”, it is more convenient to express the equation of state in the form (4.12) already used in Lemma 4.3,

$$p = p(\rho, \hat{\varepsilon}).$$

For instance, for a thermally (not necessarily calorically) perfect gas, one has

$$p = \rho R T(\varepsilon).$$

We have seen, moreover, that  $p$  satisfies the identity (see Chap. III, Sect. 1.2, (1.21))

$$p = \rho p_\rho + \tilde{\varepsilon} p_{\tilde{\varepsilon}}, \quad (4.31)$$

which generalizes (4.1) in that  $p_\rho$  and  $p_{\tilde{\varepsilon}}$  are no longer constant. In the general case, the previous method of construction of a Roe-type linearization is no longer available since there does not exist in general a parameter vector  $\mathbf{W}$  such that both  $\mathbf{U}$  and  $\mathbf{F}(\mathbf{U})$  are homogeneous quadratic functions of  $\mathbf{W}$ . For instance, if we define  $\mathbf{W}$  by (4.8), only the first two components of  $\mathbf{U}$  and the first and last components of  $\mathbf{F}(\mathbf{U})$  satisfy the requirement. A more direct approach must be used. Nonetheless, we still look for average quantities  $(\bar{u}, \bar{H}, \bar{\chi}, \bar{\kappa})$  (but not necessarily an average state  $\bar{\mathbf{U}}$ ) such that

$$\mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) = \bar{\mathbf{A}}(\mathbf{U}_R - \mathbf{U}_L). \quad (4.32)$$

Here  $\bar{\mathbf{A}}$  is the matrix (4.16) computed for the values  $(\bar{u}, \bar{H}, \bar{\chi}, \bar{\kappa})$ , i.e.,

$$\bar{\mathbf{A}} = \mathbf{A}(\bar{u}, \bar{H}, \bar{\chi}, \bar{\kappa}), \quad (4.33)$$

with obvious notations since in the expression (4.16) of  $\mathbf{A}(\mathbf{U})$ , only the variables that appear explicitly, i.e.,  $(u, H, \chi, \kappa)$ , are needed. The identity (4.32) gives the system

$$\left\{ \begin{array}{l} \Delta\rho u = \Delta\rho u, \\ \Delta(\rho u^2) + \Delta p = \left( \bar{\chi} + \frac{1}{2}\bar{\kappa}\bar{u}^2 - \bar{u}^2 \right) \Delta\rho + (2 - \bar{\kappa})\bar{u}\Delta(\rho u) \\ \qquad + \bar{\kappa}\Delta\left(\tilde{\varepsilon} + \frac{\rho u^2}{2}\right), \\ \Delta(Hq) = \bar{u}\left(\bar{\chi} + \frac{1}{2}\bar{\kappa}\bar{u}^2 - \bar{H}\right)\Delta\rho + (\bar{H} - \bar{\kappa}\bar{u}^2)\Delta(\rho u) \\ \qquad + (1 + \bar{\kappa})\bar{u}\Delta\left(\tilde{\varepsilon} + \frac{\rho u^2}{2}\right). \end{array} \right. \quad (4.34)$$

The first equation is trivial. Let us look at the second equation, which reads

$$\Delta p - (\bar{\chi}\Delta\rho + \bar{\kappa}\Delta\tilde{\varepsilon}) = \left( \frac{\bar{\kappa}}{2} - 1 \right) \{ \Delta(\rho u^2) + \bar{u}^2\Delta\rho - 2\bar{u}\Delta(\rho u) \}. \quad (4.35)$$

A natural choice for the average values  $\bar{u}, \bar{\chi}$ , and  $\bar{\kappa}$  is such that both members in (4.35) vanish (which holds for a gas with equation of state of Grüneisen type (4.1))

$$\Delta p = \bar{\chi}\Delta\rho + \bar{\kappa}\Delta\tilde{\varepsilon}, \quad (4.36)$$

$$\Delta(\rho u^2) + \bar{u}^2\Delta\rho = 2\bar{u}\Delta(\rho u). \quad (4.37)$$

Assume that we can define  $\bar{\chi}$  and  $\bar{\kappa}$  such that (4.36) holds. Then, using (4.25), an easy computation shows that (4.37), which is quadratic in  $\bar{u}$ , is satisfied if we take for  $\bar{u}$  the average obtained in the preceding section, i.e., (4.21) or

(4.28):

$$\bar{u} = m(u), \quad u = (u_L, u_R),$$

where the averaging operator  $m$  is defined by (4.23).

Let us now consider the third equation in (4.34). Assuming (4.36), we notice that

$$\bar{u}\bar{\chi}\Delta\rho + \bar{\kappa}\bar{u}\Delta\tilde{\varepsilon} + \bar{u}\Delta\tilde{\varepsilon} = \bar{u}\Delta p + \bar{u}\Delta\tilde{\varepsilon} = \bar{u}\Delta(p + \tilde{\varepsilon}) = \bar{u}\Delta\left(\rho H - \rho\frac{u^2}{2}\right).$$

There remains

$$\begin{aligned} \Delta(Hq) &= \bar{u}\left(\bar{\kappa}\frac{\bar{u}^2}{2} - \bar{H}\right)\Delta\rho + (\bar{H} - \bar{\kappa}\bar{u}^2)\Delta(\rho u) + (1 + \bar{\kappa})\bar{u}\Delta\left(\rho\frac{u^2}{2}\right) \\ &\quad + \bar{u}\Delta\left(\rho H - \rho\frac{u^2}{2}\right) \\ &= \bar{u}\bar{\kappa}\left\{\frac{\bar{u}^2}{2}\Delta\rho - \bar{u}\Delta(\rho u) + \Delta\left(\rho\frac{u^2}{2}\right)\right\} - \bar{u}\bar{H}\Delta\rho + \bar{H}\Delta(\rho u) + \bar{u}\Delta(\rho H). \end{aligned}$$

Together with (4.37), we get

$$\Delta(H\rho u) = -\bar{u}\bar{H}\Delta\rho + \bar{H}\Delta(\rho u) + \bar{u}\Delta(\rho H),$$

or using once more (4.26),

$$\Delta(H\rho u) = \bar{H}\hat{m}(\rho)\Delta u + \bar{u}\Delta(\rho H),$$

which is indeed satisfied if  $\bar{H}$  is defined by (4.20) and (4.28), i.e.,

$$\bar{H} = m(H), \quad H = (H_L, H_R).$$

In order to define  $\bar{c}$ , we also need the specific enthalpy. From

$$H = h + \frac{u^2}{2},$$

we get

$$\bar{h} = \bar{H} - \frac{\bar{u}^2}{2} = m\left(h + \frac{u^2}{2}\right) - \frac{1}{2}(m(u))^2$$

and thus

$$\bar{h} = m(h) + (\rho_L\rho_R)^{\frac{1}{2}}\frac{(\Delta u)^2}{(\sqrt{\rho_L} + \sqrt{\rho_R})^2}. \quad (4.38)$$

The sound speed is then defined by (4.14),

$$\bar{c}^2 = \bar{\chi} + \bar{\kappa}\bar{h}. \quad (4.39)$$

The remaining steps for computing Roe's scheme are unchanged, only replacing in (4.30)  $\kappa$ , which is no longer constant, by  $\bar{\kappa}$ . Indeed, if (4.36) and (4.39) hold, the computations of Lemma 4.7 are still valid, in particular

$$\alpha_2 = \Delta\rho - \frac{\Delta p}{\bar{c}^2}.$$

In short, (4.32) holds with the same averages for  $\bar{u}$  and  $\bar{H}$  as were found in the case of a polytropic ideal gas (or more generally for an equation of state (4.1)) provided we choose average values of  $\bar{\kappa}$  and  $\bar{\chi}$  such that (4.36) is satisfied.

*Theorem 4.2*

Assume an equation of state of the form (4.12). There exists a Roe-type linearization  $\mathbf{A}(\mathbf{U}_R, \mathbf{U}_L) = \bar{\mathbf{A}} = \mathbf{A}(\bar{u}, \bar{H}, \bar{\chi}, \bar{\kappa})$ , where the quantities  $(\bar{u}, \bar{H})$  are given by formulas (4.21) provided we define  $(\bar{\chi}, \bar{\kappa})$  such that (4.36) is satisfied. Finally, Roe's scheme is defined by (4.6) where the coefficients  $\alpha_k$  are given by (4.29).

For the equation of state (4.1),  $\chi$  and  $\kappa$  are constant so that  $\bar{\chi} = c_{\text{ref}}^2$  and  $\bar{\kappa} = \gamma - 1$ . For an arbitrary equilibrium gas,  $\bar{\kappa}$  and  $\bar{\chi}$  are not uniquely determined by (4.36). Arguments for a precise definition of  $\bar{\chi}$  and  $\bar{\kappa}$  in terms of the thermodynamic states only are developed in Vinokur and Montagné [1173] (see also Glaister [523, 524]). In the case of a mixture of thermally perfect nonreacting gases, and for the extension to chemical equilibrium mixtures and then to chemical and vibrational nonequilibrium mixtures, we refer to Abgrall (1989–1990) [2] and [7], Liu and Vinokur [829], Dubroca and Morreeuw [443], and Fernandez and Larroutuou [469] (see also Saurel et al. [1012] for related computations) [736].

*Remark 4.4.* Simple identities are obtained by defining  $\bar{\rho}$  formally by

$$\bar{\rho} = \hat{m}(\rho) = (\rho_L \rho_R)^{\frac{1}{2}}.$$

Indeed, we have, for instance,

$$\begin{aligned}\Delta(\rho u) &= \bar{\rho} \Delta u + \bar{u} \Delta \rho, \\ \Delta(\rho u^2) &= 2\bar{\rho} \bar{u} \Delta u = \bar{u}^2 \Delta \rho, \\ \Delta(\rho H) &= \bar{\rho} \Delta H + \bar{H} \Delta \rho, \\ \Delta(H \rho u) &= \bar{H} \bar{\rho} \Delta u + \bar{u} \Delta(H \rho),\end{aligned}$$

and so on. However, in the case of a perfect gas, by (4.8),  $\rho = w_1^2$  implies that the density of the average state  $\bar{\mathbf{U}}$  is

$$\bar{\rho} = w_1^{*2} = \left( \frac{1}{2}(w_{1L} + w_{1R}) \right)^2 = \left( \frac{1}{2}(\sqrt{\rho_L} + \sqrt{\rho_R}) \right)^2,$$

which is not equal to  $\sqrt{\rho_L \rho_R}$ . □

*Remark 4.5.* Let us show that if  $\mathbf{U}_L$  and  $\mathbf{U}_R$  are connected by a single shock, (4.28), (4.38), and (4.39) are compatible with the Rankine–Hugoniot condition, i.e., are satisfied at the shock. We first note that (4.36) implies

$$\frac{\Delta p}{\Delta \rho} = \bar{\chi} + \bar{\kappa} \frac{\Delta \tilde{\varepsilon}}{\Delta \rho}.$$

Thus, if we assume

$$\bar{h} = \frac{\Delta \tilde{\varepsilon}}{\Delta \rho}, \quad (4.40)$$

we have

$$(\bar{c}^2) = \bar{\chi} + \bar{\kappa} \bar{h} = \frac{\Delta p}{\Delta \rho}, \quad (4.41)$$

and these expressions are the discrete analogs of  $h = \partial \tilde{\varepsilon}(\rho, s)/\partial \rho$  and  $c^2 = \partial p(\rho, s)/\partial \rho$  (see the proof of Lemma 4.3). Let us check that (4.40) and (4.41) are satisfied when  $\mathbf{U}_L$  and  $\mathbf{U}_R$  are the left and right states of a shock wave. Indeed, in that case, the Rankine–Hugoniot conditions derived in Chap. III, Sect. 2, (2.3) give

$$\Delta p = \frac{(\Delta(\rho u))^2}{\Delta \rho} - \Delta(\rho u^2).$$

An easy computation using the properties of the averaging operators  $m$  and  $\hat{m}$  then implies

$$\Delta p = \hat{m}(\rho)^2 \frac{(\Delta u)^2}{\Delta \rho}.$$

Introducing the equation of the Hugoniot curve (2.12) of Chap. III,

$$\Delta \varepsilon + \frac{1}{2}(p_L + p_R)\Delta \tau = 0, \quad \tau = \frac{1}{\rho},$$

and writing

$$\frac{1}{2}(p_L + p_R) = p_L + \frac{\Delta p}{2} = p_R - \frac{\Delta p}{2},$$

we get for a shock connecting  $\mathbf{U}_L$  and  $\mathbf{U}_R$

$$\frac{\Delta \tilde{\varepsilon}}{\Delta \rho} = \frac{h_L \rho_L + h_R \rho_R}{\rho_L + \rho_R} = \mathcal{M}(h), \quad (4.42)$$

where we define the average  $\mathcal{M}$  by

$$\mathcal{M}(a) = \frac{a_L \rho_L + a_R \rho_R}{\rho_L + \rho_R}, \quad a = (a_L, a_R). \quad (4.43)$$

We can check that, for a shock, the expression for  $\bar{h}$  defined by (4.38) is also equal to  $\mathcal{M}(h)$ . In short,

$$\bar{h} = \frac{\Delta \tilde{\varepsilon}}{\Delta \rho} = \mathcal{M}(h),$$

$$\bar{c}^2 = \frac{\Delta p}{\Delta \rho} = \bar{\chi} + \bar{\kappa} \bar{h}.$$

Note that if (4.41) holds, we have in (4.29)  $\alpha_2 = 0$ . Also

$$\begin{aligned}\alpha_1 &= \frac{1}{2} \left( \Delta \rho - \hat{m}(\rho) \frac{\Delta u}{\bar{c}} \right) = \frac{\Delta p - \hat{m}(\rho) \bar{c} \Delta u}{2 \bar{c}^2}, \\ \alpha_3 &= \frac{1}{2} \left( \Delta \rho + \hat{m}(\rho) \frac{\Delta u}{\bar{c}} \right) = \frac{\Delta p + \hat{m}(\rho) \bar{c} \Delta u}{2 \bar{c}^2}.\end{aligned}$$

These are the discrete forms of the characteristic variables (see Chap. II, Sect. 2, Remark 2.1).  $\square$

### 4.3 A Roe-Type Linearization Based on Shock Curve Decomposition

In Roe's method, as we have already observed, we solve a linear Riemann problem, and the initial discontinuity breaks up into  $p$  discontinuity waves that propagate with speed  $a_k(\mathbf{u}_L, \mathbf{u}_R)$ ,  $1 \leq k \leq p$ . We can now think of a new Roe-type linearization where the wave speeds (i.e., the eigenvalues) are chosen to satisfy more closely the properties of the exact ones. This new Riemann solver is based on shock curve decomposition. Indeed, assume at first that when solving exactly the Riemann problem,  $\mathbf{U}_L$  and  $\mathbf{U}_R$  are connected by discontinuity waves only (1-shock, 2-contact discontinuity, and 3-shock). Since  $u$  and  $p$  are constant across the 2-wave, we determine the intermediate states  $u^*$  and  $p^*$  as the intersection of the shock curves in the  $(u, p)$ -plane (see Chap. III, Sect. 3). An easy computation using the Rankine–Hugoniot conditions shows that the speed of the 1-shock can be given uniquely in terms of  $\mathbf{U}_L$  and  $u^*$  and  $p^*$  by

$$\sigma_1 = u_L + \frac{(p^* - p_L)}{\rho_L(u^* - u_L)}.$$

Similarly, the speed of the 3-shock is equal to

$$\sigma_3 = u_R + \frac{(p^* - p_R)}{\rho_R(u^* - u_R)}.$$

Now, in the general case, we replace the eventual rarefaction waves by (non-admissible) shock waves. This is achieved by considering the whole shock curve in the  $(u, p)$ -plane and not only the admissible part, replacing the rarefaction curve by the nonadmissible part of the shock curve (remember that the two curves are osculatory). In the proof of Theorem 3.1, Chap. III,

the values  $u^*$  and  $p^*$  are thus obtained as the intersection of the shock curves. They determine in turn the wave speeds  $\sigma_1$  and  $\sigma_3$  and the values  $\rho_L^*, \rho_R^*$  of the density (which we had denoted  $\rho_I$  and  $\rho_{II}$  in Chap. III, Sect. 3), i.e., the whole states  $\mathbf{U}_L^*$  and  $\mathbf{U}_R^*$  on each side of the contact discontinuity.

Let us rewrite this method in terms of a Roe-type linearization. Define  $\mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)$  as the matrix whose eigenvalues  $\sigma_i, i = 1, 2, 3$ , are the above computed wave speeds

$$\sigma_1 = u_L + \frac{(p^* - p_L)}{\rho_L(u^* - u_L)}, \quad \sigma_2 = u^*, \quad \sigma_3 = u_R + \frac{(p^* - p_R)}{\rho_R(u^* - u_R)}, \quad (4.44)$$

(we have omitted the dependence in  $\mathbf{U}_L, \mathbf{U}_R$ ) and eigenvectors

$$\begin{cases} \mathbf{r}_1 = \left( 1, u_L + \frac{(p^* - p_L)}{\rho_L(u^* - u_L)}, H_L + u^* \frac{(p^* - p_L)}{\rho_L(u^* - u_L)} \right)^T, \\ \mathbf{r}_2 = \left( 1, u^*, \frac{(\rho\varepsilon)_R^* - (\rho\varepsilon)_L^*}{\rho_R^* - \rho_L^*} + \frac{(u^*)^2}{2} \right)^T, \\ \mathbf{r}_3 = \left( 1, u_R + \frac{(p^* - p_R)}{\rho_R(u^* - u_R)}, H_R + u^* \frac{(p^* - p_R)}{\rho_R(u^* - u_R)} \right)^T. \end{cases} \quad (4.45)$$

The eigenvectors have been defined in such a way that they satisfy the following identity.

*Lemma 4.8*

The vectors  $\mathbf{r}_i(\mathbf{U}_L, \mathbf{U}_R), i = 1, 2, 3$ , defined by (4.45) satisfy

$$\begin{aligned} \mathbf{U}_L^* - \mathbf{U}_L &= (\rho_L^* - \rho_L)\mathbf{r}_1(\mathbf{U}_L, \mathbf{U}_R), \\ \mathbf{U}_R^* - \mathbf{U}_L^* &= (\rho_R^* - \rho_L^*)\mathbf{r}_2(\mathbf{U}_L, \mathbf{U}_R), \\ \mathbf{U}_R - \mathbf{U}_R^* &= (\rho_R - \rho_R^*)\mathbf{r}_3(\mathbf{U}_L, \mathbf{U}_R). \end{aligned}$$

*Proof.* The formula for  $\mathbf{r}_1$  results from the fact that since  $\mathbf{U}_L^*$  and  $\mathbf{U}_L$  are connected by a 1-shock (which may be nonadmissible), the Rankine–Hugoniot condition (Chap. III, (2.10)–(2.12)) gives

$$\frac{p^* - p_L}{u^* - u_L} = -M = \frac{(u^* - u_L)\rho_L^*\rho_L}{(\rho_L^* - \rho_L)},$$

Hence

$$\begin{aligned} (\rho u)_L^* - (\rho u)_L &= \rho_L^* u^* - (\rho u)_L = (\rho_L^* - \rho_L) \left\{ u_L + \frac{(u^* - u_L)\rho_L^*}{(\rho_L^* - \rho_L)} \right\} \\ &= (\rho_L^* - \rho_L) \left\{ u_L + \frac{(p^* - p_L)}{\rho_L(u^* - u_L)} \right\}, \end{aligned}$$

which gives the second component of  $\mathbf{r}_1(\mathbf{U}_L, \mathbf{U}_R)$ . We study similarly the last component of  $\mathbf{r}_1$ , which is

$$\begin{aligned} (\rho e)_L^* - (\rho e)_L &= (\rho H)_L^* - (\rho H)_L - (p^* - p_L) \\ &= (\rho_L^* - \rho_L) \left\{ H_L + \rho_L^* \frac{(H_L^* - H_L)}{(\rho_L^* - \rho_L)} - \frac{(p^* - p_L)}{(\rho_L^* - \rho_L)} \right\}. \end{aligned}$$

Writing

$$H = \varepsilon + \frac{u^2}{2} + \frac{p}{\rho}$$

and using once more the Rankine–Hugoniot condition (Chap. III, Sect. 2, (2.12))

$$\varepsilon_L^* - \varepsilon_L + (\tau_L^* - \tau_L) \frac{(p^* + p_L)}{2} = 0,$$

we check easily that

$$\rho_L^*(H_L^* - H_L) - (p^* - p_L) = u^* \rho_L^*(u^* - u_L),$$

which implies as expected

$$\frac{\rho_L^*(H_L^* - H_L)}{(\rho_L^* - \rho_L)} - \frac{(p^* - p_L)}{(\rho_L^* - \rho_L)} = u^* \frac{(p^* - p_L)}{\rho_L(u^* - u_L)}.$$

The computations for  $\mathbf{r}_3$  are identical. The formula for  $\mathbf{r}_2$  is obvious; it is obtained by writing

$$\begin{aligned} (\rho e)_R^* - (\rho e)_L^* &= (\rho \varepsilon)_R^* - (\rho \varepsilon)_L^* + (\rho_R^* - \rho_L^*) \frac{(u^*)^2}{2} \\ &= (\rho_R^* - \rho_L^*) \left\{ \frac{(\rho \varepsilon)_R^* - (\rho \varepsilon)_L^*}{\rho_R^* - \rho_L^*} + \frac{(u^*)^2}{2} \right\}, \end{aligned}$$

which ends the calculations.  $\square$

Let us note that the formulas for  $\mathbf{r}_1$  and  $\mathbf{r}_3$  give the discrete analog of (4.18). More precisely, we can prove the following result.

*Lemma 4.9*

The matrix  $\mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)$  defined by (4.44) and (4.45) satisfies

$$\begin{aligned} \mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) &= \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)(\mathbf{U}_R - \mathbf{U}_L), \\ \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R) \rightarrow \mathbf{A}(\mathbf{U}_L) &\rightarrow \mathbf{A}(\mathbf{U}) \text{ as } \mathbf{U}_R \rightarrow \mathbf{U}_L = \mathbf{U}. \end{aligned}$$

*Proof.* First, we have by the definition of  $\sigma_i$  and by the Rankine–Hugoniot conditions

$$\begin{aligned}\mathbf{F}(\mathbf{U}_L^*) - \mathbf{F}(\mathbf{U}_L) &= \sigma_1(\mathbf{U}_L^* - \mathbf{U}_L), \\ \mathbf{F}(\mathbf{U}_R^*) - \mathbf{F}(\mathbf{U}_L^*) &= \sigma_2(\mathbf{U}_R^* - \mathbf{U}_L^*), \\ \mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_R^*) &= \sigma_3(\mathbf{U}_R - \mathbf{U}_R^*).\end{aligned}$$

Together with Lemma 4.8, this implies

$$\begin{aligned}\mathbf{F}(\mathbf{U}_R) - \mathbf{F}(\mathbf{U}_L) &= \sigma_1(\rho_L^* - \rho_L)\mathbf{r}_1 + u^*(\rho_R^* - \rho_L^*)\mathbf{r}_2 + \sigma_3(\rho_R - \rho_R^*)\mathbf{r}_3 \\ &= \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)\{(\rho_L^* - \rho_L)\mathbf{r}_1 + (\rho_R^* - \rho_L^*)\mathbf{r}_2 + (\rho_R - \rho_R^*)\mathbf{r}_3\} \\ &= \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)(\mathbf{U}_R - \mathbf{U}_L).\end{aligned}$$

Now, let us check that

$$\frac{(p^* - p_L)}{\rho_L(u^* - u_L)} \rightarrow -c \text{ as } \mathbf{U}_R \rightarrow \mathbf{U}_L = \mathbf{U}.$$

We write as above the Rankine–Hugoniot condition for the 1-shock connecting  $\mathbf{U}_L^*$  and  $\mathbf{U}_L$ ,

$$\frac{p^* - p_L}{u^* - u_L} = -M,$$

and

$$M^2 = \frac{p^* - p_L}{\tau^* - \tau_L}.$$

Since

$$\frac{p^* - p_L}{\tau^* - \tau_L} \rightarrow -\frac{\partial p}{\partial \tau} = \frac{c^2}{\tau^2}$$

(along the Hugoniot curve with center  $A_L = (\tau_L, p_L)$ ,  $s' = 0$  at  $A_L$ ), we obtain

$$\sigma_1 \rightarrow u - c = a_1, \quad \mathbf{r}_1(\mathbf{U}_L, \mathbf{U}_R) \rightarrow \mathbf{r}_1(\mathbf{U})$$

given by (4.18). Similarly,

$$\sigma_3 \rightarrow u + c = a_3, \quad \mathbf{r}_3(\mathbf{U}_L, \mathbf{U}_R) \rightarrow \mathbf{r}_3(\mathbf{U}).$$

Then, since  $p^* = p(\rho_L^*, \rho\varepsilon_L^*) = p(\rho_R^*, \rho\varepsilon_R^*)$ , we have

$$\begin{aligned}\frac{(\rho\varepsilon)_R^* - (\rho\varepsilon)_L^*}{\rho_R^* - \rho_L^*} &= \frac{(p(\rho_R^*, \tilde{\varepsilon}_L^*) - p(\rho_L^*, \tilde{\varepsilon}_L^*))/(\rho_R^* - \rho_L^*)}{(p(\rho_R^*, \tilde{\varepsilon}_L^*) - p(\rho_R^*, \tilde{\varepsilon}_R^*))/((\rho\varepsilon)_R^* - (\rho\varepsilon)_L^*)} \\ &\rightarrow -\frac{p_\rho}{p_\varepsilon} = -\frac{\chi}{\kappa}.\end{aligned}$$

We get by (4.14)

$$\begin{aligned} \frac{(\rho\varepsilon)_R^* - (\rho\varepsilon)_L^*}{(\rho_R^* - \rho_L^*)} + \frac{(u^*)^2}{2} &\rightarrow -\frac{\chi}{\kappa} + \frac{u^2}{2} \\ &= h - \frac{c^2}{\kappa} + \frac{u^2}{2} = H - \frac{c^2}{\kappa} \end{aligned}$$

and

$$\sigma_2 \rightarrow u = a_2, \quad \mathbf{r}_2(\mathbf{U}_L, \mathbf{U}_R) \rightarrow \mathbf{r}_2(\mathbf{U}),$$

which ends the proof.  $\square$

Note that for a thermally perfect gas,  $\chi = 0$  and

$$(\rho\varepsilon)_R^* = (\rho\varepsilon)_L^* = \frac{p^*}{\kappa} = \frac{p^*}{\gamma - 1}.$$

The extension of this construction to a multicomponent fluid made up of perfect and real gases is developed in Mehlmann [860]. In this situation, which is of practical importance, one can show that the associated scheme respects the positivity of the mass fractions and the local proportion of atoms; see also Colella and Glaz [329] and Larroutuou [735].

#### 4.4 Another Roe-Type Linearization Associated with a Path

Using the nonconservative form of system (2.1), Toumi [1129] has proposed a generalized Roe-type linearization associated with a path  $\Phi$ . Indeed, let  $\Phi$  be a sufficiently smooth function,

$$\Phi : [0, 1] \times \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^p \quad \text{with } \Phi(0; \mathbf{u}, \mathbf{v}) = \mathbf{u}, \quad \Phi(1; \mathbf{u}, \mathbf{v}) = \mathbf{v}.$$

One can define a Roe-type linearization depending on the path  $\Phi$ , i.e., a matrix  $\mathbf{A}(\mathbf{u}, \mathbf{v})_\Phi$  that together with properties (3.30), (3.31) satisfies

$$\mathbf{A}(\mathbf{u}, \mathbf{v})_\Phi(\mathbf{u} - \mathbf{v}) = \int_0^1 \mathbf{A}(\Phi(s; \mathbf{u}, \mathbf{v})) \frac{\partial \Phi}{\partial s}(s; \mathbf{u}, \mathbf{v}) ds. \quad (4.46)$$

In the present case where  $\mathbf{A} = \mathbf{f}'$ , the integral on the right-hand side does not depend on the path  $\Phi$  connecting  $\mathbf{u}$  and  $\mathbf{v}$ ,

$$\begin{aligned} \int_0^1 \mathbf{f}'(\Phi(s; \mathbf{u}, \mathbf{v})) \frac{\partial \Phi}{\partial s}(s; \mathbf{u}, \mathbf{v}) ds &= \mathbf{f}(\Phi(1; \mathbf{u}, \mathbf{v})) - \mathbf{f}(\Phi(0; \mathbf{u}, \mathbf{v})) \\ &= \mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u}). \end{aligned}$$

However, the matrix  $\mathbf{A}(\mathbf{u}, \mathbf{v})_\Phi$  depends on the path. Setting  $\mathbf{u}(\theta) = \mathbf{u} + \theta(\mathbf{v} - \mathbf{u})$ , we have

$$\mathbf{A}(\mathbf{u}, \mathbf{v})_{\Phi} = \int_0^1 \int_0^1 \mathbf{A}(\Phi(s; \mathbf{u}, \mathbf{v})) \frac{\partial^2 \Phi}{\partial s \partial \mathbf{u}}(s; \mathbf{u}, \mathbf{u}(\theta)) d\theta ds. \quad (4.47)$$

If the “canonical” straight line is chosen, we obtain

$$\mathbf{A}(\mathbf{u}, \mathbf{v})_{\Phi} = \int_0^1 \mathbf{A}(\mathbf{u} + s(\mathbf{v} - \mathbf{u})) ds.$$

Let us check that the Roe matrix obtained via the parameter vectors in the case of a perfect gas corresponds to the straight path in the parameter vector variables.

*Lemma 4.10*

Assume the hypotheses of Lemma 4.2. The matrix (4.47) associated with the following path,

$$\Phi = \Phi_W(s; \mathbf{U}_L, \mathbf{U}_R) = \mathbf{U}(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)), \quad (4.48)$$

coincides with the Roe-type linearization (4.4).

*Proof.* The result follows directly from the property that  $\mathbf{U}(\mathbf{W})$  and  $\mathbf{G}(\mathbf{W}) = \mathbf{F}(\mathbf{U}(\mathbf{W}))$  are homogeneous quadratic functions of  $\mathbf{W}$ , together with the fact that integrating a linear function of  $\mathbf{W}$  along the path  $\Phi$  gives

$$\int_0^1 (\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds = \frac{1}{2}(\mathbf{W}_R + \mathbf{W}_L) = \mathbf{W}^*. \quad (4.49)$$

We have, on the one hand,

$$\begin{aligned} & \int_0^1 \mathbf{A}(\Phi(s; \mathbf{U}_L, \mathbf{U}_R)) \frac{\partial \Phi}{\partial s}(s; \mathbf{U}_L, \mathbf{U}_R) ds \\ &= \int_0^1 \mathbf{A}(\mathbf{U}(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))) \\ & \quad \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) \cdot (\mathbf{W}_R - \mathbf{W}_L) ds. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathbf{U}_R - \mathbf{U}_L &= \mathbf{U}(\mathbf{W}_R) - \mathbf{U}(\mathbf{W}_L) \\ &= \int_0^1 \frac{d}{ds} \mathbf{U}(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds \\ &= \int_0^1 \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds \cdot (\mathbf{W}_R - \mathbf{W}_L). \end{aligned}$$

Thus, we get

$$\begin{aligned} \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)_{\Phi_W} &= \left( \int_0^1 \mathbf{A}(\mathbf{U}(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))) \right. \\ &\quad \left. \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds \right) \cdot \left( \int_0^1 \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds \right)^{-1}. \end{aligned} \quad (4.50)$$

Now we compute successively each integral in (4.50); this requires expressions for  $\mathbf{U}'(\mathbf{W})$  and  $\mathbf{G}'(\mathbf{W}) = \mathbf{A}(\mathbf{U}(\mathbf{W}))\mathbf{U}'(\mathbf{W})$ . From (4.9), the Jacobian matrix  $\mathbf{U}'$  for a perfect gas gives

$$\mathbf{U}'(\mathbf{W}) = \begin{pmatrix} 2w_1 & 0 & 0 \\ w_2 & w_1 & 0 \\ w_3/\gamma & w_2(\gamma-1)/\gamma & w_1/\gamma \end{pmatrix} \quad (4.51)$$

which is linear in the  $w_i$ ; by (4.49), it yields

$$\int_0^1 \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds = \mathbf{U}'(\mathbf{W}^*).$$

Now, by Lemma 4.2,

$$\mathbf{G}'(\mathbf{W}) = \begin{pmatrix} w_2 & w_1 & 0 \\ w_3(\gamma-1)/\gamma & w_2(\gamma+1)/\gamma & w_1(\gamma-1)/\gamma \\ 0 & w_3 & w_2 \end{pmatrix},$$

which is also linear in the  $w_i$ 's; thus, using (4.49) again,

$$\int_0^1 \mathbf{G}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds = \mathbf{G}'(\mathbf{W}^*). \quad (4.52)$$

It is then easy to check that

$$\mathbf{A}_{\Phi_W}(\mathbf{U}_L, \mathbf{U}_R) = \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R) = \mathbf{A}(\bar{\mathbf{U}}),$$

where the matrix  $\mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)$  was computed above in (4.16) and (4.21) via the parameter vectors, since by (4.4)

$$\mathbf{A}(\bar{\mathbf{U}}) = \mathbf{G}'(\mathbf{W}^*)\mathbf{U}'(\mathbf{W}^*)^{-1},$$

which thus coincides with (4.50).  $\square$

These considerations lead us to define a generalized Roe matrix for a real gas by formula (4.50) associated with the same straight path (4.48) in the parameter vector variables (4.8), i.e.,

$$\begin{aligned} \mathbf{A}_{\Phi_W}(\mathbf{U}_L, \mathbf{U}_R) &= \int_0^1 \mathbf{G}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))ds \\ &\quad \left( \int_0^1 \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))ds \right)^{-1}. \end{aligned} \quad (4.53)$$

Let us just sketch the ideas for the computation of  $\mathbf{A}_{\Phi_W}$ . As we have already observed, in (4.9) Lemma 4.2, the last component  $w_1 w_3 - p$  of  $\mathbf{U}$  is no longer a quadratic function of  $\mathbf{W}$ , and we must keep the expression of  $\mathbf{U}$  in the form

$$\mathbf{U}(\mathbf{W}) = (w_1^2, w_1 w_2, w_1 w_3 - p)^T.$$

Thus, in the expression (4.51) of  $\mathbf{U}'(\mathbf{W})$ , the last line is replaced by

$$w_3 - p_{w_1}, \quad -p_{w_2}, \quad w_1 - p_{w_3}.$$

Similarly,  $\mathbf{G}(\mathbf{W}) = \mathbf{F}(\mathbf{U}(\mathbf{W})) = (w_1 w_2, p + w_2^2, w_2 w_2)^T$ , and in  $\mathbf{G}'(\mathbf{W})$ , the second line is replaced by

$$p_{w_1}, \quad p_{w_2} + 2w_2, \quad p_{w_3}.$$

In order to compute the integrals in (4.53), we proceed as follows:

(i) We set

$$\tilde{p}_{w_i} = \int_0^1 p_{w_i}(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))ds,$$

which gives immediately

$$\int_0^1 \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))ds = \begin{pmatrix} 2w_1^* & 0 & 0 \\ w_2^* & w_1^* & 0 \\ w_3^* - \tilde{p}_{w_1} & -\tilde{p}_{w_2} & w_1^* - \tilde{p}_{w_3} \end{pmatrix},$$

and

$$\int_0^1 \mathbf{G}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))ds = \begin{pmatrix} w_2^* & w_1^* & 0 \\ \tilde{p}_{w_1} & \tilde{p}_{w_2} + 2w_2^* & \tilde{p}_{w_3} \\ 0 & w_3^* & w_2^* \end{pmatrix}.$$

Then formula (4.53) gives the matrix  $\mathbf{A}_{\Phi_W}(\mathbf{U}_L, \mathbf{U}_R)$ .

(ii) In order to express this matrix in a more practical form, we introduce the partial derivatives

$$\frac{\partial p}{\partial \rho}|_{q,E} = P_\rho, \quad \frac{\partial p}{\partial q}|_{p,E} = P_q, \quad \frac{\partial p}{\partial E}|_{\rho,q} = P_E.$$

In fact  $P_\rho = K$  given by (4.17), as we observed at the end of the proof of Lemma 4.4, and we rewrite the matrix  $\mathbf{A}$  given by (4.16) in terms of  $P_\rho, P_q$ , and  $P_E$

$$\mathbf{A}(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 \\ P_\rho - u^2 & 2u + P_q & P_E \\ u(P_\rho - H) & H + uP_q & (1 + P_E)u \end{pmatrix}. \quad (4.54)$$

*Lemma 4.11*

The matrix  $\mathbf{A}_{\Phi_W}(\mathbf{U}_L, \mathbf{U}_R)$  defined by (4.53) is the Jacobian matrix (4.55)  $\mathbf{A}(\bar{u}, \bar{H}, \tilde{P}_\rho, \tilde{P}_q, \tilde{P}_E)$  evaluated at Roe's averaged states  $\bar{u}, \bar{H}$  given by (4.21) and the averaged values of the partial derivatives.

*Proof.* First we express the partial derivatives  $P_\rho$ ,  $P_q$ , and  $P_E$  in terms of the  $p_{w_i}$ ; we get the relations

$$\begin{aligned} P_\rho &= K = p_\rho + p_\varepsilon \frac{u^2}{2} = \chi + \kappa \frac{u^2}{2} \\ P_\rho &= \frac{w_1 p_{w_1} - w_2 p_{w_2} - w_3 p_{w_3}}{2w_1(w_1 - p_{w_3})}, \end{aligned} \quad (4.55)$$

$$P_q = -\kappa u = \frac{p_{w_2}}{(w_1 - p_{w_3})} = -uP_E, \quad (4.56)$$

and

$$P_E = \kappa = \frac{p_{w_3}}{(w_1 - p_{w_3})}.$$

Since  $p$  is a function of only two thermodynamic variables, we note that the derivatives are not independent and that the equality  $P_q = -uP_E$  implies in turn

$$p_{w_2} = -p_{w_3} \frac{w_2}{w_1}. \quad (4.57)$$

Next, we set

$$\begin{cases} \tilde{P}_\rho = \frac{w_1^* \tilde{p}_{w_1} - w_2^* \tilde{p}_{w_2} - w_3^* \tilde{p}_{w_3}}{2w_1^*(w_1^* - \tilde{p}_{w_3})}, \\ \tilde{P}_q = \frac{\tilde{p}_{w_2}}{w_1^* - \tilde{p}_{w_3}}, \\ \tilde{P}_E = \frac{\tilde{p}_{w_3}}{w_1^* - \tilde{p}_{w_3}}. \end{cases} \quad (4.58)$$

These expressions are those obtained by replacing  $w_i$  by  $w_i^*$  and  $P_\rho$  by the corresponding expression  $\tilde{P}_\rho$ , where  $p_{w_i}$  is replaced by  $\tilde{p}_{w_i}$ .

By analogy with the identity obtained for an ideal gas,

$$\begin{aligned} \mathbf{A}(\bar{\mathbf{U}}) &= \mathbf{G}'(\mathbf{W}^*) \mathbf{U}'(\mathbf{W}^*)^{-1} \\ &= \int_0^1 \mathbf{G}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds \left( \int_0^1 \mathbf{U}'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L)) ds \right)^{-1}, \end{aligned}$$

we check that we obtain for the matrix  $\mathbf{A}_{\Phi_W}(\mathbf{u}, \mathbf{v})$  the expression

$$\mathbf{A}(\mathbf{u}, \mathbf{v})_{\Phi_W} = \begin{pmatrix} 0 & 1 & 0 \\ \tilde{P}_\rho - \bar{u}^2 & 2\bar{u} + \tilde{P}_q & \tilde{P}_E \\ \bar{u}(\tilde{P}_\rho - \bar{H}) & \bar{H} + \bar{u}\tilde{P}_q & (1 + \tilde{P}_E)\bar{u} \end{pmatrix}, \quad (4.59)$$

which is the Jacobian matrix (4.54) evaluated at  $\bar{u}, \bar{H}$  and the averaged values  $\tilde{P}_{\rho, m, E}$ .  $\square$

*Remark 4.6.* We can define  $\tilde{\chi}$  and  $\tilde{\kappa}$  from formulas (4.55)–(4.58),

$$\begin{aligned} \tilde{\kappa} &= \tilde{P}_E, \\ \tilde{\chi} &= \tilde{P}_\rho + \tilde{P}_q \frac{w_2^*}{w_1^*}. \end{aligned}$$

We might ask whether condition (4.36) of the preceding section is satisfied. First, since

$$\begin{aligned} \Delta p &= \int_0^1 \frac{d}{ds} (p(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))) ds \\ &= \int_0^1 \{p'(\mathbf{W}_L + s(\mathbf{W}_R - \mathbf{W}_L))\} \cdot (\mathbf{W}_R - \mathbf{W}_L) ds, \end{aligned}$$

we have

$$\Delta p = \sum \tilde{p}_{w_i} \Delta_{w_i}.$$

Similarly,

$$\begin{aligned} \tilde{P}_\rho \Delta \rho + \tilde{P}_q \Delta q + \tilde{P}_E \Delta(\rho H) &= \tilde{P}_\rho (2w_1^* \Delta w_1) + \tilde{P}_q (w_2^* \Delta w_1 + w_1^* \Delta w_2) \\ &\quad + \tilde{P}_E (w_3^* \Delta w_1 + w_1^* \Delta w_3). \end{aligned}$$

Hence, by (4.55)–(4.58),

$$\begin{aligned} \tilde{P}_\rho \Delta \rho + \tilde{P}_q \Delta q + \tilde{P}_E \Delta(\rho H) \\ &= (\tilde{p}_{w_1} \Delta w_1 + \tilde{p}_{w_2} \Delta w_2 + \tilde{p}_{w_3} \Delta w_3) \frac{w_1^*}{(w_1^* - \tilde{p}_{w_3})} \\ &= (1 + \tilde{P}_E) \Delta p, \end{aligned}$$

which in turn yields

$$\Delta p = \tilde{P}_\rho \Delta \rho + \tilde{P}_q \Delta q + \tilde{P}_E \Delta E.$$

Then, in general, for a “real gas,”

$$\Delta p \neq \tilde{\chi} \Delta \rho + \tilde{\kappa} \Delta(\rho \varepsilon).$$

This results from the fact that (4.57) does not hold for the mean values

$$\tilde{p}_{w_2} \neq -\tilde{p}_{w_3} \frac{w_2^*}{w_1^*}.$$

For more details, we refer to Toumi [1129] (see also Glaister [526] and Coquel and Liou [359]).  $\square$

*Remark 4.7.* The above construction of  $\mathbf{A}(\mathbf{u}, \mathbf{v})_\Phi$  is linked to the definition of a nonconservative product  $(\mathbf{A}(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial x})_\Phi$  associated with a path (see Dal Maso et al. [387]). This was introduced to study systems in nonconservation form (see Chap. II, Sect. 5, Remark 5.2). The “canonical” straight line corresponds to the nonconservative product of Volpert.

The above approach has since been extended by C. Parés and coauthors who define *path conservative* schemes for nonconservative problems [245]. However, the numerical approximation of nonconservative products remains a difficult issue as illustrated in [12].  $\square$

## 4.5 The Case of the Gas Dynamics System in Lagrangian Coordinates

Recall that the gas dynamics system written in Lagrangian coordinates is of the form

$$\frac{\partial \mathbf{V}}{\partial t} + \frac{\partial}{\partial m} \mathbf{f}(\mathbf{V}) = \mathbf{0} \quad (4.60)$$

where

$$\mathbf{V} = (\tau, u, e)^T, \quad \mathbf{f}(\mathbf{V}) = (-u, p, pu)^T. \quad (4.61)$$

As usual, we supplement (4.60) with the equation of state

$$p = p(\tau, \varepsilon), \quad \varepsilon = e - \frac{1}{2}u^2. \quad (4.62)$$

The Jacobian matrix  $\mathbf{A}(\mathbf{V})$  of the flux function  $\mathbf{f}(\mathbf{V})$  is then given by

$$\mathbf{A}(\mathbf{V}) = \begin{pmatrix} 0 & -1 & 0 \\ \frac{\partial p}{\partial \tau} & -u \frac{\partial p}{\partial \varepsilon} & \frac{\partial p}{\partial \varepsilon} \\ u \frac{\partial p}{\partial \tau} & p - u^2 \frac{\partial p}{\partial \varepsilon} & u \frac{\partial p}{\partial \varepsilon} \end{pmatrix}. \quad (4.63)$$

Let us now construct Godunov-type methods of approximation of system (4.60).

### 4.5.1 Roe's Method in Lagrangian Coordinates

We begin by introducing a fairly general Roe's method of solution which extends the approach of the preceding sections. For any pair  $(\mathbf{V}_L, \mathbf{V}_R)$ , we look for a matrix  $\bar{\mathbf{A}} = \mathbf{A}(\mathbf{V}_L, \mathbf{V}_R)$  such that

$$\mathbf{f}(\mathbf{V}_R) - \mathbf{f}(\mathbf{V}_L) = \mathbf{A}(\mathbf{V}_L, \mathbf{V}_R)(\mathbf{V}_R - \mathbf{V}_L)$$

or in short

$$\Delta \mathbf{f} = \bar{\mathbf{A}} \Delta \mathbf{V}.$$

We are given an incremental form of the equation of state (4.62)

$$\Delta p = \mu \Delta \tau + \xi \Delta u + \kappa \Delta e \quad (4.64)$$

where  $\mu$ ,  $\xi$ , and  $\kappa$  are average values of  $\frac{\partial p}{\partial \tau}$ ,  $-u \frac{\partial p}{\partial \varepsilon}$ , and  $\frac{\partial p}{\partial \varepsilon}$ , respectively, which have to be chosen conveniently. Note that the notation  $\kappa$  which we use presently for the few following lines may be misleading since it does not correspond to that already used above:  $p$  is presently the function  $p(\tau, \varepsilon)$ , and the notation  $\kappa$  was used in the preceding section for  $\frac{\partial p(\rho, \rho\varepsilon)}{\partial(\rho\varepsilon)}$  where the pressure  $p$  was considered as a function of  $(\rho, \rho\varepsilon)$ .

Setting for any  $\alpha \in [0, 1]$  and any pair  $(\varphi_L, \varphi_R)$

$$\varphi_\alpha = \alpha \varphi_L + \beta \varphi_R, \quad \beta = 1 - \alpha,$$

i.e., a convex combination of  $(\varphi_L, \varphi_R)$ , we have

$$\Delta(pu) = u_\alpha \Delta p + p_\beta \Delta u$$

and this is analogous to results stated in Lemma 4.6 so that (4.64) yields

$$\Delta(pu) = \mu u_\alpha \Delta \tau + (\xi u_\alpha + p_\beta) \Delta u + \kappa u_\alpha \Delta e.$$

In that way, we obtain a Roe matrix

$$\bar{\mathbf{A}} = \begin{pmatrix} 0 & -1 & 0 \\ \mu & \xi & \kappa \\ \mu u_\alpha & \xi u_\alpha + p_\beta & \kappa u_\alpha \end{pmatrix} \quad (4.65)$$

which depends on the parameter  $\alpha$ . The eigenvalues of  $\bar{\mathbf{A}}$  are the solutions of the characteristic polynomial

$$\lambda(\lambda^2 - \lambda(\xi + \kappa u_\alpha) + \mu - \kappa p_\beta) = 0.$$

Since  $\mu - \kappa p_\beta$  is an average value of  $\frac{\partial p}{\partial \tau} - p \frac{\partial p}{\partial \varepsilon} = -C^2$  where  $C = \frac{c}{\tau}$  is the Lagrangian sound speed, it is natural to suppose that  $\mu$ ,  $\xi$ , and  $\kappa$  are chosen in such a way that

$$\mu - \kappa \beta < 0.$$

Then the eigenvalues  $\bar{a}_k = a_k(\mathbf{V}_L, \mathbf{V}_R)$ ,  $1 \leq k \leq 3$ , of the matrix  $\bar{\mathbf{A}}$  are

$$\bar{a}_1 = -C_L < \bar{a}_2 = 0 < \bar{a}_3 = C_R \quad (4.66)$$

provided  $C_L, C_R > 0$  satisfy

$$\begin{cases} C_R - C_L = \xi + \kappa u_\alpha, \\ -C_L C_R = \mu - \kappa p_\beta. \end{cases} \quad (4.67)$$

One can easily check that the corresponding eigenvectors  $\bar{\mathbf{r}}_k = \mathbf{r}_k(\mathbf{V}_L, \mathbf{V}_R)$ ,  $1 \leq k \leq 3$ , may be chosen as

$$\bar{\mathbf{r}}_1 = \begin{pmatrix} -1 \\ -C_L \\ p_\beta - C_L u_\alpha \end{pmatrix}, \quad \bar{\mathbf{r}}_2 = \begin{pmatrix} \kappa \\ 0 \\ -\mu \end{pmatrix}, \quad \bar{\mathbf{r}}_3 = \begin{pmatrix} -1 \\ C_R \\ p_\beta + C_R u_\alpha \end{pmatrix}. \quad (4.68)$$

Moreover, we have

$$\Delta \mathbf{V} = \sum_{k=1}^3 \bar{\alpha}_k \bar{\mathbf{r}}_k, \quad \bar{\alpha}_k = \alpha_k(\mathbf{V}_L, \mathbf{V}_R) \quad (4.69)$$

with

$$\begin{cases} \bar{\alpha}_1 = \frac{1}{C_L(C_L + C_R)} (\Delta p - C_R \Delta u), \\ \bar{\alpha}_3 = \frac{1}{C_R(C_L + C_R)} (\Delta p + C_L \Delta u). \end{cases} \quad (4.70)$$

Hence the Riemann solver associated with the Roe matrix  $\bar{\mathbf{A}}$  is given by

$$\tilde{\mathbf{w}}\left(\frac{m}{t}; \mathbf{V}_L, \mathbf{V}_R\right) = \begin{cases} \mathbf{V}_L, & \frac{m}{t} < C_L, \\ \mathbf{V}_L^* = \mathbf{V}_L + \bar{\alpha}_1 \bar{\mathbf{r}}_1, & -C_L < \frac{m}{t} < 0, \\ \mathbf{V}_R^* = \mathbf{V}_R - \bar{\alpha}_3 \bar{\mathbf{r}}_3, & 0 < \frac{m}{t} < C_R, \\ \mathbf{V}_R, & \frac{m}{t} > C_R. \end{cases} \quad (4.71)$$

In particular, we obtain from (4.71)

$$\begin{cases} \tau_L^* = \tau_L - \frac{1}{C_L(C_L + C_R)} (\Delta p - C_R \Delta u), \\ \tau_R^* = \tau_R + \frac{1}{C_R(C_L + C_R)} (\Delta p + C_L \Delta u) \end{cases} \quad (4.72)$$

and

$$u_L^* = u_R^* = u^* = \frac{1}{C_L + C_R} (C_L u_L + C_R u_R - \Delta p). \quad (4.73)$$

which can also be written as

$$u^* = u_L + \frac{1}{C_L + C_R} (C_R \Delta u - \Delta p) = u_R - \frac{1}{C_L + C_R} (C_L \Delta u + \Delta p)$$

On the other hand, setting

$$p^* = \frac{1}{C_L + C_R} (C_R p_L + C_L p_R - C_L C_R \Delta u), \quad (4.74)$$

one checks that the numerical flux  $\mathbf{g}(\mathbf{V}_L, \mathbf{V}_R)$  can be written in the form

$$\mathbf{g}(\mathbf{V}_L, \mathbf{V}_R) = \begin{pmatrix} -u^* \\ p^* \\ (pu)_{\frac{1}{2}} + p_\beta X + u_\alpha Y \end{pmatrix} \quad (4.75)$$

where we have set

$$\begin{aligned} X &= \frac{1}{C_L + C_R} (-\Delta p + \frac{1}{2} (C_R - C_L) \Delta u), \\ Y &= -\frac{1}{C_L + C_R} (C_L C_R \Delta u + \frac{1}{2} (C_R - C_L) \Delta p) \end{aligned} \quad (4.76)$$

and used the notation for any pair  $(\varphi_L, \varphi_R)$

$$\varphi_{\frac{1}{2}} = \frac{1}{2} (\varphi_L + \varphi_R).$$

*Remark 4.8.* Note that, given the wave speeds  $C_L$  and  $C_R$ , the triple  $(\mu, \xi, \kappa)$  may be viewed as a solution of the linear system consisting of Eqs. (4.64) and (4.67) and depending on the parameter  $\alpha$ . Since its determinant is equal to

$$-\left\{ (\alpha - \frac{1}{2}) (\Delta p \Delta \tau + (\Delta u)^2) + p_{\frac{1}{2}} \Delta \tau + \Delta \varepsilon \right\},$$

this system is singular for any  $\alpha$  if and only if

$$\Delta p \Delta \tau + (\Delta u)^2 = p_{\frac{1}{2}} \Delta \tau + \Delta \varepsilon = 0.$$

But it is an easy matter to check that the above relations hold if and only if the states  $\mathbf{V}_L$  and  $\mathbf{V}_R$  are connected by a shock. As a consequence, provided that  $\mathbf{V}_L$  and  $\mathbf{V}_R$  are not connected by a shock, one can always find an  $\alpha$  and a triple  $(\mu, \xi, \kappa)$  in such a way that  $\bar{\mathbf{A}}$  is a Roe matrix whose eigenvalues are indeed  $-C_L$ , 0, and  $C_R$ . If the two states are connected by a shock, the jump formulas obtained above for the intermediate states are exactly satisfied if one takes the exact shock speed for  $-C_L$ , for a 1-shock (resp.  $C_R$  for a 3-shock),

and there is no constraint on the other speed. This expresses the fact that for a shock propagating at speed  $\sigma$ , the jump relation  $\Delta\mathbf{f} = \sigma\Delta\mathbf{V}$  together with the relation for a Roe matrix  $\Delta\mathbf{f} = \bar{\mathbf{A}}\Delta\mathbf{V}$  yields that  $\Delta\mathbf{V}$  is an eigenvector of  $\bar{\mathbf{A}}$  associated with the eigenvalue  $\sigma$ .  $\square$

#### 4.5.2 Toward a Positively Conservative and Entropy Satisfying Godunov-Type Scheme

We now want to cure the above drawbacks of Roe's method. Following Remark 4.8, we first observe that the Riemann solver (4.71) depends on the coefficients  $\mu$ ,  $\xi$ , and  $\kappa$  only through the wave speeds  $C_L$  and  $C_R$ . Thus, with any pair  $(C_L, C_R)$  of positive numbers and any  $\alpha \in [0, 1]$ , we associate an approximate Riemann solver of the form (4.71). We first check that it indeed defines a Godunov-type method, i.e., it satisfies

$$\Delta\mathbf{f} = -C_L(\mathbf{V}_L^* - \mathbf{V}_L) + C_R(\mathbf{V}_R - \mathbf{V}_R^*)$$

or more explicitly

$$\begin{cases} -\Delta u = -C_L(\tau_L^* - \tau_L) + C_L(\tau_R - \tau_R^*), \\ \Delta p = -C_L(u^* - u_L) + C_L(u_R - u^*), \\ \Delta(pu) = -C_L(e_L^* - e_L) + C_L(e_R - e_R^*). \end{cases} \quad (4.77)$$

The first two relations (4.77) follow from (4.72) and (4.73), respectively. On the other hand, since

$$\begin{cases} e_L^* - e_L = \frac{1}{C_L(C_L + C_R)}(\Delta p - C_R\Delta u)(p_\beta - C_L u_\alpha), \\ e_R^* - e_R = \frac{1}{C_R(C_L + C_R)}(\Delta p + C_L\Delta u)(p_\beta + C_R u_\alpha), \end{cases} \quad (4.78)$$

we have

$$-C_L(e_L^* - e_L) + C_R(e_R - e_R^*) = u_\alpha \Delta p + p_\beta \Delta u = \Delta(pu).$$

Next, a “canonical” choice of the parameter  $\alpha$  is suggested by the following result:

*Proposition 4.1*

Let us choose

$$\alpha = \frac{C_L}{C_L + C_R}. \quad (4.79)$$

Then the numerical flux of the Godunov-type method associated with the Riemann solver (4.71) is given by

$$\mathbf{g}(\mathbf{V}_L, \mathbf{V}_R) = (-u^*, p^*, p^* u^*)^T \quad (4.80)$$

where  $u^*$  and  $p^*$  are defined by (4.73) and (4.74), respectively.

*Proof.* We need only to check that

$$g_3(\mathbf{V}_L, \mathbf{V}_R) = p^* u^*.$$

Simple computations show that

$$u^* = u_{\frac{1}{2}} + X, \quad p^* = p_{\frac{1}{2}} + Y$$

and therefore by (4.76)

$$g_3(\mathbf{V}_L, \mathbf{V}_R) - p^* u^* = (pu)_{\frac{1}{2}} + p_\beta X + u_\alpha Y - (p_{\frac{1}{2}} + Y)(u_{\frac{1}{2}} + X).$$

Since

$$\begin{aligned} (pu)_{\frac{1}{2}} + p_\beta X + u_\alpha Y - (p_{\frac{1}{2}} + Y)(u_{\frac{1}{2}} + X) &= \\ &= \frac{1}{4} \Delta p \Delta u - XY + (\alpha - \frac{1}{2})(\Delta p X - \Delta u Y) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{4} \Delta p \Delta u - XY &= -\frac{C_R - C_L}{2(C_L + C_R)^2} (\Delta p - C_R \Delta u)(\Delta p + C_L \Delta u) \\ \Delta p X - \Delta u Y &= -\frac{1}{C_L + C_R} (\Delta p - C_R \Delta u)(\Delta p + C_L \Delta u), \end{aligned}$$

we obtain

$$g_3(\mathbf{V}_L, \mathbf{V}_R) - p^* u^* = -\left(\frac{C_R - C_L}{2(C_L + C_R)^2} + \frac{\alpha - 1/2}{C_L + C_R}\right) (\Delta p - C_R \Delta u)(\Delta p + C_L \Delta u).$$

This expression vanishes for all pair  $(\Delta u, \Delta p)$  if and only if

$$\alpha - \frac{1}{2} + \frac{C_R - C_L}{2(C_L + C_R)} = 0$$

i.e., if and only if (4.79) holds.  $\square$

Clearly, in order to satisfy the above equation (or equivalently (4.79)), the simplest choice of the parameters  $C_L$ ,  $C_R$ , and  $\alpha$  consists in taking

$$C_L = C_R = \bar{C}, \quad \alpha = \frac{1}{2} \quad (4.81)$$

which leads to

$$\begin{cases} u^* = u_{\frac{1}{2}} - \frac{1}{2\bar{C}} \Delta p, \\ p^* = p_{\frac{1}{2}} - \frac{\bar{C}}{2} \Delta u. \end{cases} \quad (4.82)$$

In addition, we check

*Lemma 4.12*

Under the assumption (4.81), we have

$$\tau_L^* = \tau_L + \frac{1}{\bar{C}}(u^* - u_L), \quad \tau_R^* = \tau_R - \frac{1}{\bar{C}}(u^* - u_R), \quad (4.83)$$

$$u^* = u_L - \frac{1}{\bar{C}}(p^* - p_L) = u_R + \frac{1}{\bar{C}}(p^* - p_R), \quad (4.84)$$

$$e_L^* = e_L - \frac{1}{\bar{C}}(p^* u^* - p_L u_L), \quad e_R^* = e_R + \frac{1}{\bar{C}}(p^* u^* - p_R u_R), \quad (4.85)$$

$$\varepsilon_L^* = \varepsilon_L + \frac{1}{2\bar{C}^2}(p^{*2} - p_L^2), \quad \varepsilon_R^* = \varepsilon_R + \frac{1}{2\bar{C}^2}(p^{*2} - p_R^2). \quad (4.86)$$

*Proof.* Assume (4.81). Then (4.72) gives

$$\tau_L^* = \tau_L - \frac{1}{2\bar{C}^2}(\Delta p - \bar{C} \Delta u), \quad \tau_R^* = \tau_R + \frac{1}{2\bar{C}^2}(\Delta p + \bar{C} \Delta u),$$

and (4.83) follows from the first Eq. (4.82). Next, (4.84) follows at once from the second Eq. (4.82). On the other hand, (4.78) gives

$$\begin{aligned} e_L^* - e_L &= \frac{1}{2\bar{C}^2}(\Delta p - \bar{C} \Delta u)(p_{\frac{1}{2}} - \bar{C} u_{\frac{1}{2}}) = \\ &= \frac{1}{2\bar{C}^2}(p_{\frac{1}{2}} \Delta p - \bar{C} \Delta(pu) + \bar{C}^2 u_{\frac{1}{2}} \Delta u). \end{aligned}$$

while (4.82) yields

$$\begin{aligned} p^* u^* - p_L u_L &= (p_{\frac{1}{2}} - \frac{\bar{C}}{2} \Delta u)(u_{\frac{1}{2}} - \frac{1}{2\bar{C}} \Delta p) - p_L u_L + \\ &= -\frac{1}{2\bar{C}}(p_{\frac{1}{2}} \Delta p - \bar{C} \Delta(pu) + \bar{C}^2 u_{\frac{1}{2}} \Delta u) \end{aligned}$$

and the first Eq. (4.85) follows. The second Eq. (4.85) is proved in a similar way. Let us now evaluate  $\varepsilon_L^* = e_L^* - \frac{1}{2} u^{*2}$ . We can write

$$\varepsilon_L^* = \varepsilon_L + \frac{1}{2}(u_L^2 - u^{*2}) - \frac{1}{\bar{C}}(p^* u^* - p_L u_L).$$

Replacing in the above equation  $u^*$  by its first expression (4.84) yields the expression (4.86) of  $\varepsilon_L^*$ . We analogously evaluate  $\varepsilon_R^*$ .  $\square$

*Remark 4.9.* Note that Eqs. (4.83)–(4.85) can be written as

$$-\bar{C} \begin{pmatrix} \tau_L^* - \tau_L \\ u^* - u_L \\ e_L^* - e_L \end{pmatrix} = \begin{pmatrix} -(u^* - u_L) \\ p^* - p_L \\ p^* u^* - p_L u_L \end{pmatrix}, \quad \bar{C} \begin{pmatrix} \tau_R - \tau_R^* \\ u_R - u^* \\ e_R - e^* \end{pmatrix} = \begin{pmatrix} -(u_R - u^*) \\ p_R - p^* \\ p_R u_R - p^* u^* \end{pmatrix}.$$

These equations may be indeed interpreted as Rankine–Hugoniot jump relations across the waves with speeds  $\pm\bar{C}$ . More precisely, in the case (4.81), we can derive independently the approximate Riemann solver in the following way. We associate with the gas dynamics system (4.60) the system

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial \Pi}{\partial m} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial}{\partial m}(\Pi u) = 0 \end{cases}$$

where we have replaced the pressure  $p$  by a new variable  $\Pi$ ; this enables us to relax the nonlinearity due to the pressure law, so that the corresponding sound waves are linearly degenerate. Then, the system must also be complemented by an equation for  $\Pi$ ; it is done in such a way that one gets a linearly degenerate system with eigenvalues  $\pm\bar{C}, 0$ . Then setting

$$\mathbf{Z} = \begin{pmatrix} \mathbf{V} \\ \Pi \end{pmatrix}, \quad \Pi_L = p(\tau_L, \varepsilon_L), \quad \Pi_R = p(\tau_R, \varepsilon_R),$$

we introduce the approximate Riemann solver

$$\tilde{\mathbf{w}}_L\left(\frac{m}{t}; \mathbf{V}_L, \mathbf{V}_R\right) = \begin{cases} \mathbf{Z}_L, & \frac{m}{t} < -\bar{C}, \\ \mathbf{Z}_L^*, & -\bar{C} < \frac{m}{t} < 0, \\ \mathbf{Z}_R^*, & 0 < \frac{m}{t} < \bar{C}, \\ \mathbf{Z}_R, & \frac{m}{t} > \bar{C} \end{cases}$$

and we determine the pair  $(\mathbf{Z}_L^*, \mathbf{Z}_R^*)$  by requiring that the Rankine–Hugoniot relations are to be satisfied when crossing the waves with speeds  $-\bar{C}, 0, \bar{C}$ . Indeed, we obtain

$$u_L^* = u_R^* = u^*, \quad \Pi_L^* = \Pi_R^* = p^*$$

where  $u^*, p^*$  are given by (4.82), while  $\tau_L^*, \tau_R^*, e_L^*$ , and  $e_R^*$  are obtained from Eqs. (4.83) and (4.85). It is also worthwhile to notice that  $p^*$  differs in general from  $p(\tau_L^*, \varepsilon_L^*)$  and  $p(\tau_R^*, \varepsilon_R^*)$ .

This approach, where the approximate Riemann solver is defined as the exact solver for an “approximate system,” is followed to derive the so-called *relaxation schemes* and will be given more attention in the last section of this chapter.  $\square$

In the sequel, for the sake of simplicity, we will restrict ourselves to the choice (4.81). Then the corresponding scheme reads

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \frac{\Delta t}{\Delta m} (u_{j+1/2}^n - u_{j-1/2}^n), \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta m} (p_{j+1/2}^n - p_{j-1/2}^n), \\ e_j^{n+1} = e_j^n - \frac{\Delta t}{\Delta m} (p_{j+1/2}^n u_{j+1/2}^n - p_{j-1/2}^n u_{j-1/2}^n) \end{cases} \quad (4.87)$$

with

$$\begin{cases} u_{j+1/2}^n = \frac{1}{2} (u_j^n + u_{j+1}^n) - \frac{1}{2C_{j+1/2}^n} (p_{j+1}^n - p_j^n), \\ p_{j+1/2}^n = \frac{1}{2} (p_j^n + p_{j+1}^n) - \frac{C_{j+1/2}^n}{2} (u_{j+1}^n - u_j^n) \end{cases} \quad (4.88)$$

and the CFL condition

$$\frac{\Delta t}{\Delta m} C_{j+1/2}^n \leq \frac{1}{2}. \quad (4.89)$$

It remains to choose the wave speed  $\bar{C}$  (and therefore the  $C_{j+1/2}^n$ 's) in order to guarantee that the Riemann solver is positively conservative and entropy satisfying. We first require the specific volumes  $\tau_L^*$  and  $\tau_R^*$  to be positive. Using (4.72) with  $C_L = C_R = \bar{C}$ , we find the conditions

$$\begin{cases} 2\bar{C}^2 \tau_L + \bar{C} \Delta u - \Delta p > 0, \\ 2\bar{C}^2 \tau_R + \bar{C} \Delta u + \Delta p > 0 \end{cases}$$

or equivalently

$$\bar{C} > \begin{cases} \frac{1}{4\tau_L} (\sqrt{(\Delta u)^2 + 8\tau_L \Delta p} - \Delta u) & \text{if } (\Delta u)^2 + 8\tau_L \Delta p \geq 0, \\ \frac{1}{4\tau_R} (\sqrt{(\Delta u)^2 - 8\tau_R \Delta p} - \Delta u) & \text{if } (\Delta u)^2 - 8\tau_R \Delta p \geq 0. \end{cases} \quad (4.90)$$

Next, we suppose that the equation of state (4.62) derives from a complete equation of state  $\varepsilon = \varepsilon(\tau, s)$  which implies  $p = -\frac{\partial \varepsilon}{\partial \tau}(\tau, s)$ . Denoting by  $I(a, b)$  the interval with end points  $a$  and  $b$ , we can now state

*Lemma 4.13*

Assume that  $\bar{C}$  satisfies the inequalities (4.90) and

$$\begin{cases} -\frac{\partial p}{\partial \tau}(\tau, s_L) \leq \bar{C}^2 \text{ for all } \tau \in I(\tau_L, \tau_L^*), \\ -\frac{\partial p}{\partial \tau}(\tau, s_R) \leq \bar{C}^2 \text{ for all } \tau \in I(\tau_R, \tau_R^*). \end{cases} \quad (4.91)$$

Then, we have

$$\varepsilon_L^* \geq \varepsilon(\tau_L^*, s_L) > 0, \quad \varepsilon_R^* \geq \varepsilon(\tau_R^*, s_R) > 0. \quad (4.92)$$

*Proof.* Let us check the first inequality (4.92). Using (4.86), we have

$$\varepsilon_L^* - \varepsilon(\tau_L^*, s_L) = \varepsilon_L + \frac{1}{2\bar{C}^2}(p^{*2} - p_L^2) - \varepsilon(\tau_L^*, s_L)$$

or equivalently

$$\begin{aligned} \varepsilon_L^* - \varepsilon(\tau_L^*, s_L) &= \varepsilon_L - \frac{p_L^2}{2\bar{C}^2} - \varepsilon(\tau_L^*, s_L) + \frac{p(\tau_L^*, s_L)^2}{2\bar{C}^2} + \\ &+ \frac{1}{\bar{C}^2}p(\tau_L^*, s_L)(p^* - p(\tau_L^*, s_L)) + \frac{1}{2\bar{C}^2}(p^* - p(\tau_L^*, s_L))^2. \end{aligned}$$

Since by (4.83)

$$\tau_L - \frac{p_L}{\bar{C}^2} = \tau_L^* - \frac{p^*}{\bar{C}^2}$$

we can write

$$\begin{aligned} \varepsilon_L^* - \varepsilon(\tau_L^*, s_L) &= \varepsilon_L - \frac{p_L^2}{2\bar{C}^2} - \varepsilon(\tau_L^*, s_L) + \frac{p(\tau_L^*, s_L)^2}{2\bar{C}^2} + \\ &+ p(\tau_L^*, s_L)(\tau_L + \frac{p_L}{\bar{C}^2} - \tau_L^* - \frac{p(\tau_L^*, s_L)}{\bar{C}^2}) + \frac{1}{2\bar{C}^2}(p^* - p(\tau_L^*, s_L))^2. \end{aligned}$$

In order to prove that  $\varepsilon_L^* \geq \varepsilon(\tau_L^*, s_L)$ , it is enough to check that

$$\varepsilon_L - \frac{p_L^2}{2\bar{C}^2} - \varepsilon(\tau_L^*, s_L) + \frac{p(\tau_L^*, s_L)^2}{2\bar{C}^2} + p(\tau_L^*, s_L)(\tau_L + \frac{p_L}{\bar{C}^2} - \tau_L^* - \frac{p(\tau_L^*, s_L)}{\bar{C}^2}) \geq 0.$$

Let us then introduce the function

$$\begin{aligned} \varphi(\tau) &= \varepsilon(\tau, s_0) - \frac{p(\tau, s_0)^2}{2\bar{C}^2} - \varepsilon(\tau_0, s_0) + \frac{p(\tau_0, s_0)^2}{2\bar{C}^2} + \\ &+ p(\tau_0, s_0)(\tau + \frac{p(\tau, s_0)}{\bar{C}^2} - \tau_0 - \frac{p(\tau_0, s_0)}{\bar{C}^2}) \end{aligned}$$

where  $\tau_0$  and  $s_0$  are fixed. Clearly  $\varphi(\tau_0) = 0$ . On the other hand, since  $\frac{\partial \varepsilon}{\partial \tau}(\tau, s) = -p(\tau, s)$ , we have

$$\varphi'(\tau) = -(p(\tau, s_0) - p(\tau_0, s_0)) \left( 1 + \frac{1}{\bar{C}^2} \frac{\partial p}{\partial \tau}(\tau, s_0) \right).$$

Under the condition

$$-\frac{\partial p}{\partial \tau}(\tau, s_0) \leq \bar{C}^2,$$

$\varphi'(\tau)$  has the sign of  $p(\tau_0, s_0) - p(\tau, s_0)$  and therefore the sign of  $\tau - \tau_0$  since  $\tau \mapsto p(\tau, s)$  is a decreasing function. Hence  $\varphi$  reaches its minimum at  $\tau_0$  so that  $\varphi(\tau) \geq 0$ . By applying this result with  $\tau_0 = \tau_L^*$ ,  $s_0 = s_L$ , and  $\tau = \tau_L$ , we obtain

$$-\frac{\partial p}{\partial \tau}(\tau, s_L) \leq \bar{C}^2 \text{ in } I(\tau_L, \tau_L^*) \Rightarrow \varepsilon_L^* \geq \varepsilon(\tau_L^*, s_L).$$

The second inequality (4.92) is proved in a similar way.  $\square$

### Theorem 4.3

Assume that (4.81) holds. Then the approximate Riemann solver (4.71) is positively conservative and entropy satisfying provided that  $\bar{C}$  satisfies the conditions (4.90) and (4.91).

*Proof.* Under the conditions (4.90) and (4.91), the intermediate states  $\mathbf{V}_L^*$  and  $\mathbf{V}_R^*$  in (4.71) belong to the set of states  $\Omega = \{\mathbf{V}; \tau > 0, \varepsilon > 0\}$ . It remains to prove that the Riemann solver is entropy satisfying. Since  $(-s, 0)$  is an entropy pair, we have to check the inequality

$$-\bar{C}(s(\tau_L^*, \varepsilon_L^*) - s_L) + \bar{C}(s_R - s(\tau_R^*, \varepsilon_R^*)) \leq 0$$

which holds as soon as

$$s(\tau_L^*, \varepsilon_L^*) \geq s_L, \quad s(\tau_R^*, \varepsilon_R^*) \geq s_R.$$

Since  $\frac{\partial \tau}{\partial \varepsilon}(\tau, \varepsilon) > 0$ , this indeed follows from (4.92).  $\square$

#### 4.5.3 A Simple Riemann Solver in Eulerian Coordinates

The corresponding solver in Eulerian coordinates is a direct consequence of the equivalence result derived in Theorem 3.3 and both schemes share the same properties. From (4.71), and using the correspondence between formula (3.24) and (3.20), the Riemann solver writes

$$\tilde{\mathbf{w}}\left(\frac{x}{t}; \mathbf{U}_L, \mathbf{U}_R\right) = \begin{cases} \mathbf{U}_L, & \frac{x}{t} < u_L - \bar{C}\tau_L, \\ \mathbf{U}_L^* = \mathbf{U}(\mathbf{V}_L^*), & u_L - \bar{C}\tau_L < \frac{x}{t} < u^*, \\ \mathbf{U}_R^* = \mathbf{U}(\mathbf{V}_R^*), & u^* < \frac{x}{t} < u_R + \bar{C}\tau_R, \\ \mathbf{U}_R, & \frac{x}{t} > u_R + \bar{C}\tau_R. \end{cases}$$

where we use the notation  $\mathbf{U}_L^* = \mathbf{U}(\mathbf{V}_L^*) = \frac{1}{\tau_L^*}(1, u_L^*, e_L^*)^T$  together with the formulas of Lemma 4.12; similarly for  $\mathbf{U}_R^* = \mathbf{U}(\mathbf{V}_R^*)$ . The resulting scheme is thus entropy satisfying and positively conservative.

## 5 Flux Vector Splitting Methods

### 5.1 General Formulation

The upwind difference scheme is easy to implement in the nonlinear case when the eigenvalues of  $\mathbf{A}(\mathbf{u})$  are all of one sign. When they are of mixed sign, we want to give a direct generalization of the upwind difference scheme (1.10) obtained in the linear case  $\mathbf{f}(\mathbf{u}) = \mathbf{A}\mathbf{u}$ , which corresponds to the numerical flux function (2.17),

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{A}^+ \mathbf{u} + \mathbf{A}^- \mathbf{v}.$$

In fact, we look for a flux splitting of the form

$$\mathbf{f}(\mathbf{u}) = \mathbf{f}^+(\mathbf{u}) + \mathbf{f}^-(\mathbf{u}), \quad (5.1)$$

and we set

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{f}^+(\mathbf{u}) + \mathbf{f}^-(\mathbf{v}). \quad (5.2)$$

We shall require that the Jacobian matrix  $\mathbf{A}_+(\mathbf{u}) \equiv \mathbf{f}'^+(\mathbf{u})$  (resp.  $\mathbf{A}_-(\mathbf{u}) \equiv \mathbf{f}'^-(\mathbf{u})$ ) has positive (resp. negative) eigenvalues.

Consider first the *scalar* case. If the function  $f$  is strictly convex, we have already found such a decomposition when deriving the Engquist–Osher’s scheme (see Sect. 3.4). Let  $\bar{u} \in \mathbb{R}$  be the only sonic point,  $a(\bar{u}) = f'(\bar{u}) = 0$ . We can set

$$f^+(v) = f(\max(v, \bar{u})), \quad f^-(v) = f(\min(v, \bar{u}))$$

so that

$$f(v) = f^+(v) + f^-(v) - f(\bar{u}),$$

and then (up to an additive constant)

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

Let us consider now the case of a system whose flux function is a *homogeneous* function of degree 1, i.e.,

$$\mathbf{f}(\mu \mathbf{u}) = \mu \mathbf{f}(\mathbf{u}), \quad \forall \mu \in \mathbb{R},$$

which is indeed the case for ideal gas dynamics. In that case, Euler’s identity for homogeneous functions gives

$$\mathbf{f}(\mathbf{u}) = \mathbf{A}(\mathbf{u})\mathbf{u},$$

so that following Steger and Warming, we define

$$\mathbf{f}^+(\mathbf{u}) = \mathbf{A}^+(\mathbf{u})\mathbf{u}, \quad \mathbf{f}^-(\mathbf{u}) = \mathbf{A}^-(\mathbf{u})\mathbf{u}, \quad (5.3)$$

where  $\mathbf{A}^\pm$  is defined by (1.9). We obtain the numerical flux

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{f}^+(\mathbf{u}) + \mathbf{f}^-(\mathbf{v}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v})) - \frac{1}{2}(|\mathbf{A}(\mathbf{v})|\mathbf{v} - |\mathbf{A}(\mathbf{u})|\mathbf{u}).$$

*Remark 5.1.* Let us see why the present flux vector splitting scheme is relatively viscous near stagnation points. The “viscosity” or dissipation term  $\mathbf{D}(\mathbf{u}, \mathbf{v})$  such that

$$\mathbf{g}(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v})) - \frac{1}{2}\mathbf{D}(\mathbf{u}, \mathbf{v})$$

is given by

$$\mathbf{D}(\mathbf{u}, \mathbf{v}) = (\mathbf{f}^+(\mathbf{v}) - \mathbf{f}^-(\mathbf{v})) - (\mathbf{f}^+(\mathbf{u}) - \mathbf{f}^-(\mathbf{u})),$$

which in the particular case of Steger and Warming’s scheme gives

$$\mathbf{D}(\mathbf{u}, \mathbf{v}) = |\mathbf{A}(\mathbf{v})|\mathbf{v} - |\mathbf{A}(\mathbf{u})|\mathbf{u}.$$

Now  $\mathbf{f}(\mathbf{v}) = \mathbf{f}(\mathbf{u})$  does not imply  $\mathbf{f}^+(\mathbf{v}) - \mathbf{f}^-(\mathbf{v}) = \mathbf{f}^+(\mathbf{u}) - \mathbf{f}^-(\mathbf{u})$ , roughly because  $v = -u$  does not imply  $|v|v = |u|u$ ! For nearby values of  $u$  and  $v$ ,  $v = -u$  implies that  $u$  and  $v$  are near 0. It follows that a flux vector splitting scheme does not give a good resolution near “sonic” points such that  $a_k = 0$ , since there may exist nearby states  $\mathbf{u}, \mathbf{v}$  such that  $\operatorname{sgn} a_k(\mathbf{u}) \neq \operatorname{sgn} a_k(\mathbf{v})$ . This happens for stationary contact discontinuities in the case of Euler equations (see Coquel and Liou [358]).

Also, after rearrangement, we can write Steger and Warming’s viscosity term as

$$\mathbf{D}(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(|\mathbf{A}(\mathbf{u})| + |\mathbf{A}(\mathbf{v})|)(\mathbf{v} - \mathbf{u}) + \frac{1}{2}(|\mathbf{A}(\mathbf{v})| - |\mathbf{A}(\mathbf{u})|)(\mathbf{u} + \mathbf{v}).$$

A scheme satisfying

$$\mathbf{D}(\mathbf{u}, \mathbf{v}) = \left| \mathbf{A} \left( \frac{\mathbf{u} + \mathbf{v}}{2} \right) \right| (\mathbf{v} - \mathbf{u}) + \mathbf{o}(|\mathbf{u} - \mathbf{v}|)$$

is called an *upstream scheme* by Harten et al. [595]. Thus, the above scheme is of upstream form except near “sonic” points.  $\square$

If we set

$$\mathbf{u} = \sum_{k=1}^p \alpha_k(\mathbf{u}) \mathbf{r}_k(\mathbf{u}),$$

where  $(\mathbf{r}_k)$  is a basis of eigenvectors of  $\mathbf{A}$  and the  $\alpha_k(\mathbf{u})$  are the characteristic variables, we have

$$\mathbf{f}(\mathbf{u}) = \sum_{k=1}^p a_k(\mathbf{u}) \alpha_k(\mathbf{u}) \mathbf{r}_k(\mathbf{u})$$

and

$$\mathbf{f}^\pm(\mathbf{u}) = \sum_{k=1}^p a_k^\pm(\mathbf{u}) \alpha_k(\mathbf{u}) \mathbf{r}_k(\mathbf{u}). \quad (5.4)$$

*Remark 5.2.* Let us note that, for any  $\mu \in \mathbb{R}$ ,

$$\mathbf{A}(\mu \mathbf{u}) = \mathbf{A}(\mathbf{u}),$$

and hence by (1.9)

$$\mathbf{A}^\pm(\mu \mathbf{u}) = \mathbf{A}^\pm(\mathbf{u}) \quad \text{and} \quad \mathbf{f}^\pm(\mu \mathbf{u}) = \mu \mathbf{f}^\pm(\mathbf{u}).$$

By Euler's identity, this implies

$$\mathbf{f}^\pm(\mathbf{u}) = \mathbf{f}^{\pm'}(\mathbf{u}) \mathbf{u} \equiv \mathbf{A}_\pm(\mathbf{u}) \mathbf{u}.$$

However,  $\mathbf{A}_+(\mathbf{u}) = \mathbf{f}^{+'}(\mathbf{u})$  is not equal to  $\mathbf{A}^+(\mathbf{u})$  (defined by (1.9)), though  $\mathbf{A}_+(\mathbf{u}) \mathbf{u} = \mathbf{A}^+(\mathbf{u}) \mathbf{u}$  (see Lerat [762] for details). Also, the notations may be misleading since  $\mathbf{A}_+(\mathbf{u})$  and  $\mathbf{A}_-(\mathbf{u})$  do not necessarily have positive or negative eigenvalues, and one will have to check their sign.  $\square$

## 5.2 Application to the Gas Dynamics Equations: (I) Steger and Warming's Approach

Let us check the flux homogeneity property in the case of the gas dynamics equations.

*Lemma 5.1*

We assume that the equation of state satisfies

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

Then, the flux function is homogeneous of degree 1.

*Proof.* We have from (2.19)

$$\mathbf{U} = \begin{pmatrix} \rho \\ q \\ E \end{pmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} q \\ q^2/\rho + p \\ (E+p)m/\rho \end{pmatrix}, \quad p = p\left(\rho, \frac{E}{\rho} - \frac{q^2}{2\rho^2}\right).$$

Setting

$$p(\mathbf{U}) = p(\rho, \varepsilon) = p\left(\rho, \frac{E}{\rho} - \frac{q^2}{2\rho^2}\right),$$

we note that by assumption

$$p(\mu\mathbf{U}) = p(\mu\rho, \varepsilon) = \mu p(\rho, \varepsilon).$$

Hence

$$\mathbf{F}(\mu\mathbf{U}) = \begin{pmatrix} \mu q \\ \mu q^2/\rho + p(\mu\rho, \varepsilon) \\ (\mu E + p(\mu\rho, \varepsilon))q/\rho \end{pmatrix} = \mu \begin{pmatrix} q \\ q^2/\rho + p(\rho, \varepsilon) \\ (E + p(\rho, \varepsilon))q/\rho \end{pmatrix},$$

and

$$\mathbf{F}(\mu\mathbf{U}) = \mu\mathbf{F}(\mathbf{U}),$$

which proves the homogeneity of  $\mathbf{F}$ .  $\square$

*Remark 5.3.* Lemma 5.1 applies in the case of a polytropic ideal gas since

$$p = (\gamma - 1)\rho\varepsilon.$$

More generally, it also applies for an equation of state of a thermally (not necessarily calorically) perfect gas,

$$p = \rho T(\varepsilon). \quad (5.5)$$

For such an equation,  $p$  satisfies (4.31) (see Chap. III, Sect. 1.2, (1.21))

$$p = \rho p_\rho + \tilde{\varepsilon} p_{\tilde{\varepsilon}}.$$

Now, an easy computation using the expression (4.16) of  $\mathbf{A}(\mathbf{U})$  found in the preceding section shows that

$$\mathbf{F}(\mathbf{U}) = \mathbf{A}(\mathbf{U})\mathbf{U} + (p - \rho p_\rho - \tilde{\varepsilon} p_{\tilde{\varepsilon}})(0, 1, u)^T.$$

Hence, the homogeneity property

$$\mathbf{F}(\mathbf{U}) = \mathbf{A}(\mathbf{U})\mathbf{U}$$

is equivalent to requiring that  $p$  satisfies (4.31). Note that the flux Jacobian matrix remains homogeneous for nonequilibrium mixtures of thermally perfect gases (see Grossman and Cinnella [309, 562]; see also [880]).  $\square$

Let us express the split fluxes, given by the formulas (5.4),

$$\mathbf{U} = \sum_{k=1}^3 \alpha_k(\mathbf{U}) \mathbf{r}_k(\mathbf{U}),$$

and

$$\mathbf{F}^\pm(\mathbf{U}) = \sum_{k=1}^3 a_k^\pm(\mathbf{U}) \alpha_k(\mathbf{U}) \mathbf{r}_k(\mathbf{U}),$$

where again we have  $a_1 = u - c, a_2 = u, a_3 = u + c$ , and the vectors  $\mathbf{r}_k$  are given by (4.18).

We first consider the case of a thermally perfect gas where the flux is a homogeneous function. Solving the system for the  $(\alpha_k)$  as in Lemma 4.7, we get for a thermally perfect gas (5.5)

$$\alpha_1 = \alpha_3 = \frac{\rho}{2\gamma}, \quad \alpha_2 = p \frac{(\gamma - 1)}{c^2} = \rho \frac{(\gamma - 1)}{\gamma}. \quad (5.6)$$

Setting

$$\mathbf{F}_k(\mathbf{U}) = a_k(\mathbf{U}) \alpha_k(\mathbf{U}) \mathbf{r}_k(\mathbf{U}), \quad k = 1, 2, 3,$$

we have

$$\mathbf{F}(\mathbf{U}) = \mathbf{F}_1(\mathbf{U}) + \mathbf{F}_2(\mathbf{U}) + \mathbf{F}_3(\mathbf{U}).$$

From (4.18), the  $\mathbf{F}_k(\mathbf{U})$  are given by

$$\mathbf{F}_1(\mathbf{U}) = a_1 \frac{\rho}{2\gamma} \begin{pmatrix} 1 \\ u - c \\ H - uc \end{pmatrix}, \quad \mathbf{F}_2(\mathbf{U}) = a_2 \frac{\rho(\gamma - 1)}{\gamma} \begin{pmatrix} 1 \\ u \\ H - \frac{c^2}{(\gamma - 1)} \end{pmatrix}, \quad (5.7a)$$

$$\mathbf{F}_3(\mathbf{U}) = a_3 \frac{\rho}{2\gamma} \begin{pmatrix} 1 \\ u + c \\ H + uc \end{pmatrix}. \quad (5.7b)$$

(Note that  $H - \frac{c^2}{(\gamma - 1)} = \frac{u^2}{2}$  for a polytropic ideal gas.)

We can make precise the expressions of  $\mathbf{F}^\pm$  in terms of the  $\mathbf{F}_k$  since we are in the following situation: each eigenvalue has a single zero that we denote by  $\bar{u}_i$  ( $\bar{u}_1 = c, \bar{u}_2 = 0, \bar{u}_3 = -c$ ). We obtain

$$\begin{cases} u \geq c & \Rightarrow a_k \geq 0 \text{ and } \mathbf{F}^+(\mathbf{U}) = \mathbf{F}(\mathbf{U}), \mathbf{F}^-(\mathbf{U}) = 0, \\ u \leq -c & \Rightarrow a_k \leq 0 \text{ and } \mathbf{F}^-(\mathbf{U}) = \mathbf{F}(\mathbf{U}), \mathbf{F}^+(\mathbf{U}) = 0. \end{cases} \quad (5.8a)$$

Thus, in supersonic regions, which are by definition those where the Mach number  $|M| = |\frac{u}{c}|$  satisfies  $|M| \geq 1$ , we recover the upwind scheme as expected.

Now, for  $-c < u < c$ ,

$$\begin{cases} -c < u < 0 \Rightarrow a_3 > 0 > a_2 > a_1 \\ \text{and } \mathbf{F}^+(\mathbf{U}) = \mathbf{F}_3(\mathbf{U}), \mathbf{F}^-(\mathbf{U}) = \mathbf{F}_1(\mathbf{U}) + \mathbf{F}_2(\mathbf{U}), \\ 0 < u < c \Rightarrow a_3 > a_2 > 0 > a_1 \\ \text{and } \mathbf{F}^+(\mathbf{U}) = \mathbf{F}_2(\mathbf{U}) + \mathbf{F}_3(\mathbf{U}), \mathbf{F}^-(\mathbf{U}) = \mathbf{F}_1(\mathbf{U}). \end{cases} \quad (5.8b)$$

The eigenvalues of  $\mathbf{F}'^+$  and  $\mathbf{F}'^-$  have the correct sign (i.e., resp.  $\geq 0$  and  $\leq 0$ ) for a perfect gas, but this is not straightforward to prove (see Lerat [762], who proves the result for  $1 < \gamma < \frac{5}{3}$  and Vinokur and Montagné [1173] who plot their numerical values). Note that the eigenvalues of  $\mathbf{A}_+(\mathbf{U}) = \mathbf{F}'^+(\mathbf{U})$  are not continuous in general at a zero  $\bar{u}_i$ .

We turn to an arbitrary equilibrium gas for which the homogeneity property is not satisfied. There no longer exists a natural flux splitting. Nonetheless, following the approach of Sanders and Prendergast [958] (see also Vinokur and Montagné [1173]), we can extend the above decomposition, i.e., write  $\mathbf{F}(\mathbf{U})$  in the form

$$\mathbf{F}(\mathbf{U}) = \mathbf{F}_1(\mathbf{U}) + \mathbf{F}_2(\mathbf{U}) + \mathbf{F}_3(\mathbf{U}),$$

and then define  $\mathbf{F}^\pm(\mathbf{U})$  as in (5.8). The idea is to look for convective fluxes of the form

$$\mathbf{F}_k(\mathbf{U}) = a_k(\mathbf{U})\mathbf{U}_k$$

with the constraints

$$\begin{cases} \sum_{k=1}^3 \mathbf{U}_k = \mathbf{U} = \rho \left( 1, u, \varepsilon + \frac{u^2}{2} \right)^T, \\ \sum_{k=1}^3 \mathbf{F}_k(\mathbf{U}) = \mathbf{F}(\mathbf{U}) = \rho \left( u, u^2 + \frac{p}{\rho}, \left( e + \frac{p}{\rho} \right) u \right)^T. \end{cases} \quad (5.9)$$

Now, we set

$$\mathbf{U}_k = \rho_k \left( 1, a_k, \varepsilon_k + \frac{a_k^2}{2} \right)^T, \quad 1 \leq k \leq 3,$$

which means that the velocity in  $\mathbf{U}_k$  is equal to  $a_k$ , as was indeed the case in (5.7). The six unknowns  $\rho_k$ ,  $\varepsilon_k$  are determined from Eqs. (5.9), which represent only five equations since one of them is trivial (the variable  $\rho u$  is present in both  $\mathbf{U}$  and  $\mathbf{F}$ ). We easily obtain as in (5.6)

$$\rho_1 = \rho_3 = \alpha_1 = \alpha_3 = \frac{\rho}{2\gamma}, \quad \rho_2 = \alpha_2 = \frac{\rho(\gamma-1)}{\gamma},$$

where  $\gamma$  is defined by

$$\gamma = \rho c^2 / p. \quad (5.10)$$

Also,

$$\varepsilon_1 = \varepsilon_3, \quad \sum_{k=1}^3 \varepsilon_k \rho_k = \rho \left( \varepsilon - \frac{c^2}{2\gamma} \right).$$

There remains one degree of freedom, which in Vinokur and Montagné [1173] is taken in the form of a nondimensional parameter  $\beta$  such that

$$\varepsilon_2 = \varepsilon - (1 - \beta) \frac{c^2}{\gamma(\gamma - 1)}.$$

Since by (5.10)

$$H - \frac{c^2}{(\gamma - 1)} = H - \frac{p}{\rho} - \frac{c^2}{\gamma(\gamma - 1)} = \frac{u^2}{2} + \varepsilon - \frac{c^2}{\gamma(\gamma - 1)},$$

the expression has been determined so that the case  $\beta = 0$  corresponds to a thermally perfect gas. It is then proved by arguing on the eigenvalues of  $\mathbf{A}_+$  that a convenient choice is indeed

$$\varepsilon_2 = \varepsilon - \frac{c^2}{\gamma(\gamma - 1)},$$

thus leading to

$$\varepsilon_1 = \varepsilon_3 = \varepsilon + \frac{c^2(2 - \gamma)}{\gamma}$$

and to the same expressions for the fluxes  $\mathbf{F}_k$  as above.

In short, we can still take the same expressions (5.8) for  $\mathbf{F}^\pm$  in order to define a generalized Steger and Warming flux vector splitting (see [1072]).

### 5.3 Application to the Gas Dynamics Equations: (II) Van Leer's Approach

As we have already observed, Steger and Warming's decomposition is not continuous in general at a sonic point (see Liang and Chan [799]). Another splitting of the fluxes in the range  $|u| < c$  has been derived by van Leer under the requirement that the split fluxes and their Jacobians be continuous. Besides the usual conditions ("consistency"  $\mathbf{F}^+ + \mathbf{F}^- = \mathbf{F}$ , and the correct sign for the eigenvalues of  $\mathbf{F}^\pm$ ), further requirements are:

- (i) A symmetry property,

$$\mathbf{F}^+(u) = \pm \mathbf{F}^-(u) \quad \text{if } \mathbf{F}(u) = \pm \mathbf{F}(-u)$$

(all other quantities but  $u$  constant).

- (ii)  $\mathbf{F}^\pm$  is a polynomial in  $u$  (of lowest possible degree).
- (iii)  $\mathbf{F}^\pm'$  has a vanishing eigenvalue for  $|M| < 1$ .

Note that (i) is linked to the fact that  $a_1(u) = u - c = -a_3(-u)$  and is satisfied by the splitting of the preceding section. For what concerns condition (ii), in the system of gas dynamics, each component of  $\mathbf{F}$  is a polynomial in  $u$  of degree  $\leq 3$ . Condition (iii) is imposed in the hope that the eigenvalues of  $\mathbf{F}^\pm'$  will have the correct sign (since the one that is most likely to have the wrong sign is “forced” to vanish). We can seek cubic polynomials in  $u$  for the components of  $\mathbf{F}^\pm$  too. For supersonic flow,  $|M| > 1$ ,  $\mathbf{F}^+ = \mathbf{F}^- = 0$ , and the continuity requirement on  $\mathbf{F}^+$  requires that the polynomials include a factor  $(u \pm c)^2$ . Thus, for each component  $\mathbf{F}_i^\pm$  of  $\mathbf{F}^\pm$ , we postulate in the range  $|u| < c$

$$F_i^\pm = (u \pm c)^2(au + b), \quad i = 1, 2, 3,$$

where the coefficients of the remaining first-order factor are obtained by using the symmetry property (i) and the “consistency”  $F_i = F_i^+ + F_i^-$ . (The notation of the components  $F_i$  is to be distinguished from the fluxes  $\mathbf{F}_i$  of the preceding section.) For the splitting of the first two components of  $\mathbf{F}$  (mass flux  $F_1 = q = \rho u$  and momentum flux  $F_2 = qu + p$ ), using the identity

$$q = \rho u = \frac{\rho}{4c} \{(u + c)^2 - (u - c)^2\},$$

we easily obtain

$$F_1^\pm = \pm \frac{\rho}{4c} (u \pm c)^2, \quad (5.11)$$

and by (5.10) ( $p = \frac{\rho c^2}{\gamma}$ ),

$$F_2^\pm = F_1^\pm \frac{\{(\gamma - 1)u \pm 2c\}}{\gamma}. \quad (5.12)$$

In fact, the splitting of the third component (energy flux  $F_3 = (E + p)u = \rho u^3/2 + pu + \rho \varepsilon u$ ) obtained in this way would differ from that derived by van Leer for an ideal gas in that (iii) is not satisfied. If  $\gamma$  is constant in (5.10),  $\varepsilon = \frac{c^2}{\gamma(\gamma - 1)}$ , and the third component of  $\mathbf{F}$  can be written as

$$F_3 = q \left\{ \frac{u^2}{2} + \frac{c^2}{\gamma(\gamma - 1)} \right\}.$$

Then, it is easy to check that a convenient splitting is

$$\begin{aligned} F_3^\pm &= F_1^\pm \frac{((\gamma - 1)u \pm 2c)^2}{2(\gamma^2 - 1)} \\ &= \frac{\gamma^2}{2(\gamma^2 - 1)} \frac{F_2^{\pm 2}}{F_1^\pm}. \end{aligned} \quad (5.13)$$

When  $\gamma$  is not constant, we express the flux in the form

$$F_3^\pm = F_1^\pm \left\{ \frac{((\gamma - 1)u \pm 2c)^2}{2(\gamma^2 - 1)} + \left( \varepsilon - \frac{c^2}{\gamma(\gamma - 1)} \right) + \beta(u \mp c)^2 \right\}.$$

Let us note the following simple facts:

- For a polytropic ideal gas, choosing  $\beta = 0$  reduces to van Leer's splitting. The sign of the two corresponding nonzero eigenvalues of  $\mathbf{F}^\pm'$  is studied in van Leer [1152] and Vinokur and Montagné [1173]. In particular, we get the right sign in the range  $1 \leq \gamma \leq 3$ .
- For  $\beta = \frac{1}{(\gamma+1)}$ , we get

$$F_3^\pm = F_1^\pm H,$$

or with more explicit notation,

$$(Eu + pu)^\pm = q^\pm H,$$

which means that the total enthalpy is preserved (see also Liou et al. [818]).

- For nonconstant  $\gamma$ , condition (iii) cannot be identically satisfied for any choice of  $\beta$ . For a thermally perfect gas, it can be shown that the choice  $\beta = 0$  leads to a negative eigenvalue in the whole range  $|M| < 1$ , but since it is very small, this is not a major drawback, and the split energy flux for nonconstant  $\gamma$  will indeed be taken as in the ideal gas case,

$$F_3^\pm = F_1^\pm \left\{ \frac{((\gamma - 1)u \pm 2c)^2}{2(\gamma^2 - 1)} + \varepsilon - \frac{c^2}{\gamma(\gamma - 1)} \right\}.$$

In short, in the range  $|\frac{u}{c}| \geq 1$ , we keep the upwind scheme, and the decomposition is given by formulas (5.8a), while in the range  $|\frac{u}{c}| < 1$ , we define a splitting

$$\mathbf{F} = \mathbf{F}^+ + \mathbf{F}^-, \quad \mathbf{F}^\pm = (F_i^\pm)$$

by the formulas (5.11)–(5.13). This splitting satisfies the requirements (i)–(iii). Moreover, for an ideal gas, the Jacobian matrix  $\mathbf{F}^+'$  (resp.  $\mathbf{F}^-'$ ) has positive (resp. negative) eigenvalues.

For other studies of the real gas case, we refer to Larroutuou and Fezoui [739], Grossman and Cinella [562], Shuen et al. [818], Müller [880], and Montagné [872] and also Roe [986].

We shall study in Sect. 7 the kinetic schemes that are also flux vector splitting methods.

Other papers related to flux splitting are Liou and Steffen [816, 817], Radespiel and Kroll [964], Coquel and Liou [359], and Chen and LeFloch [290] and for the relation with parabolized schemes Chang and Merkle [277]).

## 6 Van Leer's Second-Order Method

Van Leer's method generalizes Godunov's method to obtain a second-order scheme. Following the ideas of G.R., Chapter 4, Section 3 [539], we present it for a space grid that is not necessarily uniform. Let us recall the three main steps of the method:

- (i) A “reconstruction step,” which consists in constructing a piecewise linear function  $\tilde{\mathbf{v}}$  from given cell averages  $\mathbf{v}_j^n$ ,

$$\tilde{\mathbf{v}}^n(x) = \mathbf{v}_j^n + (x - x_j) \frac{\mathbf{S}_j^n}{\Delta x_j}, \quad x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}}, \quad (6.1)$$

where  $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ ; the slopes  $\mathbf{S}_j^n$  are to be detailed later on.

- (ii) An “evolution step,” which involves either an exact or an approximate Riemann solver. One solves

$$\begin{cases} \frac{\partial \mathbf{w}}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{w})}{\partial x} = \mathbf{0}, & 0 \leq t \leq \Delta t, \\ \mathbf{w}(x, 0) = \tilde{\mathbf{v}}^n(x), \end{cases} \quad (6.2)$$

which gives  $\mathbf{w}(., \Delta t)$ .

- (iii) A “cell averaging step” (or projection step), which gives  $\mathbf{v}_j^{n+1}$ . As in (2.5), we project (in the sense of  $\mathbf{L}^2$ ) the solution  $\mathbf{w}(., \Delta t)$  onto the piecewise constant functions

$$\mathbf{v}_j^{n+1} = \frac{1}{\Delta x_j} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \mathbf{w}(x, \Delta t) dx. \quad (6.3)$$

### 6.1 Van Leer's Method for Systems

Provided we assume some convenient CFL condition so that the waves issued from the points  $x_{j-\frac{1}{2}}$  and  $x_{j+\frac{1}{2}}$  do not interact, the solution of (6.2) is in fact obtained by juxtaposition of local generalized Riemann problems. In order to derive a more explicit form of the scheme, we integrate Eq. (6.2) over a cell  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times (0, \Delta t)$ ,

$$\int_0^{\Delta t} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left( \frac{\partial \mathbf{w}}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{w})}{\partial x} \right) dx dt = \mathbf{0}.$$

We obtain

$$\begin{aligned} & \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} (\mathbf{w}(x, \Delta t) - \mathbf{w}(x, 0)) dx \\ & + \int_0^{\Delta t} (\mathbf{f}(\mathbf{w}(x_{j+\frac{1}{2}}, t)) - \mathbf{f}(\mathbf{w}(x_{j-\frac{1}{2}}, t))) dt = \mathbf{0} \end{aligned}$$

since the flux is continuous, and then by (6.3)

$$\Delta x_j (\mathbf{v}_j^{n+1} - \mathbf{v}_j^n) + \int_0^{\Delta t} \{\mathbf{f}(\mathbf{w}(x_{j+\frac{1}{2}}, t)) - \mathbf{f}(\mathbf{w}(x_{j-\frac{1}{2}}, t))\} dt = \mathbf{0}.$$

We are left with the evaluation of the numerical flux

$$\mathbf{g}_{j+\frac{1}{2}}^n = \frac{1}{\Delta t} \int_0^{\Delta t} \mathbf{f}(\mathbf{w}(x_{j+\frac{1}{2}}, t)) dt. \quad (6.4)$$

To begin with, using the midpoint rule, we write

$$\frac{1}{\Delta t} \int_0^{\Delta t} \mathbf{f}(\mathbf{w}(x_{j+\frac{1}{2}}, t)) dt = \mathbf{f}\left(\mathbf{w}\left(x_{j+\frac{1}{2}}, \frac{\Delta t}{2}\right)\right) + \mathcal{O}(\Delta t^2).$$

As in the scalar case (G.R., Chapter 4, Sections 3.2 and 3.3), we now give two ways for approximating  $\mathbf{f}(\mathbf{w}(x_{j+\frac{1}{2}}, \frac{\Delta t}{2}))$ .

(1) One can use the asymptotic expansion of the solution of the G.R.P. (6.2) associated with the piecewise linear data (6.1),

$$\mathbf{w}(x, t) = \mathbf{u}^0\left(\frac{x - x_{j+\frac{1}{2}}}{t} r\right) + t \mathbf{u}^1\left(\frac{x - x_{j+\frac{1}{2}}}{t}\right) + \mathcal{O}(t^2),$$

where the derivation of  $\mathbf{u}^0$  and  $\mathbf{u}^1$  will be detailed in the next section (in particular,  $\mathbf{u}^0$  is a solution of a classical Riemann problem). It implies

$$\mathbf{w}(x_{j+\frac{1}{2}}, t) = \mathbf{u}^0(0) + t \mathbf{u}^1(0) + \mathcal{O}(t^2) = \mathbf{w}^0(x_{j+\frac{1}{2}}) + t \mathbf{w}^1(x_{j+\frac{1}{2}}) + \mathcal{O}(t^2), \quad (6.5)$$

and the scheme reads

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n + \frac{\Delta t}{\Delta x_j} \{\mathbf{g}_{j+\frac{1}{2}}^n - \mathbf{g}_{j-\frac{1}{2}}^n\}.$$

where the numerical flux is given by  $\mathbf{g}_{j+\frac{1}{2}}^n = \mathbf{f}(\mathbf{u}^0(0) + \frac{\Delta t}{2} \mathbf{u}^1(0))$ , i.e.,

$$\mathbf{g}_{j+\frac{1}{2}}^n = \mathbf{f}\left(\mathbf{w}^0(x_{j+\frac{1}{2}}) + \frac{\Delta t}{2} \mathbf{w}^1(x_{j+\frac{1}{2}})\right). \quad (6.6)$$

The argument may be double-valued, but the flux is single-valued. In fact, we can take an approximation, for example,

$$\mathbf{f}\left(\mathbf{u}^0(0) + \frac{\Delta t}{2} \mathbf{u}^1(0)\right) \approx \mathbf{f}(\mathbf{u}^0(0)) + \frac{\Delta t}{2} \mathbf{f}'(\mathbf{u}^0(0)) \mathbf{u}^1(0),$$

or any other more convenient approximation. In the next section, for example, we shall use for the third component  $pu$  of  $\mathbf{f}$

$$(pu)^0 + t(pu)^1 = p^0 u^0 + t(p^0 u^1 + p^1 u^0) \approx (p^0 + tp^1)(u^0 + tu^1).$$

(2) One uses a predictor–corrector scheme. Following an idea of Hancock, we define the updated values  $\mathbf{v}_{j+\frac{1}{2},\pm}^{n+\frac{1}{2}}$  at time  $t_n + \frac{\Delta t}{2}$  by

$$\begin{cases} \mathbf{v}_{j+\frac{1}{2},-}^{n+\frac{1}{2}} = \mathbf{v}_{j+\frac{1}{2},-}^n - \frac{\Delta t}{2\Delta x_j} (\mathbf{f}(\mathbf{v}_{j+\frac{1}{2},-}^n) - \mathbf{f}(\mathbf{v}_{j-\frac{1}{2},+}^n)), \\ \mathbf{v}_{j+\frac{1}{2},+}^{n+\frac{1}{2}} = \mathbf{v}_{j+\frac{1}{2},+}^n - \frac{\Delta t}{2\Delta x_j} (\mathbf{f}(\mathbf{v}_{j+\frac{3}{2},-}^n) - \mathbf{f}(\mathbf{v}_{j+\frac{1}{2},+}^n)), \end{cases}$$

where

$$\begin{cases} \mathbf{v}_{j+\frac{1}{2},-}^n = \tilde{\mathbf{v}}(x_{j+\frac{1}{2}} - 0) = \mathbf{v}_j^n + \frac{\mathbf{S}_j^n}{2}, \\ \mathbf{v}_{j+\frac{1}{2},+}^n = \tilde{\mathbf{v}}(x_{j+\frac{1}{2}} + 0) = \mathbf{v}_{j+1}^n - \frac{\mathbf{S}_{j+1}^n}{2}. \end{cases}$$

Then, we solve the Riemann problem at the point  $x_{j+\frac{1}{2}}$  with piecewise constant initial data  $\mathbf{v}_{j+\frac{1}{2},\pm}^{n+\frac{1}{2}}$

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = \mathbf{0}, \\ \mathbf{u}(x, 0) = \begin{cases} \mathbf{v}_{j+\frac{1}{2},-}^{n+\frac{1}{2}}, & x < x_{j+\frac{1}{2}}, \\ \mathbf{v}_{j+\frac{1}{2},+}^{n+\frac{1}{2}}, & x > x_{j+\frac{1}{2}}, \end{cases} \end{cases}$$

whose solution is noted as usual  $\mathbf{w}_R\left(\frac{x-x_{j+\frac{1}{2}}}{t}; \mathbf{v}_{j+\frac{1}{2},-}^{n+\frac{1}{2}}, \mathbf{v}_{j+\frac{1}{2},+}^{n+\frac{1}{2}}\right)$ . We replace

$$\mathbf{w}\left(x_{j+\frac{1}{2}}, \frac{\Delta t}{2}\right)$$

by

$$\mathbf{w}_R(0; \mathbf{v}_{j+\frac{1}{2},-}^{n+\frac{1}{2}}, \mathbf{v}_{j+\frac{1}{2},+}^{n+\frac{1}{2}}),$$

and thus we take

$$\mathbf{g}_{j+\frac{1}{2}}^n = \mathbf{f}\left(\mathbf{w}_R(0; \mathbf{v}_{j+\frac{1}{2},-}^{n+\frac{1}{2}}, \mathbf{v}_{j+\frac{1}{2},+}^{n+\frac{1}{2}})\right). \quad (6.7)$$

*Remark 6.1.* The updated values  $\mathbf{v}_{j+\frac{1}{2},\pm}^{n+\frac{1}{2}}$  can be interpreted as follows:

$$\mathbf{v}_{j+\frac{1}{2},\pm}^{n+\frac{1}{2}} \approx \mathbf{w}\left(x_{j+\frac{1}{2}} \pm 0, \frac{\Delta t}{2}\right),$$

and  $\mathbf{w}(x_{j+\frac{1}{2}} \pm 0, \frac{\Delta t}{2})$  is evaluated by

$$\begin{cases} \mathbf{w}(x_{j+\frac{1}{2}} \pm 0, \frac{\Delta t}{2}) \approx \mathbf{w}(x_{j+\frac{1}{2}} \pm 0, 0) + \frac{\Delta t}{2} \frac{\partial \mathbf{u}}{\partial t} \\ = \mathbf{w}(x_{j+\frac{1}{2}} \pm 0, 0) - \frac{\Delta t}{2} \frac{\partial \mathbf{f}(\mathbf{u})}{\partial x}. \end{cases}$$

Using a finite difference approximation for  $\frac{\partial \mathbf{f}(\mathbf{u})}{\partial x}$  yields

$$\begin{aligned} \mathbf{w}\left(x_{j+\frac{1}{2}} - 0, \frac{\Delta t}{2}\right) &\approx \mathbf{v}_{j+\frac{1}{2},-}^n - \frac{\Delta t}{2\Delta x_j} \left( \mathbf{f}(\mathbf{v}_{j+\frac{1}{2},-}^n) - \mathbf{f}(\mathbf{v}_{j-\frac{1}{2},+}^n) \right), \\ \mathbf{w}\left(x_{j+\frac{1}{2}} + 0, \frac{\Delta t}{2}\right) &\approx \mathbf{v}_{j+\frac{1}{2},+}^n - \frac{\Delta t}{2\Delta x_j} \left( \mathbf{f}(\mathbf{v}_{j+\frac{3}{2},-}^n) - \mathbf{f}(\mathbf{v}_{j+\frac{1}{2},+}^n) \right). \end{aligned}$$

The last step corresponds to applying Godunov's scheme to these updated values.  $\square$

The reconstruction step consists then in defining new slopes  $\mathbf{S}_j^{n+1}$ . We also proceed by prediction–correction. We can use, for instance, central differencing to define the predictor

$$\hat{\mathbf{S}}_j^{n+1} = \Delta x_j \frac{\mathbf{v}_{j+1}^{n+1} - \mathbf{v}_{j-1}^{n+1}}{x_{j+1} - x_{j-1}}. \quad (6.8)$$

Otherwise, if we follow the first approach (1),  $\mathbf{S}_j^{n+1}$  can be predicted by

$$\hat{\mathbf{S}}_j^{n+1} = \mathbf{w}(x_{j+\frac{1}{2}} - 0, \Delta t) - \mathbf{w}(x_{j-\frac{1}{2}} + 0, \Delta t). \quad (6.9)$$

where  $\mathbf{w}(x_{j+\frac{1}{2}} \pm 0, \Delta t)$  is computed from the first terms of the expansion (6.5),

$$\mathbf{u}(0 \pm, \Delta t) \approx \mathbf{u}^0(0 \pm) + \Delta t \mathbf{u}^1(0 \pm) = \mathbf{w}^0(x_{j+\frac{1}{2}} \pm 0) + \Delta t \mathbf{w}^1(x_{j+\frac{1}{2}} \pm 0),$$

or even more simply by retaining the first-order term of the expansion (which corresponds to the solution of a classical Riemann problem),

$$\hat{\mathbf{S}}_j^{n+1} = \mathbf{w}^0(x_{j+\frac{1}{2}} - 0, \Delta t) - \mathbf{w}^0(x_{j-\frac{1}{2}} + 0, \Delta t).$$

In the second approach,

$$\hat{\mathbf{S}}_j^{n+1} = v_{j+\frac{1}{2},+}^{n+\frac{1}{2}} - v_{j+\frac{1}{2},-}^{n+\frac{1}{2}}.$$

In the correction steps, we limit the slope componentwise following formula (3.11) in G.R., Chapter 4, which ensures that  $\tilde{\mathbf{v}}^{n+1}(x_{j \pm \frac{1}{2}})$  does not take values outside the range spanned by the neighboring mesh averages. Omitting the time superscript for each component  $\delta v_{k,j}$  of  $\mathbf{S}_j$ ,  $k = 1, \dots, p$ , we impose

$$\delta v_{k,j} = \begin{cases} s \min\{2|\Delta v_{k,j-\frac{1}{2}}|, |\hat{\delta}v_{k,j}|, 2|\Delta v_{k,j+\frac{1}{2}}|\} \\ \text{if } s = \operatorname{sgn} \Delta v_{k,j-\frac{1}{2}} = \operatorname{sgn} \Delta v_{k,j+\frac{1}{2}} = \operatorname{sgn} \hat{\delta}v_{k,j}, \\ 0 \quad \text{otherwise.} \end{cases} \quad (6.10)$$

There are other possibilities for limiting the slope that are not so easy to interpret. However, this correction is very important in practice.

## 6.2 Solution of the Generalized Riemann Problem

The first method (6.6) involves a generalized Riemann problem (G.R.P.)

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = \mathbf{0}, \\ \mathbf{u}(x, 0) = \begin{cases} \mathbf{u}_L(x), & x < 0, \\ \mathbf{u}_R(x), & x > 0. \end{cases} \end{cases} \quad (6.11)$$

The solution is not self-similar, since the discontinuity curves are no longer straight lines, but it still consists of a sequence of shocks, rarefactions, and contact discontinuities (see Fig. 6.1).

Following the arguments of G.R., Chapter 4, Section 3.3, we can use an asymptotic expansion near the origin in order to derive an approximate solution, which is obtained as a perturbation of the (classical) Riemann problem

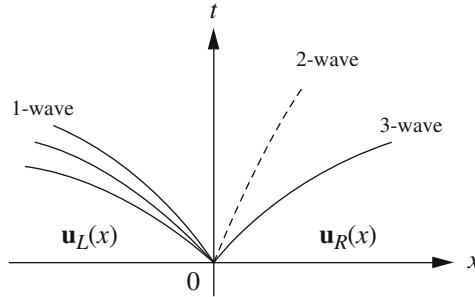
$$\begin{cases} \frac{\partial \mathbf{u}^0}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}^0) = \mathbf{0}, \\ \mathbf{u}^0(x, 0) = \begin{cases} \mathbf{u}_L(0), & x < 0, \\ \mathbf{u}_R(0), & x > 0, \end{cases} \end{cases} \quad (6.12)$$

Indeed, we have

$$\mathbf{u}(x, t) = \mathbf{u}^0\left(\frac{x}{t}\right) + t\mathbf{u}^1\left(\frac{x}{t}\right) + \mathcal{O}(t^2), \quad (6.13)$$

where  $\mathbf{u}^0(\xi) = \mathbf{w}_R(\xi; \mathbf{u}_L(0), \mathbf{u}_R(0))$  is the solution of (6.12) and the computation of  $\mathbf{u}^1$  is less straightforward. The derivation of  $\mathbf{u}^0$  follows essentially the lines of the scalar case. We shall sketch the main ideas of the derivation of  $\mathbf{u}^1$ . We determine  $\mathbf{u}^1$  in the regions of smoothness of  $\mathbf{u}$  and its jumps across the transition curves that separate these regions (see Fig. 6.1) as well as the conditions at infinity. First setting  $\xi = \frac{x}{t}$ , we obtain easily that  $\mathbf{u}^1$  satisfies, in each smoothness region,

$$-\xi \frac{d\mathbf{u}^1}{d\xi} + \frac{d}{d\xi} (\mathbf{A}(\mathbf{u}^0)\mathbf{u}^1) + \mathbf{u}^1 = \mathbf{0}. \quad (6.14)$$



**Fig. 6.1** Solution of the G.R.P.

In order to determine the jump conditions satisfied by  $\mathbf{u}^1$ , let us consider a discontinuity curve  $\sum$  with equation

$$\sigma(t) = \sigma_0 t + \frac{\sigma_1 t^2}{2} + \mathcal{O}(t^3)$$

separating two smoothness regions of  $\mathbf{u}$ . On the one hand, if  $\mathbf{u}^0$  is continuous across  $\sum$ ,

$$[\mathbf{u}^1] + \sigma_1 \left[ \frac{d\mathbf{u}^0}{d\xi} \right] = \mathbf{0}. \quad (6.15)$$

On the other hand, if  $\mathbf{u}^0$  is discontinuous,

$$[(\mathbf{A}(\mathbf{u}^0) - \sigma_0)\mathbf{u}^1] = \sigma_1 [\mathbf{u}^0]. \quad (6.16)$$

The conditions as  $\xi \rightarrow \pm\infty$  are derived as in the scalar case,

$$\lim_{\xi \rightarrow -\infty} \frac{d\mathbf{u}^1}{d\xi} = \frac{d\mathbf{u}_L}{d\xi}(0), \quad \lim_{\xi \rightarrow +\infty} \frac{d\mathbf{u}^1}{d\xi} = \frac{d\mathbf{u}_R}{d\xi}(0). \quad (6.17)$$

### Proposition 6.1

If  $|\mathbf{u}_R(0) - \mathbf{u}_L(0)|$  is small enough, there exists a unique function  $\mathbf{u}^1$  that satisfies (6.14) in the domains of smoothness of  $\mathbf{u}$  together with the conditions (6.15)–(6.17).

*Proof.* Let us write  $\mathbf{u}^1$  in the basis  $(\mathbf{r}_k(\mathbf{u}^0))$ ,

$$\mathbf{u}^1(\xi) = \sum_{k=1}^p \alpha_k(\xi) \mathbf{r}_k(\mathbf{u}^0). \quad (6.18)$$

We study successively the regions where  $\mathbf{u}^0$  is constant, those where  $\mathbf{u}^0$  is smooth, and then the discontinuities of  $\mathbf{u}^0$ .

- (i) Outside the rarefaction waves and the discontinuity curves,  $\mathbf{u}^0$  is constant, and the system (6.14) becomes

$$\mathbf{u}^1 + (\mathbf{A}(\mathbf{u}^0) - \xi) \frac{d\mathbf{u}^1}{d\xi} = \mathbf{0}.$$

Thus, we get

$$\alpha_k + (\lambda_k(\mathbf{u}^0) - \xi)\alpha'_k = 0,$$

and we can solve each scalar equation as in the scalar case,

$$\alpha_k = a_k(\xi - \lambda_k(\mathbf{u}^0)), \quad (6.19)$$

where  $a_k$  is a constant. We obtain that  $\mathbf{u}^1$  is an *affine* function of  $\xi$  in each region of smoothness where  $\mathbf{u}^0$  is constant.

- (ii) If  $\mathbf{u}^0$  contains a  $k$ -rarefaction wave, we have

$$\begin{cases} \frac{d}{d\xi} \mathbf{u}^0(\xi) = \mathbf{r}_k(\mathbf{u}^0(\xi)), \\ \xi = \lambda_k(\mathbf{u}^0(\xi)), \end{cases}$$

and (6.14) becomes

$$(\mathbf{A}(\mathbf{u}^0) - \xi) \frac{d\mathbf{u}^1}{d\xi} + \left( \frac{d}{d\xi} (\mathbf{A}(\mathbf{u}^0)) \right) \mathbf{u}^1 + \mathbf{u}^1 = \mathbf{0}. \quad (6.20)$$

We obtain again a linear differential system in the  $\alpha_i$  (which can be solved explicitly in the case of the gas dynamics equations, as we shall see below). In particular, for the  $k$ th component, the first term in (6.20) cancels, and we get a pure algebraic equation,

$$\alpha_k + \sum_j \beta_{jk}(\mathbf{u}^0) \alpha_j = 0, \quad (6.21)$$

where the  $\beta_{j\ell}$  are defined by

$$\frac{d}{d\xi} (\mathbf{A}(\mathbf{u}^0)) \mathbf{r}_j(\mathbf{u}^0) = \sum_\ell \beta_{j\ell}(\mathbf{u}^0) \mathbf{r}_\ell(\mathbf{u}^0).$$

(It is easy to prove that  $\beta_{kk} = 1$ .)

- (iii) Now, assume that  $\mathbf{u}^0$  is continuous at a point  $\xi = \sigma_0$  but  $\frac{d\mathbf{u}^0}{d\xi}$  is discontinuous: this corresponds to the boundary of a rarefaction wave. By (6.15), we get for a  $k$ -rarefaction

$$\mathbf{u}^1(\sigma_0+) - \mathbf{u}^1(\sigma_0-) = \pm \sigma_1 \mathbf{r}_k(\mathbf{u}^0(\sigma_0)), \quad (6.22a)$$

where the sign + (resp.  $-$ ) holds if the constant state is located on the right-hand side (resp. left) of the rarefaction wave. This shows that  $\alpha_k$  is discontinuous at  $\sigma_0$  but

$$\alpha_i, \quad k \neq i, \text{ is continuous at } \xi = \sigma_0. \quad (6.22b)$$

- (iv) Then, if  $\mathbf{u}^0$  is discontinuous at a point  $\xi = \sigma_0$ , which corresponds to a shock wave or a contact discontinuity, (6.17) gives a system of  $p$  equations in the  $(2p + 1)$  variables  $\alpha_k(\sigma_0 \pm)$  and  $\sigma_1$ .
- (v) Finally, following the arguments of the scalar case, the conditions at  $\pm\infty$  (6.18) give for  $-\xi$  large enough

$$\mathbf{u}^1(\xi) = (\xi - \mathbf{A}(\mathbf{u}_L(0))) \frac{d\mathbf{u}_L}{d\xi}(0), \quad (6.23a)$$

and for  $\xi$  large enough

$$\mathbf{u}^1(\xi) = (\xi - \mathbf{A}(\mathbf{u}_R(0))) \frac{d\mathbf{u}_R}{d\xi}(0). \quad (6.23b)$$

We then check that (6.17) and (6.22)–(6.23) determine  $\mathbf{u}^1$  in a unique way (see Bourgeade et al. [184, 756]).  $\square$

*Remark 6.2.* The theory above applies to a system with a source term

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{F}(x, t, \mathbf{u}) = \mathbf{G}(x, t, \mathbf{u});$$

we refer again to Bourgeade et al. [184].  $\square$

### 6.3 The G.R.P. for the Gas Dynamics Equations in Lagrangian Coordinates

We start from Eqs. (2.25)

$$\begin{cases} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial p}{\partial m} = 0, \\ \frac{\partial e}{\partial t} + \frac{\partial}{\partial m}(pu) = 0, \end{cases}$$

with  $p = p(\tau, \varepsilon) = p(\tau, e - \frac{u^2}{2})$ , which can be written in the general form

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial}{\partial m} \mathbf{F}(\mathbf{U}) = 0, \text{ where}$$

$$\mathbf{U} = (\tau, u, e)^T, \quad \mathbf{F}(\mathbf{U}) = (-u, p, pu)^T.$$

We shall just sketch the resolution of the G.R.P. The solution of the classical Riemann problem is detailed in Chap. III, Sect. 3, and we may suppose that  $u^0, p^0, \tau_\pm^0$ , and  $s_\pm^0$  are determined explicitly. We must then compute  $\mathbf{u}^1$ . We solve Eqs. (6.17)–(6.21), i.e., we determine the coefficients  $\alpha_k$  of  $\mathbf{u}^1(\xi)$  in the basis  $(\mathbf{r}_k(\mathbf{u}^0))$ . In fact, we use the nonconservative variables  $\mathbf{v} = (\tau, u, s)^T$  (Chap. II, Example 2.3) with  $p = p(\tau, s)$  since the formulas are easier to handle. Denoting for ease of notation

$$g = g(\mathbf{v}) = \sqrt{-\frac{\partial p}{\partial \tau}} = \frac{c}{\tau}, \quad (6.24a)$$

where  $p = p(\tau, s)$  and

$$q = -\frac{\left(\frac{\partial p}{\partial \tau}\right)}{\left(\frac{\partial p}{\partial s}\right)}, \quad (6.24b)$$

the system (2.25) can be written in nonconservative form

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{B}(\mathbf{v}) \frac{\partial \mathbf{v}}{\partial m} = \mathbf{0},$$

where the matrix  $\mathbf{B}$  is given by

$$\mathbf{B} = \begin{pmatrix} 0 & -1 & 0 \\ -g^2 & 0 & g^2/q \\ 0 & 0 & 0 \end{pmatrix}.$$

The contact discontinuity in Lagrangian coordinates corresponds to  $\xi = \lambda_2(\mathbf{v}) = 0$ , while the other eigenvalues are

$$\lambda_1(\mathbf{v}) = -g, \quad \lambda_3(\mathbf{v}) = g,$$

and the eigenvectors can be chosen as

$$\mathbf{r}_1(\mathbf{v}) = (1, g, 0)^T, \quad \mathbf{r}_2(\mathbf{v}) = (1, 0, q)^T, \quad \mathbf{r}_3(\mathbf{v}) = (1, -g, 0)^T.$$

We shall assume for simplicity an equation of state of Grüneisen type, of the form (4.1)

$$p = (\gamma - 1) \frac{\varepsilon}{\tau} + c_{\text{ref}}^2 \frac{(\tau_{\text{ref}} - \tau)}{\tau_{\text{ref}} \tau}.$$

Then we compute

$$\frac{\partial p(\tau, s)}{\partial \tau} = -\frac{c^2}{\tau^2} = -\gamma(p + p_\infty)/\tau,$$

where  $p_\infty$  is defined as previously by  $p_\infty = c_{\text{ref}}^2 \rho_{\text{ref}} / \gamma$ . Thus  $(p + p_\infty) \tau^\gamma$  is a function of  $s$  only and we can take

$$p + p_\infty = \frac{s}{\tau^\gamma}.$$

Then, we have

$$g^2 = \gamma \frac{s}{\tau^{\gamma+1}}, \quad q = \gamma \frac{s}{\tau}.$$

Given  $\mathbf{v}^0 = (\tau_0, u_0, s_0)^T$  a solution of the classical Riemann problem, the coefficients  $\alpha_k(\mathbf{v})$  of  $\mathbf{v} = (\tau, u, s)^T$  in the basis  $(\mathbf{r}_k(\mathbf{v}^0))$  defined by

$$\mathbf{v} = \sum_{k=1}^3 \alpha_k(\mathbf{v}) \mathbf{r}_k(\mathbf{v}^0) \quad (6.25a)$$

satisfy

$$\begin{cases} \tau = \alpha_1 + \alpha_2 + \alpha_3, \\ u = (\alpha_1 - \alpha_3) g^0, \\ s = q^0 \alpha_2, \end{cases} \quad (6.25b)$$

or, equivalently,

$$\begin{cases} \alpha_1 = \frac{1}{2} \left( \frac{u}{g^0} + \tau - \frac{s}{q^0} \right), \\ \alpha_2 = \frac{s}{q^0}, \\ \alpha_3 = \frac{1}{2} \left( \frac{-u}{g^0} + \tau - \frac{s}{q^0} \right). \end{cases} \quad (6.25c)$$

Following the ideas of the preceding section, we determine the coefficients  $\alpha_k(\xi)$  of  $\mathbf{v}^1$  in each zone.

- (i) Definition of the coefficients  $a_{K,i}$ ; computation of  $a_{0,k}$  and  $a_{III,k}$ . In each region  $K$  where  $\mathbf{v}^0$  is constant, we know from (6.22) that the  $\alpha_k$  are affine and more precisely that there exist constants  $a_{K,i}$ ,  $i = 1, 2, 3$ , such that

$$\begin{cases} \alpha_1(\xi) = a_{K,1}(\xi + g^0), \\ \alpha_2(\xi) = a_{K,2}\xi, \\ \alpha_3(\xi) = a_{K,3}(\xi - g^0). \end{cases} \quad (6.26)$$

Assume for specificity that the (classical) Riemann problem corresponds to case 3 of Fig. 3.8, Chap. III (1-rarefaction, 3-shock). In Lagrangian coordinates, we have a centered rarefaction wave propagating to the left and a shock traveling to the right. The zone  $K = 0$  corresponds to the initial condition  $\mathbf{v}_L^0 = \mathbf{v}_L(0)$  and is bordered on the right by the line  $\xi = -g_L^0 = -g(\mathbf{v}_L^0)$ ; zone  $I$  lies between the tail of the rarefaction wave  $\xi = -g(\mathbf{v}_I^0) = -g_I^0$  and the contact discontinuity  $x = 0$ , from which in

turn starts zone  $II$ , which ends at  $\xi = \sigma_3^0$  (the shock wave), and zone  $III$  corresponds to the right state  $\mathbf{v}_R^0 = \mathbf{v}_R(0)$  (see Fig. 6.2). Now, defining the coefficients  $\alpha_{L,k}$  and  $\alpha_{R,k}$  by

$$\mathbf{v}_L^1 = \frac{d\mathbf{v}_L}{dx}(0) = \sum_{k=1}^3 \alpha_{L,k} \mathbf{r}_k(\mathbf{v}_L^0), \quad \mathbf{v}_L^0 = \mathbf{v}_L(0), \quad (6.27a)$$

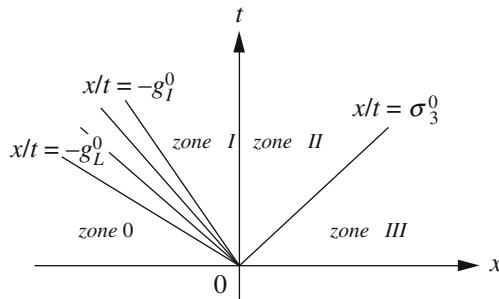
$$\mathbf{v}_R^1 = \frac{d\mathbf{v}_R}{dx}(0) = \sum_{k=1}^3 \alpha_{R,k} \mathbf{r}_k(\mathbf{v}_R^0), \quad \mathbf{v}_R^0 = \mathbf{v}_R(0), \quad (6.27b)$$

and using the conditions (6.18) at  $\xi = \pm\infty$ , we have

$$a_{0,k} = \alpha_{L,k}, \quad a_{III,k} = \alpha_{R,k}, \quad k = 1, 2, 3, \quad (6.28)$$

which determines completely the coefficients  $\alpha_k$  of  $\mathbf{v}^1$  in zone 0 and in zone III. The other coefficients will be determined below.

- (ii) Determination of  $\alpha_2(\xi)$  in the rarefaction; computation of  $a_{I,2}$ . Between zones 0 and I (i.e., in the 1-rarefaction that is bordered by the lines  $\xi = -g_L^0$  and  $\xi = -g_I^0$ ), we solve the system (6.23), which consists of one algebraic equation giving  $\alpha_1(\xi)$  in terms of  $\alpha_2(\xi)$  and  $\alpha_3(\xi)$  and a system of two linear differential equations for  $\alpha_2(\xi)$  and  $\alpha_3(\xi)$ . For an equation of state of the form (4.1) and taking into account the jump conditions (6.21), we can solve explicitly the differential system. We use the fact that for a rarefaction



**Fig. 6.2** The four zones in the Riemann problem

$$\frac{d}{d\xi} s^0(\xi) = 0, \quad g^0(\xi) = -\lambda_1(\mathbf{v}(\xi)) = \xi.$$

Then, we note that by (6.22b),  $\alpha_2(\xi)$  and  $\alpha_3(\xi)$  are continuous across the line  $\xi = -g_L^0$ ; hence the coefficients  $\alpha_{L,k} = a_{0,k}$  appear, and for  $-g_L^0 < \xi < -g_I^0$ , we get

$$\begin{cases} \alpha_1(\xi) = \frac{\alpha_2(\xi)}{2(\gamma+1)} - \frac{\alpha_3(\xi)}{2}, \\ \alpha_2(\xi) = -\left(\frac{-\xi}{g_L^0}\right)^{\frac{(\gamma-1)}{(\gamma+1)}} g_L^0 \alpha_{L,2}, \\ \alpha_3(\xi) = \frac{g_L^0}{3\gamma-1} \left(\frac{-g_L^0}{\xi}\right)^{\frac{1}{2}} \left\{ -1 + \left(\frac{-\xi}{g_L^0}\right)^{\frac{3\gamma-1}{2(\gamma+1)}} \right\} \alpha_{L,2} \\ \quad - 2g_L^0 \left(\frac{-g_L^0}{\xi}\right)^{\frac{1}{2}} \alpha_{L,3}. \end{cases} \quad (6.29)$$

$\alpha_2(\xi)$  and  $\alpha_3(\xi)$  are also continuous across the tail of the 1-rarefaction  $\xi = -g_I^0$ , i.e.,

$$\alpha_j(-g_I^0-) = \alpha_j(-g_I^0+), \quad j = 2, 3.$$

In particular,  $\alpha_2$  is continuous at  $\xi = -g_I^0$ . Hence, equating the corresponding expressions (6.26) and (6.29), we get

$$-a_{I,2}g_I^0 = \alpha_2(-g_I^0) = -\left(\frac{g_I^0}{g_L^0}\right)^{\frac{(\gamma-1)}{(\gamma+1)}} g_L^0 \alpha_{L,2},$$

or

$$a_{I,2} = \left(\frac{g_I^0}{g_L^0}\right)^{\frac{-2}{(\gamma+1)}} \alpha_{L,2} = \left(\frac{q_L^0}{q_I^0}\right) \alpha_{L,2}. \quad (6.30)$$

- (iii) Definition of  $p^1(\xi)$ ; continuity of  $p^1(\xi)$  at the contact discontinuity. Now, we give an easy though important result concerning the contact discontinuity: defining  $p^1(\xi)$  by

$$p^1 = \left(\frac{\partial p}{\partial \tau}\right)^0 \tau^1 + \left(\frac{\partial p}{\partial s}\right)^0 s^1, \quad (6.31)$$

we have

*Lemma 6.1*

The functions  $u^1$  and  $p^1$  are continuous at  $\xi = 0$ .

*Proof.* Recall that the contact discontinuity is characterized by

$$[u] = [p] = 0.$$

Using

$$u = u^0 + t u^1 + O(t^2), \quad p = p^0 + t p^1 + O(t^2)$$

and identifying the terms of order 0,1 yields

$$[u^0] = [p^0] = 0$$

and

$$[u^1] = [p^1] = 0,$$

which gives the result.  $\square$

We shall set

$$\begin{aligned} u_*^0 &= u^0(0), & p_*^0 &= p^0(0), \\ u_*^1 &= u^1(0), & p_*^1 &= p^1(0). \end{aligned}$$

- (iv) The jump conditions. For the 3-shock separating zones *II* and *III*, in order to specify the jump conditions, we introduce again the functions  $\tau = h_a(p) = h_a(\tau_a, u_a; p)$  and  $\Phi_a(p) = \Phi(\tau_a, u_a; p)$  of (3.2) and (3.4), Chap. III, Sect. 3. If the state  $\mathbf{u} = (\tau, u, p)$  is connected to the state  $\mathbf{u}_a = (\tau_a, u_a, p_a)$ , we write

$$\tau_a = \tau_a^0 + t\tau_a^1 + \cdots, \quad u_a = u_a^0 + tu_a^1 + \cdots$$

$$\tau^0 + t\tau^1 + \cdots = h(\tau_a, u_a; p^0 + tp^1 + \cdots)$$

and

$$u^0 + tu^1 + \cdots = u_a^0 + tu_a^1 + \cdots + \Phi(\tau_a, u_a; p^0 + tp^1 + \cdots).$$

This gives

$$u^1 = u_a^1 + \left\{ \left( \frac{\partial \Phi}{\partial \tau_a} \right)^0 \tau_a^1 + \left( \frac{\partial \Phi}{\partial p_a} \right)^0 p_a^1 + \left( \frac{\partial \Phi}{\partial p} \right)^0 p^1 \right\}, \quad (6.32a)$$

where the derivatives of  $\Phi$  are computed at  $(\tau_a^0, u_a^0, p^0)$  and a similar formula for  $\tau^1$

$$\tau^1 = \left( \frac{\partial h}{\partial \tau_a} \right)^0 \tau_a^1 + \left( \frac{\partial h}{\partial p_a} \right)^0 p_a^1 + \left( \frac{\partial h}{\partial p} \right)^0 p^1. \quad (6.32b)$$

These formulas can be made explicit for the equation of state (4.1). We have

$$\begin{cases} \frac{\partial \Phi}{\partial \tau_a}(p) = \frac{u - u_a}{2\tau_a}, \\ \frac{\partial \Phi}{\partial p_a}(p) = -W_a(p) - \frac{\mu^2}{2} \frac{(u - u_a)}{p + p_\infty + \mu^2(p_a + p_\infty)}, \\ \frac{\partial \Phi}{\partial p}(p) = W_a(p) - \frac{(u - u_a)}{2(p + p_\infty + \mu^2(p_a + p_\infty))}, \end{cases} \quad (6.33)$$

where  $\mu^2 = \frac{(\gamma-1)}{(\gamma+1)}$  and  $W_a(p)^2 = \frac{(1-\mu^2)\tau_a}{p+p_\infty+\mu^2(p_a+p_\infty)}$ . We shall apply this formula below with  $\mathbf{u} = \mathbf{u}(\sigma_3^-)$  and  $\mathbf{u}_a = \mathbf{u}(\sigma_3^0)$ .

- (v) Computation of  $u_*^1, p_*^1$ . These values of  $u^1$  and  $p^1$  at  $\xi = 0$  will in turn be used to determine the coefficients  $a_{K,i}$ ,  $K = I, II$ ; they will also be needed in the following section. In the case of the classical Riemann problem,  $u_*^0, p_*^0$  are determined by the intersection of two curves. Similarly,  $u_*^1, p_*^1$  are found as the solution of an algebraic system of two equations. Each equation is obtained by working in the zones  $I$  and  $II$ , respectively. We give first a technical result.

*Lemma 6.2*

Let  $\xi \in \text{zone } K$ ,  $\mathbf{v}^1(\xi) = (\tau^1(\xi), u^1(\xi), s^1(\xi))^T = \sum \alpha_k \mathbf{r}_k(\mathbf{v}^0)$ , and  $p^1(\xi)$  be defined by (6.31). We have

$$\begin{aligned}\tau^1(\xi) &= (a_{K,1} + a_{K,2} + a_{K,3})\xi + g^0(a_{K,1} - a_{K,3}), \\ u^1(\xi) &= g^0\{\xi(a_{K,1} - a_{K,3}) + (a_{K,1} + a_{K,3})\}, \\ p^1(\xi) &= -(g^0)^2(a_{K,1} + a_{K,3})\xi,\end{aligned}$$

where the coefficients  $a_{K,i}$  are defined in (6.26).

*Proof.* For any  $\mathbf{v}^1 = (\tau^1, u^1, s^1)^T = \sum \alpha_K \mathbf{r}_k(\mathbf{v}^0)$ , we have by (7.25)

$$\alpha_1 + \alpha_2 + \alpha_3 = \tau^1.$$

If, moreover, the  $\alpha_k$  are given by (6.26), we get

$$\tau^1(\xi) = (a_{K,1} + a_{K,2} + a_{K,3})\xi + g^0(a_{K,1} - a_{K,3}).$$

Next, by (6.25b) and (6.26),

$$\alpha_1 - \alpha_3 = \frac{u^1}{g^0},$$

and hence

$$u^1(\xi) = g^0\{\xi(a_{K,1} - a_{K,3}) + (a_{K,1} + a_{K,3})\}.$$

Then

$$\alpha_1 + \alpha_3 = \tau^1 - \frac{s^1}{q^0}.$$

By (6.24a) ( $q = -\frac{p_\tau}{p_s}$ ,  $g^2 = -p_\tau$ ) and (6.31), we have the identity

$$\tau^1 - \frac{s^1}{q^0} = -(g^0)^{-2} \left\{ \left( \frac{\partial p}{\partial \tau} \right)^0 \tau^1 + \left( \frac{\partial p}{\partial s} \right)^0 s^1 \right\} = -\frac{p^1}{(g^0)^2}. \quad (6.34)$$

Thus

$$\alpha_1 + \alpha_3 = -\frac{p^1}{(g^0)^2}$$

and

$$p^1(\xi) = -(g^0)^2(a_{K,1} + a_{K,3})\xi,$$

which ends the proof.  $\square$

In order to obtain the system of equations defining  $(u_*^1, p_*^1)$ , we consider the zones  $I$  and  $II$ . There the functions  $\alpha_k$  are affine, and the values at the boundary  $\xi = 0$  – (resp.  $\xi = 0+$ ) are written in terms of the values at the other boundary of the zone  $\xi = -g_I^0$  (resp.  $\xi = (\sigma_3^0 -)$ ). More precisely, in zone  $I$ , by definition (6.25c) of  $\alpha_3$ , setting  $\mathbf{v}^1(0-) = (\tau_I^1, u_*^1, s_I^1)$ , we get

$$\alpha_3(0-) = \frac{1}{2} \left( -\frac{u_*^1}{g_I^0} + \tau_I^1 - \frac{s_I^1}{q_I^0} \right).$$

We have, by (6.34),

$$\tau_I^1 - \frac{s_I^1}{q_I^0} = -\frac{p_*^1}{(g_I^0)^2}.$$

Hence

$$\alpha_3(0-) = -\frac{1}{2} \left( \frac{u_*^1}{g_I^0} + \frac{p_*^1}{(g_I^0)^2} \right).$$

Then, on the one hand, we know by (6.26) that since we are in zone  $I$ ,

$$\alpha_3(0-) = -a_{I,3}g_I^0$$

and

$$\alpha_3(-g_I^0+) = -2a_{I,3}g_I^0.$$

On the other hand, due to (6.29) and the continuity of  $\alpha_3$ , we have another expression for  $\alpha_3(-g_I^0+) = \alpha_3(-g_I^0-)$ , which enables us to determine the coefficients  $a_{I,3}$  and leads thus to a first equation in  $u_*^1, p_*^1$ . We have

$$u_*^1 + \frac{p_*^1}{g_I^0} = C_L, \quad (6.35)$$

where

$$C_L = (g_L^0 g_I^0)^{\frac{1}{2}} \left\{ -u_L^1 + \frac{p_L^1}{g_L^0} + \left( \frac{p_L^1}{g_L^0} + g_L^0 \tau_L^1 \right) \frac{\left( \frac{g_I^0}{g_L^0} \right)^{\frac{(3\gamma-1)}{2(\gamma+1)}} - 1}{(3\gamma-1)} \right\}.$$

In zone  $II$ , we use the jump relations (6.32) and compute separately  $u^1(\sigma_3^0 \pm)$ . First, using Lemma 6.2 in zone  $III$  for  $\mathbf{v}^1(\sigma_3^0 +)$  gives

$$\begin{cases} \tau^1(\sigma_3^0+) = \sigma_3^0(\alpha_{R,1} + \alpha_{R,2} + \alpha_{R,3}) + g_R^0(\alpha_{R,1} - \alpha_{R,3}) = \sigma_3^0\tau_R^1 + u_R^1, \\ u^1(\sigma_3^0+) = g_R^0\{\sigma_3^0(\alpha_{R,1} - \alpha_{R,3}) + g_R^0(\alpha_{R,1} + \alpha_{R,3})\} = \sigma_3^0u_R^1 - p_R^1, \\ p^1(\sigma_3^0+) = -\sigma_3^0(g_R^0)^2(\alpha_{R,1} + \alpha_{R,3}) = \sigma_3^0p_R^1. \end{cases} \quad (6.36)$$

Similarly, in zone  $II$ , for  $\mathbf{v}^1(\sigma_3^0-)$  and  $\mathbf{v}^1(0+)$ , we have four equations in  $u^1(\sigma_3^0-), p^1(\sigma_3^0-), u_*^1, p_*^1$ . Eliminating the constants  $a_{II,1}$  and  $a_{II,3}$ , we get

$$\begin{cases} u^1(\sigma_3^0-) = u_*^1 - \sigma_3^0 \frac{p_*^1}{(g_{II}^0)^2}, \\ p^1(\sigma_3^0-) = -\sigma_3^0 u_*^1 + p_*^1. \end{cases} \quad (6.37)$$

Substituting (6.36) and (6.37) in the jump relation (6.32b) yields the second equation for  $u_*^1$  and  $p_*^1$ ,

$$\begin{aligned} & \left\{ 1 - \sigma_3^0 \left( \frac{\partial \Phi_R}{\partial p} \right)_{II}^0 \right\} u_*^1 + \left\{ - \frac{\sigma_3^0}{(g_{II}^0)^2} + \left( \frac{\partial \Phi_R}{\partial p} \right)_{II}^0 \right\} p_*^1 \\ &= - \left( \frac{\partial \Phi_R}{\partial \tau_a} \right)_{II}^0 (\sigma_3^0 \tau_R^1 + u_R^1) - \left( \frac{\partial \Phi_R}{\partial p_a} \right)_{II}^0 \{ \sigma_3^0 p_R^1 - (g_R^0)^2 u_R^1 \}, \end{aligned} \quad (6.38)$$

where the derivatives of  $\Phi$ , computed at  $(\tau_R^0, p_R^0, p_*^0)$ , are explicitly given for the equation of state (4.1) by (6.33). The system (6.35) and (6.38) is always numerically solvable and gives  $u_*^1$  and  $p_*^1$ .

- (vi) Computation of  $a_{I,k}$  and  $a_{II,k}$ . The coefficients  $a_{I,k}$  and  $a_{II,k}, k = 1, 3$ , are given in terms of  $u_*^1$  and  $p_*^1$ . We have

$$a_{K,1} = \frac{(u_*^1 - \frac{p_*^1}{g_k^0})}{(g_k^0)^2}, \quad a_{K,3} = \frac{(u_*^1 - \frac{p_*^1}{g_k^0})}{(g_k^0)^2}, \quad \text{for } K = I, II. \quad (6.39)$$

As a corollary, using (6.25b), (6.26), and (6.39), we get explicitly  $\tau^1(0\pm)$

$$\begin{aligned} \tau_I^1 &= \tau^1(0-) = (a_{K,1} - a_{K,3})g_I^0 = \frac{u_*^1}{g_I^0}, \\ \tau_{II}^1 &= \tau^1(0+) = (a_{K,1} - a_{K,3})g_{II}^0 = \frac{u_*^1}{g_{II}^0}, \end{aligned}$$

so that  $\mathbf{v}^1(0\pm)$  is now known at the contact discontinuity.

There remains to compute the coefficient  $a_{II,2}$ . For instance, for a 3-shock, from (6.36), we get

$$\begin{aligned} \tau^1(\sigma_3^0+) &= \sigma_3^0\tau_R^1 + u_R^1, \\ p^1(\sigma_3^0+) &= \sigma_3^0p_R^1 - (g_R^0)^2 u_R^1, \end{aligned}$$

and from (6.37)

$$p^1(\sigma_3^0-) = -\sigma_3^0 u_*^1 + p_*^1.$$

Together with (6.32), this yields

$$\begin{cases} a_{II,2} = -u_*^1 \frac{\sigma_3^0 + g_{II}^0}{\sigma_3^0 (g_{II}^0)^2} + \left( \frac{\partial h}{\partial p} \right)_{II}^0 \left\{ -u_*^1 + \frac{p_*^1}{\sigma_3^0} \right\} \\ \quad + \left( \frac{\partial h}{\partial \tau_a} \right)_R^0 \left\{ \tau_R^1 + \frac{u_R^1}{\sigma_3^0} \right\} + \left( \frac{\partial h}{\partial p_a} \right)_R^0 \left\{ p_R^1 - (g_{II}^0)^2 \frac{u_R^1}{\sigma_3^0} \right\}. \end{cases} \quad (6.40)$$

We have thus computed all the coefficients given by (6.28), (6.30), (6.39), and (6.40) in the particular case of a 1-rarefaction and a 3-shock. The other cases are computed in a similar way.

#### 6.4 Use of the G.R.P. in van Leer's Method

According to (6.5), we have to determine  $u^0, p^0, u^1$ , and  $p^1$  at the contact discontinuity. As in Sect. 2.3, the values of  $u$  and  $p$  at the contact discontinuity between  $\mathbf{U}_j^n, \mathbf{U}_{j+1}^n$  are denoted by  $u_{j+\frac{1}{2}}^n, p_{j+\frac{1}{2}}^n$ ; we shall also denote by  $(\frac{du}{dt})_{j+\frac{1}{2}}^n$  and  $(\frac{dp}{dt})_{j+\frac{1}{2}}^n$  the corresponding values of  $u^1$  and  $p^1$ . When other values  $\varphi$  for which  $\varphi^0$  and  $\varphi^1$  are not continuous are needed, we denote the corresponding values by  $\varphi_{j+\frac{1}{2}\pm}^n$  and  $\frac{d}{dt}\varphi_{j+\frac{1}{2}\pm}^n$ . Then since by (6.6)

$$\mathbf{g}_{j+\frac{1}{2}}^n = \mathbf{f} \left( \mathbf{w}^0(0) + \frac{\Delta t}{2} \mathbf{w}^1(0) \right),$$

we first set

$$\begin{cases} u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = u_{j+\frac{1}{2}}^n + \frac{\Delta t}{2} \left( \frac{du}{dt} \right)_{j+\frac{1}{2}}^n, \\ p_{j+\frac{1}{2}}^{n+\frac{1}{2}} = p_{j+\frac{1}{2}}^n + \frac{\Delta t}{2} \left( \frac{dp}{dt} \right)_{j+\frac{1}{2}}^n, \\ (pu)_{j+\frac{1}{2}}^{n+\frac{1}{2}} = p_{j+\frac{1}{2}}^{n+\frac{1}{2}} u_{j+\frac{1}{2}}^{n+\frac{1}{2}}, \end{cases} \quad (6.41)$$

where in the last Eq. (6.41) we have approximated  $(pu)^0 + t(pu)^1$  by

$$(pu)^0 + t(pu)^1 \approx p^0 u^0 + t(p^0 u^1 + p^1 u) = (p^0 + tp^1)(u^0 + tu^1).$$

Then

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \frac{\Delta t}{\Delta m_j} (u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}}), \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta m_j} (p_{j+\frac{1}{2}}^{n+\frac{1}{2}} - p_{j-\frac{1}{2}}^{n+\frac{1}{2}}), \\ e_j^{n+1} = e_j^n + \frac{\Delta t}{\Delta m_j} ((pu)_{j+\frac{1}{2}}^{n+\frac{1}{2}} - (pu)_{j-\frac{1}{2}}^{n+\frac{1}{2}}), \end{cases} \quad (6.42)$$

where as in (2.28)

$$\Delta m_j = \rho_j^0 \Delta \xi_j \quad (6.43)$$

and

$$p_j^n = p(\tau_j^n, \varepsilon_j^n), \quad \varepsilon_j^n = e_j^n - \frac{(u_j^n)^2}{2}.$$

Let us derive another equivalent form for van Leer's scheme. We have not yet specified the motion of the grid. We solve numerically the equation

$$\frac{dx}{dt} = u.$$

Setting

$$x_{j+\frac{1}{2}}^0 = \xi_{j+\frac{1}{2}}, \quad (6.44)$$

then  $x_{j+\frac{1}{2}}^n$ , which is the Eulerian coordinate of the interface  $\xi_{j+\frac{1}{2}}$  at time  $t_n$ , is updated according to

$$x_{j+\frac{1}{2}}^{n+1} = x_{j+\frac{1}{2}}^n + \Delta t \ u_{j+\frac{1}{2}}^{n+\frac{1}{2}}. \quad (6.45)$$

We can check by induction that

$$\rho_j^n (x_{j+\frac{1}{2}}^n - x_{j-\frac{1}{2}}^n) = \Delta m_j. \quad (6.46)$$

First, this is true for  $n = 0$  by assumption. Suppose that it holds for some  $n$ ; we have

$$\begin{aligned} \Delta m_j \tau_j^{n+1} &= \Delta m_j \tau_j^n + \Delta t (u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}}) \\ &= (x_{j+\frac{1}{2}}^n - x_{j-\frac{1}{2}}^n) + \Delta t (u_{j+\frac{1}{2}}^{n+\frac{1}{2}} - u_{j-\frac{1}{2}}^{n+\frac{1}{2}}) = x_{j+\frac{1}{2}}^{n+1} - x_{j-\frac{1}{2}}^{n+1}. \end{aligned}$$

Hence

$$\rho_j^{n+1} (x_{j+\frac{1}{2}}^{n+1} - x_{j-\frac{1}{2}}^{n+1}) = \Delta m_j, \quad (6.47)$$

which proves the desired result.

The method in Lagrangian coordinates can thus be written as

$$\left\{ \begin{array}{l} \Delta m_j = \rho_j^0 (x_{j+\frac{1}{2}}^0 - x_{j-\frac{1}{2}}^0), \\ x_{j+\frac{1}{2}}^{n+1} = x_{j+\frac{1}{2}}^n + \Delta t u_{j+\frac{1}{2}}^{n+\frac{1}{2}}, \\ \rho_j^{n+1} = (x_{j+\frac{1}{2}}^{n+1} - x_{j-\frac{1}{2}}^{n+1})^{-1} \Delta m_j, \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta m_j} (p_{j+\frac{1}{2}}^{n+\frac{1}{2}} - p_{j-\frac{1}{2}}^{n+\frac{1}{2}}), \\ e_j^{n+1} = e_j^n - \frac{\Delta t}{\Delta m_j} ((pu)_{j+\frac{1}{2}}^{n+\frac{1}{2}} - (pu)_{j-\frac{1}{2}}^{n+\frac{1}{2}}). \end{array} \right. \quad (6.48)$$

*Remark 6.3.* As in Remark 2.1, van Leer's scheme can be interpreted as a finite volume method. Indeed, each of Eqs. (2.25) can be written as

$$\frac{\partial}{\partial t}(\varphi J) + J \frac{\partial f}{\partial x} = 0.$$

By integrating this equation on  $(\xi_{j-\frac{1}{2}}, \xi_{j+\frac{1}{2}})$ , we get

$$\frac{d}{dt} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \varphi J d\xi + \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \frac{\partial f}{\partial x} J d\xi = 0$$

or

$$\frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \varphi dx + (f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}) = 0$$

(the dependence of  $f, \varphi$ , and  $x$  on  $t$  is omitted). We use a midpoint rule for the time integration

$$\Delta x_j^{n+1} \varphi_j^{n+1} = \Delta x_j^n \varphi_j^n - \Delta t (f_{j+\frac{1}{2}}^{n+\frac{1}{2}} - f_{j-\frac{1}{2}}^{n+\frac{1}{2}}),$$

where  $\varphi_j$  is the average value over the cell.

This gives for  $\varphi = \rho, \rho u, \rho e$

$$\begin{cases} \Delta x_j^{n+1} \rho_j^{n+1} = \Delta x_j^n \rho_j^n = \dots = \Delta m_j, \\ \Delta x_j^{n+1} (\rho u)_j^{n+1} = \Delta x_j^n (\rho u)_j^n - \Delta t (p_{j+\frac{1}{2}}^{n+\frac{1}{2}} - p_{j-\frac{1}{2}}^{n+\frac{1}{2}}), \\ \Delta x_j^{n+1} (\rho e)_j^{n+1} = \Delta x_j^n (\rho e)_j^n - \Delta t \left\{ (pu)_{j+\frac{1}{2}}^{n+\frac{1}{2}} - (pu)_{j-\frac{1}{2}}^{n+\frac{1}{2}} \right\}, \end{cases}$$

and thus

$$\begin{cases} \Delta x_j^n \rho_j^n = \Delta m_j, \\ \Delta m_j^{n+1} u_j^{n+1} = \Delta m_j^n u_j^n - \Delta t (p_{j+\frac{1}{2}}^{n+\frac{1}{2}} - p_{j-\frac{1}{2}}^{n+\frac{1}{2}}), \\ \Delta m_j^{n+1} e_j^{n+1} = \Delta m_j^n e_j^n - \Delta t \left\{ (pu)_{j+\frac{1}{2}}^{n+\frac{1}{2}} - (pu)_{j-\frac{1}{2}}^{n+\frac{1}{2}} \right\}. \end{cases}$$

This is the above scheme, provided  $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}, p_{j+\frac{1}{2}}^{n+\frac{1}{2}}$  are defined by the expansion of the solution of the generalized Riemann problem.  $\square$

For the reconstruction step, we need to define the slopes  $\mathbf{S} = (\delta\rho, \delta u, \delta e)$ . By (6.8)

$$\begin{aligned} \hat{\mathbf{S}}_j^{n+1} &= \mathbf{w}(x_{j+\frac{1}{2}} - 0, \Delta t) - \mathbf{w}(x_{j-\frac{1}{2}} + 0, \Delta t) \\ &= \mathbf{v}_{j+\frac{1}{2}-}^{n+1} - \mathbf{v}_{j-\frac{1}{2}+}^{n+1}. \end{aligned}$$

Thus

$$\begin{aligned}(\delta\tau)_j^{n+1} &= \tau_{j+\frac{1}{2}-}^{n+1} - \tau_{j-\frac{1}{2}+}^{n+1}, \\ (\delta u)_j^{n+1} &= u_{j+\frac{1}{2}}^{n+1} - u_{j-\frac{1}{2}}^{n+1}, \\ (\delta e)_j^{n+1} &= e_{j+\frac{1}{2}-}^{n+1} - e_{j-\frac{1}{2}+}^{n+1},\end{aligned}$$

where

$$\begin{aligned}\tau_{j+\frac{1}{2}\pm}^{n+1} &= \tau_{j+\frac{1}{2}\pm}^n + \Delta t \left( \frac{d\tau}{dt} \right)_{j+\frac{1}{2}\pm}^n, \\ u_{j+\frac{1}{2}\pm}^{n+1} &= u_{j+\frac{1}{2}\pm}^n + \Delta t \left( \frac{du}{dt} \right)_{j+\frac{1}{2}}^n, \\ e_{j+\frac{1}{2}\pm}^{n+1} &= e_{j+\frac{1}{2}\pm}^n + \Delta t \left( \frac{de}{dt} \right)_{j+\frac{1}{2}\pm}^n,\end{aligned}$$

and this is followed by a correction procedure. The slope is taken to be equal to zero if the slab average reaches an extremum or varies in a way opposite to the variation of the average values and is limited so that the values at the end of the mesh (or less restrictively the average value) lie between the neighboring slab averages (see G.R., Chapter 4, Section 3, for details).

*Remark 6.4.* In fact, it is more convenient to work with  $\rho$ , and, moreover,  $\tau$  does not appear in the formulas. Thus, we can suppose that  $\rho$  is an affine function and then correct  $\delta\rho$  instead of  $\delta\tau$ . We can also think of limiting the pressure. For a polytropic ideal gas law

$$p = p(\tau, \varepsilon) = (\gamma - 1) \frac{\varepsilon}{\tau}, \quad \varepsilon = e - \frac{u^2}{2},$$

we can write

$$\frac{\delta p}{p} = \frac{\delta \varepsilon}{\varepsilon} - \frac{\delta \tau}{\tau}.$$

Limiting the physical (primitive) rather than the conservative variables may reduce the oscillations.  $\square$

As we have done for Godunov's method, we can now derive a numerical method in Eulerian coordinates that couples a Lagrangian step and then a remapping onto the fixed Eulerian grid, with nodes  $x_{j+\frac{1}{2}}$ . Let us note that in (6.1),  $\mathbf{v}_j^n$  may indeed be viewed as the cell average of  $\tilde{\mathbf{v}}^n(x)$ ,

$$\mathbf{v}_j^n = \frac{1}{\Delta x_j} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{\mathbf{v}}^n(x) dx = \frac{1}{2} (\tilde{\mathbf{v}}^n(x_{j+\frac{1}{2}}) + \tilde{\mathbf{v}}^n(x_{j-\frac{1}{2}})),$$

while  $\mathbf{S}_j^n$  is related to the first “moment”

$$\mathbf{S}_j^n = \frac{12}{(\Delta x_j)^2} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{\mathbf{v}}^n(x)(x - x_j) dx.$$

(i) The Lagrangian step.

At the beginning of each Lagrangian step, the Lagrangian and Eulerian grids coincide, and  $\xi_{j+\frac{1}{2}} = x_{j+\frac{1}{2}}$  (see Fig. 2.3). Following (2.34), we define the updated Eulerian coordinates of the Lagrangian zone by

$$x_{j+\frac{1}{2}}^* = x_{j+\frac{1}{2}} = \Delta t u_{j+\frac{1}{2}}^n,$$

where the dependence of  $x_{j+\frac{1}{2}}^*$  on  $n$  is omitted. In other words, if  $x_{j+\frac{1}{2}} = x(\xi_{j+\frac{1}{2}}, t_n)$ , then  $x_{j+\frac{1}{2}}^* \approx x(\xi_{j+\frac{1}{2}}, t_{n+1})$ . We obtain by formulas (6.48) the quantities  $\rho_*^{n+1}, u_*^{n+1}, e_*^{n+1}$ , which are piecewise affine functions in each interval. In fact, we shall assume that  $\rho_*^{n+1}$  is an affine function of  $x$  in each interval  $(x_{j-\frac{1}{2}}^*, x_{j+\frac{1}{2}}^*)$  (this avoids a costly operation during the remapping procedure; see step (iii) and Remark 6.5), while  $u_*^{n+1}, e_*^{n+1}$  are affine functions of the mass variable  $m$ . The quantities  $\rho_*^{n+1}, u_*^{n+1}, e_*^{n+1}$  are not necessarily linear in the interval  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$  of the Eulerian grid, which, at the next Lagrangian step, coincides with the Lagrangian grid (see, for instance, Fig. 6.3). Therefore, we must “remap,” i.e., replace,  $\rho_*^{n+1}, u_*^{n+1}, e_*^{n+1}$  by linear functions  $\tilde{\rho}^{n+1}, \tilde{u}^{n+1}, \tilde{e}^{n+1}$ , which share the same cell average and first moment (w.r.t.  $x$  for  $\rho$ , w.r.t.  $m$  for  $u$  and  $e$ ). This is the object of the projection step. We shall first detail the updating of the mass coordinates.

(ii) Definition of the updated mass coordinates  $m_{j+\frac{1}{2}}^{n+1}$ .

Using the density versus the space coordinates for the definition of  $m_{j+\frac{1}{2}}^{n+1}$ ,

$$\rho_*^{n+1}(x) = \rho_j^{n+1} + (x - x_j^*) \frac{(\delta\rho)_j^{n+1}}{\Delta x_j^*}, \quad x_{j-\frac{1}{2}}^* < x < x_{j+\frac{1}{2}}^*, \quad (6.49)$$

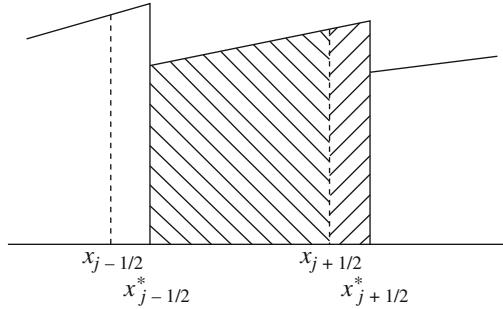
the mass  $m_{j+\frac{1}{2}}^{n+1} - m_{j+\frac{1}{2}}^n$  convected across  $x_{j+\frac{1}{2}}$  is

$$m_{j+\frac{1}{2}}^n - m_{j+\frac{1}{2}}^{n+1} = \int_{x_{j+\frac{1}{2}}}^{x_{j+\frac{1}{2}}^*} \rho_*^{n+1}(x) dx.$$

Some easy computation gives

$$m_{j+\frac{1}{2}}^n - m_{j+\frac{1}{2}}^{n+1} = (x_{j+\frac{1}{2}}^* - x_{j+\frac{1}{2}}) \left\{ \rho_j^{n+1} + \frac{(\delta\rho)_j^{n+1}}{2} \left( 1 - \frac{x_{j+\frac{1}{2}}^* - x_{j+\frac{1}{2}}}{\Delta x_j^*} \right) \right\},$$

which yields a simple equation for  $m_{j+\frac{1}{2}}^{n+1}$ .



**Fig. 6.3** Necessity of “remapping”

*Remark 6.5.* The mass coordinates  $m_{j+\frac{1}{2}}^{n+1}$  at time  $t_{n+1}$  of the Eulerian zone boundaries could be computed using  $\tau_*^{n+1}(m)$  in the (original) Lagrangian zone. Consider, for instance, the usual case of Fig. 6.3, which corresponds also to Fig. 2.3 ( $u_{j-\frac{1}{2}}^n > 0$ ,  $u_{j+\frac{1}{2}}^n > 0$ ; the Lagrangian zone has moved to the right). We can define  $m_{j+\frac{1}{2}}^{n+1}$  from the following equality obtained with the specific volume

$$\tau_*^{n+1}(m) = \tau_j^{n+1} + (m - m_j^n) \frac{(\delta\tau)_j^{n+1}}{\Delta m_j^n}, \quad m_{j-\frac{1}{2}}^n < m < m_{j+\frac{1}{2}}^n, \quad (6.50)$$

defined on the (original) Lagrangian zone:

$$\begin{aligned} \int_{j+\frac{1}{2}, n}^{j+\frac{1}{2}, n+1} \tau_*^{n+1}(m) dm &= \int \left\{ \tau_j^{n+1} + (m - m_j^n) \frac{(\delta\tau)_j^{n+1}}{\Delta m_j^n} \right\} dm \\ &= x_{j+\frac{1}{2}}^* - x_{j+\frac{1}{2}}, \end{aligned}$$

where the integral lies over  $(m_{j+\frac{1}{2}}^n, m_{j+\frac{1}{2}}^{n+1})$  and

$$m_j^n = \frac{(m_{j+\frac{1}{2}}^n + m_{j-\frac{1}{2}}^n)}{2} = m_{j+\frac{1}{2}}^n - \frac{\Delta m_j^n}{2}.$$

Using the obvious identities

$$(b - m)^2 - (a - m)^2 = (b - a)(a + b - 2m)$$

and

$$m_{j+\frac{1}{2}}^n + m_{j+\frac{1}{2}}^{n+1} - 2 \left( m_{j+\frac{1}{2}}^n - \frac{\Delta m_j^n}{2} \right) = \Delta m_j^n + m_{j+\frac{1}{2}}^{n+1} - m_{j+\frac{1}{2}}^n,$$

we compute easily

$$\begin{aligned} & \int \left\{ \tau_j^{n+1} + (m - m_j^n) \frac{(\delta\tau)_j^{n+1}}{\Delta m_j^n} \right\} dm \\ &= (m_{j+\frac{1}{2}}^{n+1} - m_{j+\frac{1}{2}}^n) \left\{ \tau_j^{n+1} + \frac{(\delta\tau)_j^{n+1}}{2} \left( 1 - \frac{m_{j+\frac{1}{2}}^{n+1} - m_{j+\frac{1}{2}}^n}{\Delta m_j^n} \right) \right\}, \end{aligned}$$

which gives a quadratic equation for the mass  $m_{j+\frac{1}{2}}^{n+1} - m_{j+\frac{1}{2}}^n$  convected across  $x_{j+\frac{1}{2}}$ .

Let us check that it is equivalent to use, in the original Lagrangian zone, the density  $\rho(x)$  given by (6.48),

$$\rho_*^{n+1}(x) = \rho_j^{n+1} + (x - x_j^*) \frac{(\delta\rho)_j^{n+1}}{\Delta x_j^*}, \quad x_{j-\frac{1}{2}}^* < x < x_{j+\frac{1}{2}}^*,$$

instead of the specific volume  $\tau(m)$  given by (6.50),

$$\tau_*^{n+1}(m) = \tau_j^{n+1} + (m - m_j) \frac{(\delta\tau)_j^{n+1}}{\Delta m_j}, \quad m_{j-\frac{1}{2}} < m < m_{j+\frac{1}{2}}$$

(we have dropped in  $m_{j+\frac{1}{2}} = m_{j+\frac{1}{2}}^n$  the superscript corresponding to the time  $t_n$ ). Indeed, consider first the averaged values; since in the pure Lagrangian method,  $x_{j+\frac{1}{2}}^{n+1}$  corresponds to what we have denoted by  $x_{j+\frac{1}{2}}^*$ , by (6.48), we have

$$\rho_j^{n+1} = (\Delta x_j^*)^{-1} \int_{x_{j-\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}^*} \rho_*^{n+1}(x) dx = (\Delta x_j^*)^{-1} \Delta m_j = \frac{1}{\tau_j^{n+1}}, \quad (6.51a)$$

where

$$\tau_j^{n+1} = (\Delta m_j)^{-1} \int_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}} \tau_*^{n+1}(m) dm$$

is the average value of  $\tau_*^{n+1}(m)$ . Then

$$\begin{aligned} (\delta\rho)_j^{n+1} &= \frac{12}{(\Delta x_j^*)^2} \int_{x_{j-\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}^*} \rho_*^{n+1}(x)(x - x_j^*) dx \\ &= \frac{12}{(\Delta x_j^*)^2} \int_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}} (x - x_j^*) d(m - m_j). \end{aligned}$$

Integrating by parts yields

$$\begin{aligned}
(\delta\rho)_j^{n+1} &= -\frac{12}{(\Delta x_j^*)^2} \int_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}} (m - m_j) \left( \frac{dx}{dm} \right) dm \\
&= -\frac{12}{(\Delta x_j^*)^2} \int_{x_{j-\frac{1}{2}}^*}^{x_{j+\frac{1}{2}}^*} (m - m_j) dx \\
&= -\frac{12}{(\Delta x_j^*)^2} \int_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}} (m - m_j) \tau_*^{n+1}(m) dm \\
&= -\frac{12}{(\Delta x_j^*)^2} \frac{(\Delta m_j)^2}{12} \delta\tau_j^{n+1},
\end{aligned}$$

so that

$$(\delta\rho)_j^{n+1} = -\frac{\delta\tau_j^{n+1}}{(\tau_j^{n+1})^2}, \quad (6.51b)$$

since  $\delta\tau_j^{n+1}$  is the first moment of  $\tau_*^{n+1} = \tau_j^{n+1} + (m - m_j)(\delta\tau)_j^{n+1}/\Delta m_j$ . Formulas (6.51) prove indeed that we can convert  $\tau$  into  $\rho$  and vice versa.

□

(iii) The projection step.

The values  $\rho_*^{n+1}, u_*^{n+1}, e_*^{n+1}$  are projected back on the Eulerian grid. We define a piecewise linear density on the Eulerian grid  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$  that shares the same average  $\rho_j^{n+1}$  and slope  $(\hat{\delta}\rho)_j^{n+1}$  as  $\rho_*^{n+1}(x)$  (the hat corresponds to the fact that we have not yet limited the slope). We compute

$$\rho_j^{n+1} = (\Delta x_j)^{-1} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \rho_*^{n+1}(x) dx$$

by taking into account the fact that the function is piecewise linear on  $(x_{j-\frac{1}{2}}^*, x_{j+\frac{1}{2}}^*)$  (in the case of Fig. 6.3, for instance, the function is linear on  $(x_{j-\frac{1}{2}}^*, x_{j+\frac{1}{2}}^*)$ , but we might have three pieces if  $u_{j-\frac{1}{2}}^n > 0, u_{j+\frac{1}{2}}^n < 0$ ). We have already noted that the slope is given by the first “moment,” i.e., we have the following obvious equality:

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} (x - x_j)(ax + b) dx = a \frac{(\Delta x_j)^3}{12},$$

from which we deduce that  $(\hat{\delta}\rho)_j^{n+1}$  is given by

$$(\hat{\delta}\rho)_j^{n+1} = \frac{12}{\Delta x_j^2} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \rho_*^{n+1}(x)(x - x_j) dx.$$

Now, having defined an updated mass coordinate  $m_{j+\frac{1}{2}}^{n+1}$ , we can project  $u$  and  $e$  with respect to the mass variable. We get

$$u_j^{n+1} = (\Delta m_j^{n+1})^{-1} \int_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}} u_*^{n+1}(m) dm,$$

where  $m_{j \pm \frac{1}{2}}$  stands for  $m_{j \pm \frac{1}{2}}^{n+1}$ , and

$$(\hat{\delta}u)_j^{n+1} = \frac{12}{(\Delta m_j^{n+1})^2} \int_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}} u_*^{n+1}(m)(m - m_j^{n+1}) dm,$$

where

$$m_j^{n+1} = \frac{1}{2}(m_{j+\frac{1}{2}}^{n+1} + m_{j-\frac{1}{2}}^{n+1}),$$

with similar formulas for  $e$ .

For practical use, the costly computation of the first moment for defining the slope can be replaced by interpolation and monotonicity correction, which we now detail.

#### (iv) Monotonicity correction.

When correcting the slopes, we can take advantage of the fact that the projection is done onto the Eulerian grid and use a milder kind of limiting than (6.10) (see van Leer [1158]). Hence, we take for limiting the slope  $\hat{\delta}v_k$  of a quantity  $v_k$  ( $v_k = \rho$ ,  $u$ , or  $e$ )

$$\hat{\delta}v_{kj} = \begin{cases} s \min \left\{ \frac{2|\Delta v_{k,j-\frac{1}{2}}|}{\sigma_{j+\frac{1}{2}}}, |\hat{\delta}v_{k,j}|, \frac{2|\Delta v_{k,j+\frac{1}{2}}|}{(1-\sigma_{j+\frac{1}{2}})} \right\} \\ \quad \text{if } s = \operatorname{sgn} \Delta v_{k,j-\frac{1}{2}} = \operatorname{sgn} \Delta v_{k,j+\frac{1}{2}} = \operatorname{sgn} \hat{\delta}v_{k,j}, \\ 0 \quad \text{otherwise,} \end{cases} \quad (6.52)$$

where

$$\sigma_{j+\frac{1}{2}} = \frac{m_{j+\frac{1}{2}}^n - m_{j+\frac{1}{2}}^{n+1}}{\Delta m_j^n}$$

is the mass fraction of the Lagrangian slab ( $m_{j-\frac{1}{2}}^n, m_{j+\frac{1}{2}}^n$ ) that has crossed the Eulerian zone boundary  $x_{j+\frac{1}{2}}$ . The limitation is to be used on the Lagrangian quantities once we have determined  $m_{j+\frac{1}{2}}^{n+1}$ , i.e., between steps (ii) and (iii).

The functions  $\tilde{\rho}^{n+1}, \tilde{u}^{n+1}, \tilde{e}^{n+1}$  are thus reconstructed, and  $\tilde{\rho}^{n+1}$  can be converted into  $\tilde{\tau}^{n+1}$  for the next Lagrangian step. The new Lagrangian zones can now be defined, coinciding again with the Eulerian grid. Other references for schemes using G.R.P. are Ben Artzi and Falcovitz [101, 103] and Ben Artzi and Birman [102] and in the case of reactive flows Falcovitz and Ben Artzi (1992–1993) and the monograph [104] and the references therein; ADER (Arbitrary high-order DERivative Riemann problem) methods involve modified G.R.P. [875, 1123] and also [545]. More recent work and related references can be found in [961].

## 7 Kinetic Schemes for the Euler Equations

The kinetic schemes we introduce in this section will eventually be written in the general form of finite volume schemes. It seems, however, essential for a better understanding of their properties to have first a quick look at the underlying kinetic theory, skipping most difficulties since it is out of the scope of this book to present a complete treatment of this theory. We refer to Cercignani [253], Cercignani et al. [254], Lions [809], and Perthame's textbook [944] and also Bouchut et al. [169] for an extended and more rigorous approach.

### 7.1 The Boltzmann Equation

First, we consider for simplicity a rarefied monatomic perfect gas and assume in this section that the dimension is  $d = 3$  (however, we keep the notation  $d$  since in dimension  $d = 1$  or  $d = 2$ , most formulas remain valid). The Boltzmann equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \operatorname{grad}_{\mathbf{x}} f = Q(f, f) \quad (7.1)$$

describes the time evolution of the one-particle distribution  $f(t, \mathbf{x}, \mathbf{v})$  for a gas in the phase space  $(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^{2d}$ , where  $\mathbf{x}$  is the position vector and  $\mathbf{v}$  the molecular velocity. One can say that  $f$  is the expected mass density in the phase space, and thus

$$\rho = \rho(\mathbf{x}, t) = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) d\mathbf{v}$$

will represent the mass per unit volume, i.e., the density in the physical space.

Equation (7.1) can be justified by considering the limit (Boltzmann–Grad limit) as  $N \rightarrow \infty$  of a system of  $N$  interacting particles: the gas is made of many identical molecules considered as hard spheres with diameter  $a$  ( $Na^2$  is fixed,  $a \rightarrow 0$ ), and we assume elastic, binary collisions and no external force fields. We have  $f = NmP = MP$  where  $P$  represents the probability density that a random molecule has velocity  $\mathbf{v} \pm d\mathbf{v}$  and position  $\mathbf{x} \pm d\mathbf{x}$ , and  $N$  is the number of molecules,  $m$  the mass of molecule, and  $M$  the total mass. Since  $P$  is a probability density,  $\int_{\mathbb{R}^{2d}} P d\mathbf{x} d\mathbf{v} = 1$  and

$$\int_{\mathbb{R}^{2d}} f d\mathbf{x} d\mathbf{v} = M = Nm.$$

Note that the expression of the collision term  $Q$  is often written as  $\frac{1}{\nu}Q$ ,  $\nu > 0$  with a parameter  $\nu$  representing the *mean free path*: as  $\nu \rightarrow \infty$ , the

equation becomes a pure transport equation, and as  $\nu \rightarrow 0$ , it forces  $f$  toward a state of equilibrium, satisfying  $Q(f, f) = 0$ .

We shall not give the precise expression for the collision operator  $Q(f, f)$ ; it is a quadratic integral operator, which acts on the velocity dependence of  $f$  and is symmetric ( $Q(f, g) = Q(g, f)$ ). We mention the main property in the following theorem.

*Theorem 7.1*

*The states of thermodynamic equilibrium characterized by  $Q(f, f) = 0$  are obtained for the Maxwellian distributions*

$$f(\mathbf{v}) = A \exp(-\beta|\mathbf{v} - \mathbf{u}|^2), \quad (7.2)$$

where  $A \in \mathbb{R}^+$ ,  $\mathbf{u} \in \mathbb{R}^d$ ,  $\beta \in \mathbb{R}^+$  are arbitrary parameters.

*Proof.* The proof of this result relies on the following important facts:

- (i) The microscopic collisional invariants, i.e., the functions  $\varphi = \varphi(\mathbf{v})$  satisfying

$$\int_{\mathbb{R}^d} Q(f, g)\varphi(\mathbf{v})d\mathbf{v} = 0, \quad \forall f, g \geq 0,$$

are exactly the functions  $\varphi(\mathbf{v}) = a + \mathbf{b} \cdot \mathbf{v} + c|\mathbf{v}|^2$  (where  $a, c$  are constants and  $\mathbf{b}$  is a constant vector). The elementary invariants are the components of the vector

$$\mathbf{K}(\mathbf{v}) = (1, \mathbf{v}, |\mathbf{v}|^2)^T \in \mathbb{R}^{d+2}. \quad (7.3)$$

Thus, for  $g = f \geq 0$ , we obtain

$$\int_{\mathbb{R}^d} Q(f, f)\mathbf{K}(\mathbf{v})d\mathbf{v} = \mathbf{0}, \quad (7.4a)$$

and for  $f$  satisfying (7.1),

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d} \mathbf{K}(\mathbf{v})f(t, \mathbf{x}, \mathbf{v})d\mathbf{v} + \int_{\mathbb{R}^d} \mathbf{v} \cdot \text{grad}_{\mathbf{x}} f(t, \mathbf{x}, \mathbf{v})\mathbf{K}(\mathbf{v})d\mathbf{v} = \mathbf{0}. \quad (7.4b)$$

- (ii) Whatever the distribution function  $f \geq 0$ , the following Boltzmann inequality holds:

$$\int_{\mathbb{R}} \log f Q(f, f)d\mathbf{v} \leq 0, \quad (7.5)$$

where equality holds if and only if  $\log f$  is an invariant (the “if” part is obvious), which implies

$$f(\mathbf{v}) = \exp(a + \mathbf{b} \cdot \mathbf{v} + c|\mathbf{v}|^2)$$

or, with an appropriate choice of constants,  $f$  is a Maxwellian,

$$f(\mathbf{v}) = \mathbf{A} \exp(-\beta(\mathbf{v} - \mathbf{u})^2),$$

which gives the result.  $\square$

As a consequence of (7.5), we find that introducing the function

$$h(r) = r \log r, \quad (7.6)$$

we have

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d} h(f) d\mathbf{v} + \int_{\mathbb{R}^d} \mathbf{v} \cdot \operatorname{grad}_{\mathbf{x}} h(f) d\mathbf{v} \leq 0, \quad (7.7)$$

where equality holds iff  $f$  is a Maxwellian (part of the H-theorem). Indeed, we write

$$\begin{aligned} \frac{\partial}{\partial t} \int_{\mathbb{R}^d} h(f) d\mathbf{v} + \int_{\mathbb{R}^d} \mathbf{v} \cdot \operatorname{grad}_{\mathbf{x}} h(f) d\mathbf{v} &= \int_{\mathbb{R}^d} h'(f) \left( \frac{\partial f}{\partial t} + \mathbf{v} \cdot \operatorname{grad}_{\mathbf{x}} f \right) d\mathbf{v} \\ &= \int_{\mathbb{R}^d} h'(f) Q(f, f) d\mathbf{v} = \int_{\mathbb{R}^d} (\log f + 1) Q(f, f) d\mathbf{v} = \int_{\mathbb{R}^d} \log f Q(f, f) d\mathbf{v} \end{aligned}$$

since 1 is a collision invariant. Now, by (7.5), the last integral is  $\leq 0$  and vanishes iff  $f$  is a Maxwellian.

The function  $h(r) = r \log r$ , which is the microscopic or kinetic entropy, is strictly convex. If we define

$$H(f) = \int_{\mathbb{R}^d} h(f) d\mathbf{v} = \int_{\mathbb{R}^d} f \log f d\mathbf{v} \quad (7.8a)$$

( $H$  is closely related to the thermodynamic entropy as we shall see below) and

$$\boldsymbol{\Psi}(f) = (\boldsymbol{\Psi}_i(f)), \quad (7.8b)$$

where

$$\boldsymbol{\Psi}_i(f) = \int_{\mathbb{R}^d} v_i h(f) d\mathbf{v} = \int_{\mathbb{R}^d} v_i f \log f d\mathbf{v},$$

then (7.7) yields

$$\frac{\partial}{\partial t} H(f) + \operatorname{div}_{\mathbf{x}} \boldsymbol{\Psi}(f) \leq 0. \quad (7.8c)$$

*Remark 7.1.* If we integrate over  $\mathbb{R}^d$  or for more realistic situations over a region  $\mathcal{R}$  filled by gas, and if we assume moreover that  $\mathcal{R}$  is bounded by nonporous solid walls and at no point of the boundary of  $\mathcal{R}$  is heat flowing from the gas, we can prove that (for any solution  $f$  of (7.1)) the quantity

$$\mathcal{H} = \int_{\mathcal{R}} H(f) d\mathbf{x} = \int_{\mathbb{R}^d \times \mathcal{R}} h(f) d\mathbf{v} d\mathbf{x}$$

decreases, i.e.,

$$\frac{d\mathcal{H}}{dt} \leq 0$$

and  $\mathcal{H}$  is constant if and only if  $f$  is a Maxwellian.

The Boltzmann equation describes the evolution (“relaxation”) toward a state of minimum for  $\mathcal{H}$ . Under the above assumption, the final state ( $t \rightarrow +\infty$ ) is a steady state and thus a Maxwellian. In particular, the distribution function in an equilibrium state (i.e., a steady state with no energy exchange with the surroundings) is a Maxwellian.  $\square$

Given a distribution function  $f$ , let us define the following macroscopic quantities (called the moments of  $f$ ):

$$\mathbf{U}(\mathbf{x}, t) = \begin{pmatrix} \rho \\ \rho \mathbf{u} \\ \rho e \end{pmatrix} (\mathbf{x}, t) = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) \begin{pmatrix} 1 \\ \mathbf{v} \\ |\mathbf{v}|^2/2 \end{pmatrix} d\mathbf{v}. \quad (7.9)$$

Because of the probabilistic meaning of  $f$ ,  $\rho$  represents the mass per unit volume, and  $\mathbf{u}$  represents the velocity. The specific total energy  $e$  satisfies

$$\rho e = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) \frac{|\mathbf{v}|^2}{2} d\mathbf{v} = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) \frac{|\mathbf{v} - \mathbf{u}|^2}{2} d\mathbf{v} + \rho \frac{|\mathbf{u}|^2}{2}.$$

Thus

$$\rho \varepsilon = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) \frac{|\mathbf{v} - \mathbf{u}|^2}{2} d\mathbf{v}$$

represents the internal energy (per unit volume). The velocity

$$\mathbf{C} = \mathbf{v} - \mathbf{u}$$

is called the peculiar velocity. We also define the stress tensor  $\pi = (\pi_{ij})_{1 \leq i, j \leq 3}$

$$\pi_{ij} = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) C_i C_j d\mathbf{v}, \quad 1 \leq i, j \leq 3.$$

We have

$$\sum_{i=1}^3 \pi_{ii} = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{u}) |\mathbf{C}|^2 d\mathbf{v} = 2\rho\varepsilon,$$

and it is convenient to identify  $\frac{1}{3}(\sum_{i=1}^3 \pi_{ii}) = \frac{2}{3}\rho\varepsilon$  with the gas pressure  $p$ ; thus

$$p = \frac{2\rho\varepsilon}{3} = \frac{1}{d} \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) |\mathbf{u} - \mathbf{v}|^2 d\mathbf{v}. \quad (7.10)$$

Indeed, for a monatomic perfect gas,  $\varepsilon$  is a function of temperature only; we get  $\frac{p}{\rho} = \text{constant}$  if temperature = constant. This property is characteristic of a perfect gas, for which Gay-Lussac and Boyle’s law,  $p = \rho RT$ , holds (R is the Boltzmann constant of the gas). With the above identification, we are led to define the kinetic temperature  $T$  by

$$T = \frac{2\varepsilon}{3R},$$

and we obtain for a monatomic perfect gas (in dimension  $d = 3$ )

$$\varepsilon = \frac{3RT}{2}. \quad (7.11)$$

(It is convenient to note the vectors of  $\mathbb{R}^{d+2}$  in transposed form, such as  $K$  in (7.3), for which the notation is easily distinguishable from the temperature.)

*Remark 7.2.* Note that with the notations of the preceding chapters (Chap. II, Example 2.1, or Chap. III, Sect. 1.2),  $p = (\gamma - 1)\rho\varepsilon$ , and here  $p = 2\rho\varepsilon/3$ ; thus the adiabatic exponent is  $\gamma = \frac{5}{3}$ , and the specific heat at constant volume is  $C_v = \frac{R}{(\gamma-1)} = 3R/2$ . Also, after some computations, the relation defining the thermodynamic entropy  $TdS = d\varepsilon + pd\tau$ , which yields  $S - S_0 = C_v \log \frac{\varepsilon}{\rho^{\gamma-1}}$ , gives

$$S = R \log \left( \frac{T^{\frac{3}{2}}}{\rho} \right) + S_0 \quad (7.12)$$

in dimension  $d = 3$  for a monatomic perfect gas. □

Last we introduce the heat flow vector  $\mathbf{Q} = (Q_i)$

$$Q_i = \frac{1}{2} \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) C_i |\mathbf{C}|^2 d\mathbf{v}, \quad 1 \leq i \leq 3.$$

Then, from (7.1) and (7.4a), we deduce the following result.

*Proposition 7.1*

Assume that  $f$  is a solution of (7.1). Then, the vector  $\mathbf{U}$  defined by (7.9) satisfies the system

$$\begin{cases} \frac{\partial \rho}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} (\rho u_j) = 0, \\ \frac{\partial \rho u_i}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} (\rho u_i u_j + \pi_{ij}) = 0, \quad 1 \leq i \leq d, \\ \frac{\partial \rho e}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \left\{ \rho u_j e + \sum_{k=1}^d \pi_{jk} u_k + Q_j \right\} = 0. \end{cases}$$

We shall detail a similar computation a little later (in the proof of Proposition 7.2). Thus, in general, the equations satisfied by the moments contain a stress tensor  $(\pi_{ij})$  and a heat flux vector  $\mathbf{Q}$ .

Assume now that  $f$  is a Maxwellian. Substituting (7.2) in (7.9), we find that the parameter  $\mathbf{u}$  in (7.2) is indeed the velocity defined by (7.9), while the other coefficients satisfy (in the case  $d = 3$ )

$$\beta = \frac{3}{4\varepsilon} = \frac{1}{2RT}, \quad A = \left( \frac{4\pi\varepsilon}{3} \right)^{-\frac{3}{2}} \rho = (2\pi RT)^{-\frac{3}{2}} \rho,$$

and thus

$$f(\mathbf{v}) = (2\pi RT)^{-\frac{3}{2}} \rho \exp\left(-\frac{|\mathbf{v} - \mathbf{u}|^2}{2RT}\right). \quad (7.13)$$

Moreover,  $\pi_{jk} = p\delta_{jk}$  and  $Q_j = 0$  from which we deduce the following corollary.

*Corollary 7.1*

Assume, moreover, that  $f$  is a Maxwellian (7.13). The associated vector  $\mathbf{U}$  satisfies the Euler equations

$$\begin{cases} \frac{\partial \rho}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} (\rho u_j) = 0, \\ \frac{\partial \rho u_i}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} (\rho u_i u_j) + \frac{\partial p}{\partial x_i} = 0, \quad 1 \leq i \leq d, \\ \frac{\partial \rho e}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} ((\rho e + p) u_j) = 0, \end{cases} \quad (7.14)$$

where, for a monatomic perfect gas in dimension  $d = 3$ ,

$$p = \rho RT = \frac{2\rho\varepsilon}{3}.$$

Let us remark that, at this level, the above assumption that  $f$  is a Maxwellian is somehow formal; it may be understood as taking the limit  $\nu \rightarrow 0$  in the Boltzmann equation with collision kernel  $\frac{1}{\nu}Q$ .

Defining the function  $H = H(\mathbf{U})$  (i.e.,  $H(M(\mathbf{v}; \rho, \mathbf{u}, T))$  by (7.8) for the Maxwellian distribution  $f = M$ ), we have

$$\begin{aligned} H &= \int_{\mathbb{R}^d} M \log(M) d\mathbf{v} \\ &= \log(\rho(2\pi RT)^{-\frac{3}{2}}) \int_{\mathbb{R}^d} M d\mathbf{v} - \frac{1}{2RT} \int M |\mathbf{v} - \mathbf{u}|^2 d\mathbf{v} \\ &= \rho \log(\rho(2\pi RT)^{-\frac{3}{2}}) - \frac{\rho\varepsilon}{RT} \\ &= \rho \left\{ \log(\rho T^{-\frac{3}{2}}) - \frac{3}{2} \log(2\pi R) - \frac{3}{2} \right\} \\ &= \rho \log(\rho T^{-\frac{3}{2}}) - \frac{3}{2} \rho \{\log(2\pi R) + 1\}, \end{aligned}$$

and thus

$$H(\mathbf{U}) = \rho \{\log(\rho T^{-\frac{3}{2}}) + C\}. \quad (7.15)$$

*Lemma 7.1*

The function  $H(\mathbf{U})$  is a strictly convex function of  $\mathbf{U}$ .

*Proof.* Since  $H$  is a function of  $(\rho, \varepsilon)$  only, following the results of Chap. III, Sect. 1, it is enough to prove that

$$\begin{aligned}\operatorname{tr} H'' &= \frac{\partial^2 H}{\partial \rho^2} + \frac{\partial^2 H}{\partial \varepsilon^2} \geq 0, \\ \det H'' &= \frac{\partial^2 H}{\partial \rho^2} \frac{\partial^2 H}{\partial \varepsilon^2} - \left( \frac{\partial^2 H}{\partial \rho \partial \varepsilon} \right)^2 \geq 0,\end{aligned}$$

where  $H''$  is the  $2 \times 2$  Hessian matrix.

By the chain rule, we compute

$$\begin{aligned}\operatorname{tr} H'' &= \int_{\mathbb{R}^d} h'(M) \left\{ \frac{\partial^2 M}{\partial \rho^2} + \frac{\partial^2 M}{\partial \varepsilon^2} \right\} d\mathbf{v} \\ &\quad + \int_{\mathbb{R}^d} h''(M) \left( \left( \frac{\partial M}{\partial \rho} \right)^2 + \left( \frac{\partial M}{\partial \varepsilon} \right)^2 \right) d\mathbf{v}.\end{aligned}$$

Since  $h$  is convex, the second integral is nonnegative, and it is easy to check that the first integral vanishes. Indeed, we can write  $h'(M)$  in the form

$$h'(M) = a(\rho, \varepsilon, \mathbf{u}) + b(\rho, \varepsilon, \mathbf{u}) |\mathbf{v}|^2$$

and

$$\begin{aligned}\int_{\mathbb{R}^d} h'(M) \left\{ \frac{\partial^2 M}{\partial \rho^2} + \frac{\partial^2 M}{\partial \varepsilon^2} \right\} d\mathbf{v} &= a \left\{ \frac{\partial^2}{\partial \rho^2} + \frac{\partial^2}{\partial \varepsilon^2} \right\} \int_{\mathbb{R}^d} M d\mathbf{v} \\ &\quad + b \left\{ \frac{\partial^2}{\partial \rho^2} + \frac{\partial^2}{\partial \varepsilon^2} \right\} \int_{\mathbb{R}^d} M |\mathbf{v}|^2 d\mathbf{v} = 0.\end{aligned}$$

This yields

$$\operatorname{tr} H'' \geq 0.$$

With a similar argument, we get

$$\begin{aligned}\det H'' &= \left( \int_{\mathbb{R}^d} h''(M) \left( \frac{\partial M}{\partial \rho} \right)^2 d\mathbf{v} \right) \left( \int_{\mathbb{R}^d} h''(M) \left( \frac{\partial M}{\partial \varepsilon} \right)^2 d\mathbf{v} \right) \\ &\quad - \left( \int_{\mathbb{R}^d} h''(M) \left( \frac{\partial M}{\partial \rho} \frac{\partial M}{\partial \varepsilon} \right) d\mathbf{v} \right)^2 \geq 0\end{aligned}$$

by the Cauchy–Schwarz inequality.  $\square$

*Remark 7.3.* We can give a more fundamental interpretation of the convexity of  $H$  (following Brenier [192] unpublished). We can write

$$H(\mathbf{U}) = \operatorname{Min} \left\{ \int_{\mathbb{R}^d} h(f) dv; f(v) \geq 0, \int_{\mathbb{R}^d} f(u) \mathbf{K}(u) dv = \mathbf{U} \right\},$$

where  $\mathbf{K}, \mathbf{U}$  are given by (7.3) and (7.9), and the minimum is reached for  $f = M$ . Since we minimize a convex functional under linear constraints, the minimum  $H(\mathbf{U})$  is a convex function of the constraints. (See also Croisille and Delorme [375] and Brenier and Osher [193].)  $\square$

The above expression (7.15) for  $H$  shows that it is related to the macroscopic entropy  $S$  given by (7.12),

$$\frac{H}{\rho} - C = \log(\rho T^{-\frac{3}{2}}) = -\frac{(S - S_0)}{R},$$

or

$$\frac{H}{\rho} = -\frac{s}{R} + C'.$$

Also, the function  $\Psi = \Psi(\mathbf{U})$  (see (7.8)) satisfies

$$\begin{aligned} \Psi &= \int_{\mathbb{R}^d} M \log(M) \mathbf{v} d\mathbf{v} \\ &= \log\left(\rho(2\pi RT)^{-\frac{3}{2}}\right) \int_{\mathbb{R}^d} M \mathbf{v} d\mathbf{v} - \frac{1}{2RT} \int M \mathbf{v} |\mathbf{v} - \mathbf{u}|^2 d\mathbf{v} \\ &= \rho \mathbf{u} \cdot \log \rho(2\pi RT)^{-\frac{3}{2}} - \mathbf{u} \left( \frac{\rho \varepsilon}{RT} \right), \end{aligned}$$

and thus

$$\Psi = \mathbf{u} H.$$

Now, from (7.5), (7.7), and the first Eq. (7.14), we obtain the entropy equality

$$\frac{\partial}{\partial t} H(\mathbf{U}) + \operatorname{div}_{\mathbf{x}} \mathbf{u} H(\mathbf{U}) = 0,$$

which is satisfied (for smooth  $\mathbf{U}$ ) since  $M$  is a Maxwellian. Together with Lemma 7.1, this proves that  $-\rho S$  is convex and is indeed an entropy function (in the sense of Lax, see the Chap. I, Sect. 5, or Chap. II, Sect. 5), with entropy flux  $-\rho \mathbf{u} S$ . It is known that the entropies may be written as  $\rho G(\frac{\rho^{\gamma-1}}{\varepsilon})$  or  $\rho G(\frac{\rho^{\gamma-1}}{T})$  (here we have  $\gamma = \frac{5}{3}$ ) for some function  $G$ .

When the gas is not assumed to be monatomic, we must increase the dimension of the phase space. Let us present a more general model where the distribution function  $f$  depends not only on  $\mathbf{x}$  and  $\mathbf{v}$  but also on an internal energy (or temperature) microscopic variable  $\theta \in \mathbb{R}_+$ , with density  $n(\theta)$ . Therefore, we have  $f = f(t, \mathbf{x}, \mathbf{v}, \theta) = f(\mathbf{v}, \theta)$  (with shorthand notations) and

$$(\rho, \rho \mathbf{u}, \rho e)^T = \int_{\mathbb{R}^d \times \mathbb{R}_+} f(\mathbf{v}, \theta) \left( 1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta \right)^T n(\theta) d\theta d\mathbf{v}, \quad (7.16a)$$

where  $e = \varepsilon + \frac{|\mathbf{u}|^2}{2}$ ,

$$\rho\varepsilon = \int_{\mathbb{R}^d \times \mathbb{R}_+} f(\mathbf{v}, \theta) \left\{ \frac{|\mathbf{v} - \mathbf{u}|^2}{2} + \theta \right\} n(\theta) d\theta d\mathbf{v} = \rho(\varepsilon_k + \varepsilon_i), \quad (7.16b)$$

and

$$\begin{cases} \rho\varepsilon_k = \int_{\mathbb{R}^d \times \mathbb{R}_+} f(\mathbf{v}, \theta) \frac{|\mathbf{v} - \mathbf{u}|^2}{2} n(\theta) d\theta d\mathbf{v}, \\ \rho\varepsilon_i = \int_{\mathbb{R}^d \times \mathbb{R}_+} f(\mathbf{v}, \theta) \theta n(\theta) d\theta d\mathbf{v}. \end{cases} \quad (7.16c)$$

We define the temperature by

$$T = \frac{2\varepsilon_k}{dR}. \quad (7.17)$$

Now, for simplicity, we consider an internal energy  $\varepsilon_i$ , which has the form

$$\varepsilon_i = \delta RT, \quad (7.18a)$$

where the coefficient  $\delta > 0$  satisfies

$$\frac{1}{\gamma - 1} = \delta + \frac{d}{2} \iff \delta = \frac{2 - d(\gamma - 1)}{2(\gamma - 1)}. \quad (7.18b)$$

Thus, from (7.16b),

$$\varepsilon = \varepsilon_i + \varepsilon_k = RT \left( \delta + \frac{d}{2} \right) = \frac{RT}{(\gamma - 1)}.$$

This particular form of  $\varepsilon_i$  is obtained for a density of energy  $n(\theta) = \theta^{\delta-1}$ , and

$$\rho\varepsilon_i = \int_{\mathbb{R}^d \times \mathbb{R}_+} f(\mathbf{v}, \theta) \theta^\delta d\theta d\mathbf{v}. \quad (7.18)$$

The Maxwellian equilibrium function satisfying

$$(\rho, \rho\mathbf{u}, \rho e)^T = \int_{\mathbb{R}^d \times \mathbb{R}_+} M(\mathbf{v}, \theta) \left( 1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta \right)^T n(\theta) d\theta d\mathbf{v}$$

is then given by

$$M(\mathbf{v}, \theta) = (2\pi RT)^{-\frac{d}{2}} \rho \exp\left(-\frac{|\mathbf{v} - \mathbf{u}|^2}{2RT}\right) \frac{1}{\theta_0} \exp\left(-\frac{\theta}{RT}\right), \quad (7.19)$$

where

$$\theta_0 = \int_{\mathbb{R}_+} \exp\left(-\frac{\theta}{RT}\right) n(\theta) d\theta.$$

The pressure is given by (see (7.10))

$$\begin{aligned} p &= \frac{1}{d} \sum_{i=1}^d \int_{\mathbb{R}^d \times \mathbb{R}^+} M(\mathbf{v}, \theta) (v_i - u_i)^2 n(\theta) d\theta d\mathbf{v}, \\ &= \frac{1}{d} \int_{\mathbb{R}^d \times \mathbb{R}^+} M(\mathbf{v}, \theta) |\mathbf{v} - \mathbf{u}|^2 n(\theta) d\theta d\mathbf{v} = \frac{2\rho\varepsilon_k}{d}, \end{aligned} \quad (7.20a)$$

i.e.,

$$p = \rho RT = (\gamma - 1)\rho\varepsilon. \quad (7.20b)$$

We note that in (7.19),  $\log M$  depends on  $\mathbf{v}$  and  $\theta$  through  $\frac{|\mathbf{v}|^2}{2} + \theta$ . The elementary invariants in (7.3) are now  $(1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta)$ , and the kinetic entropy is again the function  $h(r) = r \log r$ .

*Remark 7.4.* For an ideal monatomic gas in dimension  $d = 3$  (resp.  $d = 1$ ), we have  $\gamma = \frac{5}{3}$  (resp.  $\gamma = 3$ ) and  $RT = \frac{2\varepsilon}{3}$  (resp.  $RT = 2\varepsilon$ ), and by (7.18b),  $\delta = 0$ . Thus by (7.18a), the term  $\varepsilon_i$  vanishes, which means that we drop in that case the dependence on  $\theta$ .

Otherwise,  $\gamma = \frac{\delta+d+2}{\delta+d} < \frac{5}{3}$  (in  $d = 3$ ) if  $\delta > 0$ .  $\square$

## 7.2 The B.G.K. Model

We present now a simplified collision term that is at the source of the B.G.K. model (from P.L. Bhatnagar, E.P. Gross, and M. Krook (1954)). The B.G.K. model describes the evolution of  $f$  through the equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \operatorname{grad} f = \frac{M(\mathbf{v}) - f}{\nu}, \quad (7.21)$$

where  $M(\mathbf{v}; \rho, \mathbf{u}, T) = M(\mathbf{v})$  is the (local) Maxwellian (see (7.13)) given by

$$M(\mathbf{v}) = \left(2\pi RT(\mathbf{x}, t)\right)^{-\frac{d}{2}} \rho(\mathbf{x}, t) \exp\left(-\frac{|\mathbf{v} - \mathbf{u}|^2}{2RT}(\mathbf{x}, t)\right) \quad (7.22)$$

or by (7.19) if we consider the dependence on the internal energy  $\theta$ . In (7.21), the collision term

$$J(f) = \frac{1}{\nu}(M(\mathbf{v}) - f)$$

is constructed in order to satisfy the following properties that were proved for  $Q$ :

$$\int_{\mathbb{R}^d} J(f) \mathbf{K}(\mathbf{v}) d\mathbf{v} = 0, \quad \forall f \geq 0, \quad (7.23)$$

where  $\mathbf{K}(\mathbf{v})$  is the vector of collision invariants (7.3), and

$$\int_{\mathbb{R}^d} \log f J(f) d\mathbf{v} \leq 0, \quad (7.24)$$

with equality if and only if  $f$  is a Maxwellian  $M$ . The constant  $\nu$  is a small parameter meant to tend to 0 ( $\frac{1}{\nu}$  is the collision frequency, and  $\nu$  can also be viewed as the relaxation time). Thus, from (7.23), we get

$$\int_{\mathbb{R}^d} M(\mathbf{v}) \mathbf{K}(\mathbf{v}) d\mathbf{v} = \int_{\mathbb{R}^d} f(\mathbf{v}) \mathbf{K}(\mathbf{v}) d\mathbf{v},$$

and the parameters  $\rho, \mathbf{u}, E = \rho(\varepsilon + \frac{|\mathbf{v}|^2}{2})$  in the Maxwellian are indeed the moments of  $f$ , given by (7.9),

$$\mathbf{U}(\mathbf{x}, t) = \begin{pmatrix} \rho \\ \rho \mathbf{u} \\ \rho \varepsilon \end{pmatrix} (\mathbf{x}, t) = \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) \begin{pmatrix} 1 \\ \mathbf{v} \\ |\mathbf{v}|^2/2 \end{pmatrix} d\mathbf{v},$$

or by (7.16) if  $M$  is defined by (7.19). As  $\nu$  tends to 0,  $f$  is expected to tend to the Maxwellian distribution  $M(\mathbf{v}; \mathbf{U})$  whose moments satisfy the Euler equations (7.14), with the equation of state (7.10) or (7.11):  $p = \rho RT = 2\rho\varepsilon/3$ ,  $\varepsilon = \frac{3RT}{2}$ . (For more precise existence and stability results, we refer to Perthame [940].) As we have noticed in Remark 7.1, the entropy  $\mathcal{H}(f) = \int_{\mathbb{R}^d} H(f) d\mathbf{x} = \int_{\mathbb{R}^{2d}} h(f) d\mathbf{v} d\mathbf{x}$ , where  $h(f) = f \log f$ , decreases as  $t \rightarrow +\infty$  toward a minimum, which corresponds to a Maxwellian distribution. Thus  $M$  realizes the minimum of  $\mathcal{H}(f)$  under the constraints  $\int_{\mathbb{R}^d} f(\mathbf{v})(1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2})^T d\mathbf{v} = (\rho, \mathbf{0}, \rho\varepsilon)^T$ .

Now it is interesting, in view of the numerical applications, to consider other entropy functions and their associated “equilibrium” functions. Given a “kinetic entropy”  $h$ , we can build a BGK model, i.e., find an equilibrium function  $N$  that realizes the minimum of the entropy  $\int h$ , and consider Eq. (7.21) associated with  $N$ , i.e., with BGK kernel  $\frac{N-f}{\nu}$  in the right-hand side.

For this, given a function  $h(f, \theta)$ , a “kinetic entropy,” which is a strictly convex function of  $f$  (satisfying  $h(0, \theta) = 0$  and say  $h'(0, \theta)$  or some more stringent conditions that we do not discuss), we consider the problem of minimization,

$$H(\rho, \varepsilon) \equiv \min \left\{ \int_{\mathbb{R}^d \times \mathbb{R}_+} h(f(\mathbf{v}, \theta), \theta) n(\theta) d\theta d\mathbf{v} \right\}, \quad (7.25)$$

where the minimum is taken over all  $f \geq 0$  satisfying the constraints

$$\int_{\mathbb{R}^d \times \mathbb{R}_+} f(\mathbf{v}, \theta) \left( 1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta \right)^T n(\theta) d\theta d\mathbf{v} = (\rho, \mathbf{0}, \rho\varepsilon)^T. \quad (7.26)$$

*Lemma 7.2*

*Problem (7.25) and (7.26) admits a unique solution  $N$ .*

*Proof.* We just sketch the proof. It can be shown that  $H$  is nonincreasing in  $\varepsilon$  and has a unique minimum obtained for a function  $f = N(\mathbf{v}, \theta; \rho, \varepsilon)$  such that

$$h'(N) = a - b\left(\frac{|\mathbf{v}|^2}{2} + \theta\right)\left(\text{or}\left[a - b\left(\frac{|\mathbf{v}|^2}{2} + \theta\right)\right]_+\right),$$

where  $h' = \frac{\partial h}{\partial f}$  and the constants  $a$  and  $b$  are such that  $N$  satisfies the constraints (7.26). Indeed, provided the constants  $a$  and  $b$  are found, we have, since  $h'$  is strictly convex,

$$h(f) \geq h(N) + h'(N)(f - N),$$

which yields

$$\begin{aligned} \int_{\mathbb{R}^d \times \mathbb{R}_+} h(f)n(\theta)d\theta d\mathbf{v} &\geq \int_{\mathbb{R}^d \times \mathbb{R}_+} h(N)n(\theta)d\theta d\mathbf{v} \\ &+ \int_{\mathbb{R}^d \times \mathbb{R}_+} \left(a - b\left(\frac{|\mathbf{v}|^2}{2} + \theta\right)\right)(f - N)n(\theta)d\theta d\mathbf{v}, \end{aligned}$$

and the last integral vanishes. For details, we refer to Perthame [941] and Coron and Perthame [363].  $\square$

Let us observe that by construction,

$$\int_{\mathbb{R}^d \times \mathbb{R}_+} N(\mathbf{v} - \mathbf{u})\left(1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta\right)^T n(\theta)d\theta d\mathbf{v} = \left(\rho, \rho\mathbf{u}, \rho\varepsilon + \rho\frac{|\mathbf{u}|^2}{2}\right)^T.$$

We then consider the B.G.K. model, Eq. (7.21), associated with this function,  $N(\mathbf{v}, \theta; \rho, \varepsilon)$ ,

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \text{grad}_{\mathbf{x}} f = \frac{N(\mathbf{v} - \mathbf{u}) - f}{\nu}. \quad (7.27)$$

Setting

$$\left(\rho, \rho\mathbf{u}, \rho\varepsilon + \rho\frac{|\mathbf{u}|^2}{2}\right)^T = \int_{\mathbb{R}^d \times \mathbb{R}_+} f(\mathbf{v}, \theta)\left(1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta\right)^T n(\theta)d\theta d\mathbf{v}, \quad (7.28)$$

we have

$$\int_{\mathbb{R}^d \times \mathbb{R}_+} (f - N(\mathbf{v} - \mathbf{u}))\left(1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta\right)^T n(\theta)d\theta d\mathbf{v} = \mathbf{0}.$$

It can be proved that (7.27) admits  $h$  as a kinetic entropy, i.e., we have formally

$$\frac{d}{dt} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}_+} h(f)n(\theta)d\theta d\mathbf{v} d\mathbf{x} \leq 0$$

for any solution  $f$  of (7.27).

*Proposition 7.2.*

Let  $N$  be the unique solution of problem (7.25) and (7.26). As  $\nu$  tends to 0, the quantities  $\rho, \rho\mathbf{u}, \rho\varepsilon$  defined by (7.28) satisfy (formally) the Euler system (7.14) with the equation of state

$$p(\rho, \varepsilon) = \frac{1}{d} \int_{\mathbb{R}^d \times \mathbb{R}_+} N(\mathbf{v}, \theta; \rho, \varepsilon) |\mathbf{v} - \mathbf{u}|^2 n(\theta) d\theta d\mathbf{v}.$$

Moreover, the system admits  $H$  defined by (7.25) as a convex entropy, i.e.,

$$\frac{\partial H}{\partial t} + \operatorname{div}(H\mathbf{u}) = 0$$

for any smooth solution of the Euler equations (7.14).

By “formally” we mean that we assume that as  $\nu \rightarrow 0, f \rightarrow N$  in a convenient sense, which enables us to pass to the limit in the integrals!

*Proof.* By multiplying Eq. (7.27) by  $(1, \mathbf{v}, \frac{|\mathbf{v}|^2}{2} + \theta)$  and integrating w.r.t.  $\mathbf{v}$  and  $\theta$ , we obtain the first equation

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d \times \mathbb{R}_+} f d\mathbf{v} d\theta + \sum_{j=1}^d \frac{\partial}{\partial x_j} \int_{\mathbb{R}^d \times \mathbb{R}_+} v_j f n(\theta) d\theta d\mathbf{v} = 0,$$

and the remaining two equations are

$$\begin{cases} \frac{\partial}{\partial t} \int_{\mathbb{R}^d \times \mathbb{R}_+} f \mathbf{v} d\mathbf{v} d\theta + \sum_{j=1}^d \frac{\partial}{\partial x_j} \int_{\mathbb{R}^d \times \mathbb{R}_+} v_j f \mathbf{v} n(\theta) d\theta d\mathbf{v} = \mathbf{0}, \\ \frac{\partial}{\partial t} \int_{\mathbb{R}^d \times \mathbb{R}_+} f \left( \frac{|\mathbf{v}|^2}{2} + \theta \right) d\mathbf{v} n(\theta) d\theta \\ \quad + \sum_{j=1}^d \frac{\partial}{\partial x_j} \int_{\mathbb{R}^d \times \mathbb{R}_+} v_j f \left( \frac{|\mathbf{v}|^2}{2} + \theta \right) n(\theta) d\theta d\mathbf{v} = 0. \end{cases}$$

We want to prove that the system obtained by letting  $\nu \rightarrow 0$  is the Euler system. In fact, the first equation is exactly

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) = 0.$$

We now consider the other equations; using

$$\int_{\mathbb{R}^d \times \mathbb{R}_+} (v_i - u_i) f n(\theta) d\theta d\mathbf{v} = 0,$$

they can be written as

$$\frac{\partial}{\partial t}(\rho u_i) + \sum_{j=1}^d \frac{\partial}{\partial x_j} \left\{ \rho u_j u_i + \int_{\mathbb{R}^d \times \mathbb{R}^+} (v_j - u_j)(v_i - u_i) f n(\theta) d\theta d\mathbf{v} \right\} = 0.$$

$1 \leq i \leq d$ , and, with  $E = \rho\varepsilon + \rho \frac{|\mathbf{u}|^2}{2}$ ,

$$\frac{\partial E}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \left\{ E u_j + \int_{\mathbb{R}^d \times \mathbb{R}^+} (v_j - u_j) \left( \frac{|\mathbf{v}|^2}{2} + \theta \right) f n(\theta) d\theta d\mathbf{v} \right\} = 0$$

or

$$\begin{aligned} \frac{\partial E}{\partial t} + \operatorname{div}(E\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} \int_{\mathbb{R}^d \times \mathbb{R}^+} (v_j - u_j) & \left\{ \sum_{i=1}^d u_i (v_i - u_i) \right. \\ & \left. + \frac{|\mathbf{v} - \mathbf{u}|^2}{2} + \theta \right\} f n(\theta) d\theta d\mathbf{v} = 0. \end{aligned}$$

Now, we have assumed that as  $\nu \rightarrow 0$ ,  $f(\mathbf{v}, \theta) \rightarrow N(\mathbf{v} - \mathbf{u}, \theta)$ ; using, moreover, the fact that by symmetry

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} v_i v_j N(\mathbf{v}, \theta) d\mathbf{v} d\theta = 0, \quad i \neq j \text{ ( corresponding to } \pi_{ij}),$$

and

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} v_i \left( \frac{|\mathbf{v}|^2}{2} + \theta \right) N(\mathbf{v}, \theta) n(\theta) d\theta d\mathbf{v} = 0, \quad \forall i, \text{ ( corresponding to } Q_i),$$

we obtain, for  $1 \leq i \leq d$ ,

$$\frac{\partial}{\partial t}(\rho u_i) + \operatorname{div}(\rho \mathbf{u} u_i) + \sum_{j=1}^d \frac{\partial}{\partial x_j} \int_{\mathbb{R}^d \times \mathbb{R}^+} N(\mathbf{v}, \theta) v_j^2 n(\theta) d\theta d\mathbf{v} = 0$$

and

$$\frac{\partial E}{\partial t} + \operatorname{div}(E\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} u_j \int_{\mathbb{R}^d \times \mathbb{R}^+} v_j^2 N(\mathbf{v}, \theta) n(\theta) d\theta d\mathbf{v} = 0.$$

We finally obtain, as expected, the Euler system

$$\begin{cases} \frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) = 0, \\ \frac{\partial}{\partial t}(\rho u_i) + \operatorname{div}(\rho \mathbf{u} u_i) + \frac{\partial p}{\partial x_i} = 0, \quad 1 \leq i \leq d, \\ \frac{\partial}{\partial t} \left( \rho \left( \varepsilon + \frac{|\mathbf{u}|^2}{2} \right) \right) + \operatorname{div} \left( \rho \mathbf{u} \left( \varepsilon + \frac{|\mathbf{u}|^2}{2} \right) + p \mathbf{u} \right) = 0 \end{cases}$$

with

$$p = \frac{2}{d} \int_{\mathbb{R}^d \times \mathbb{R}_+} N(\mathbf{v}, \theta) \frac{|\mathbf{v}|^2}{2} n(\theta) d\theta d\mathbf{v} = \int_{\mathbb{R}^d \times \mathbb{R}_+} N(\mathbf{v}, \theta) v_j^2 n(\theta) d\theta d\mathbf{v}.$$

Moreover, it can be proved that the function  $H(\rho, \varepsilon)$  defined in (7.25), which is thus given by

$$H(\rho, \varepsilon) = \int_{\mathbb{R}^d \times \mathbb{R}_+} h(N(\mathbf{v}, \theta), \theta) n(\theta) d\theta d\mathbf{v},$$

is a convex function of  $(\rho, \varepsilon)$  (the proof follows the lines of Lemma 7.1) and

$$\frac{\partial H}{\partial t} + \operatorname{div}(H \mathbf{u}) = 0$$

for any smooth solution of the Euler equation (7.14), which means that  $H$  is a convex entropy for the Euler system.  $\square$

*Example 7.1.* If we choose the functions

$$h_n(f) = \frac{f^{n+1}}{(n+1)},$$

the associated  $M_n$ , which satisfy

$$h'_n(M_n) = \left[ a_n - b_n \left( \frac{|\mathbf{v}|^2}{2} + \theta \right) \right]_+,$$

are thus the functions

$$M_n(\mathbf{v}, \theta) = (a_n)^{\frac{1}{n}} \left[ 1 - \frac{b_n}{a_n} \left( \frac{|\mathbf{v}|^2}{2} + \theta \right) \right]_+^{\frac{1}{n}}$$

for some constants  $a_n, b_n$  such that (7.26) holds. As  $n \rightarrow +\infty$ ,

$$M_n(\mathbf{v}, \theta) \rightarrow a 1_{\{\frac{|\mathbf{v}|^2}{2} + \theta \leq b\}},$$

where the constants  $a$  and  $b$  are again given by the constraints (7.26) and  $1_X$  denotes the usual characteristic function of a set  $X$ .  $\square$

### 7.3 The Kinetic Scheme

In this section, we restrict ourselves to the one-dimensional case  $d = 1$ . The Maxwellian distribution is then, by (7.19),  $M(v - u, \theta; \rho, \varepsilon)$  if we set

$$M(v, \theta; \rho, \varepsilon) = \frac{\rho}{\theta_0} (2\pi RT)^{-\frac{1}{2}} \exp\left\{-\frac{v^2}{2RT} - \frac{\theta}{RT}\right\}. \quad (7.29)$$

The temperature  $T$  is a convenient variable, which is related to  $\varepsilon$  in the simple case we are considering by

$$RT = (\gamma - 1)\varepsilon$$

and

$$\varepsilon_i = \delta RT,$$

where  $\delta$  is defined by (7.18). The constant  $R$  is often incorporated in  $T$ ; in that case,  $T = (\gamma - 1)\varepsilon$  and  $p = \rho T$ .

### 7.3.1 The Time Discretization

A numerical scheme in *time* is obtained by the “operator splitting” of the transport and collision parts of the equation. Roughly, it consists of two steps (transport + relaxation); the first step solves

$$\frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} = 0 \quad (7.30)$$

(collisionless molecule transport), and the second step solves

$$J(f) = 0$$

(relaxation to thermodynamic equilibrium, whose solution is a local Maxwellian). More precisely, given initial conditions  $\rho_0 = \rho(x, 0)$ ,  $u_0 = u(x, 0)$ , and  $\varepsilon_0 = \varepsilon(x, 0)$ , we define an “equilibrium function”

$$f^0(x, v, \theta) = M(v - u_0, \theta; \rho_0, \varepsilon_0),$$

where  $M$  is the exact Maxwellian (7.29) or a function that is more convenient for numerical purposes, as we shall detail below. For the first step, we solve the linear transport equation (7.30),

$$\begin{cases} \frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} = 0, & t \in ]0, \Delta t], x \in \mathbb{R}, \\ f(x, 0; v, \theta) = f^0(x, v, \theta), & x \in \mathbb{R} \end{cases}$$

and then define the updated quantities

$$\begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix}(x, t) = \int_{\mathbb{R}} f(x, t; v, \theta) \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} + \theta \end{pmatrix} n(\theta) d\theta dv,$$

where

$$f(x, t; v, \theta) = f^0(x - vt, v, \theta).$$

For the next step, we start with  $\rho_1 = \rho(x, \Delta t)$ ,  $u_1 = u(x, \Delta t)$ , and  $\varepsilon_1 = \varepsilon(x, \Delta t)$  and follow the same procedure.

In fact, it is more convenient to drop the dependence on  $\theta$  by integrating first w.r.t.  $\theta$ , and, moreover, we get a more general formalism: we introduce instead of  $f^0$  the functions  $f_0, g_0$  defined by

$$\begin{cases} f_0(x, v) = \int_{\mathbb{R}^+} M(v - u_0, \theta; \rho_0, \varepsilon_0) n(\theta) d\theta, \\ g_0(x, v) = \int_{\mathbb{R}^+} M(v - u_0, \theta; \rho_0, \varepsilon_0) \theta n(\theta) d\theta. \end{cases} \quad (7.31)$$

We see from (7.16) and (7.18) that

$$\int_{\mathbb{R}} g_0(x, v) dv = \int_{\mathbb{R} \times \mathbb{R}^+} M(v - u_0, \theta; \rho_0, \varepsilon_0) \theta n(\theta) d\theta dv = \rho_0 \varepsilon_i = \delta \rho_0 R T_0,$$

and thus

$$\int_{\mathbb{R}} g_0(x, v) dv = \delta R T_0 \int_{\mathbb{R}} f_0(x, v) dv. \quad (7.32)$$

Also, we have

$$\begin{pmatrix} \rho_0 \\ \rho_0 u_0 \\ \rho_0 e_0 \end{pmatrix} (x) = \int_{\mathbb{R}} \begin{pmatrix} f_0 \\ v f_0 \\ \frac{v^2}{2} f_0 + g_0 \end{pmatrix} (x, v) dv. \quad (7.33)$$

Next, we solve

$$\begin{cases} \frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} = 0, & t \in ]0, \Delta t], x \in \mathbb{R}, \\ f(x, 0; v) = f_0(x, v), \end{cases} \quad (7.34a)$$

$$\begin{cases} \frac{\partial g}{\partial t} + v \cdot \frac{\partial g}{\partial x} = 0, & t \in ]0, \Delta t], x \in \mathbb{R}, \\ g(x, 0; v) = g_0(x, v), \end{cases} \quad (7.34b)$$

and then for  $t \leq \Delta t$ ,

$$\mathbf{U}(x, t) = \begin{pmatrix} \rho \\ \rho u \\ \rho e \end{pmatrix} (x, t) = \int_{\mathbb{R}} \begin{pmatrix} f \\ v f \\ (\frac{v^2}{2}) f + g \end{pmatrix} (x, t; v) dv. \quad (7.35)$$

For a general step, if  $f_n$  and  $g_n$  are known, we solve

$$\begin{cases} \frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} = 0, & t \in ]t_n, t_{n+1}], x \in \mathbb{R}, \\ f(x, t_n + 0; v) = f_n(x, v) \end{cases} \quad (7.36a)$$

and

$$\begin{cases} \frac{\partial g}{\partial t} + v \cdot \frac{\partial g}{\partial x} = 0, & t \in ]t_n, t_{n+1}], x \in \mathbb{R}, \\ g(x, t_n + 0; v) = g_n(x, v), \end{cases} \quad (7.36b)$$

which gives

$$\begin{aligned} f(x, t; v) &= f_n(x - v(t - t_n), v), & t \in ]t_n, t_{n+1}], \\ g(x, t; v) &= g_n(x - v(t - t_n), v). \end{aligned}$$

We define  $\rho_{n+1}, (\rho u)_{n+1}, (\rho e)_{n+1}$  by

$$\begin{pmatrix} \rho_{n+1} \\ (\rho u)_{n+1} \\ (\rho e)_{n+1} \end{pmatrix}(x) = \int_{\mathbb{R}} \begin{pmatrix} f_n \\ v f_n \\ \frac{v^2}{2} f_n + g_n \end{pmatrix}(x - v\Delta t, v) dv, \quad (7.37)$$

and we associate with these quantities the equilibrium function  $M(v - u_{n+1}, \theta; \rho_{n+1}, \varepsilon_{n+1})$ . Then, we define

$$\begin{aligned} f_{n+1}(x, v) &= \int_{\mathbb{R}^+} M(v - u_{n+1}, \theta; \rho_{n+1}, \varepsilon_{n+1}) d\theta, \\ g_{n+1}(x, v) &= \int_{\mathbb{R}^+} M(v - u_{n+1}, \theta; \rho_{n+1}, \varepsilon_{n+1}) \theta d\theta. \end{aligned}$$

This says that in the second (collision) step, mass, momentum, and energy are conserved.

*Lemma 7.3*

The quantities  $\rho_{n+1}, (\rho u)_{n+1}, (\rho e)_{n+1}$  defined by (7.36) and (7.37) are first-order approximations (in time) of the solution of the Euler equation (7.14).

*Proof.* The result means that if  $(\rho_n, (\rho u)_n, (\rho e)_n) = (\rho, \rho u, \rho e)(x, t_n)$  is the exact solution of the Euler equation at time  $t_n$ , the quantities defined above are first-order approximations of  $(\rho, \rho u, \rho e)(x, t_{n+1})$ . Let us first note that we can define the quantities  $(\rho, \rho u, \rho e)$  not only at time  $t_{n+1}$  but at any time, for instance,

$$(\rho, \rho u)^T(x, t) = \int_{\mathbb{R}} f(x, v, t)(1, v)^T dv,$$

where  $f$  is the solution of (7.36). It is then easy to see that  $\rho$  is indeed the solution of (7.14). (The first equation in (7.14) is solved exactly.) The argument follows the proof of Proposition 7.2. Indeed, let us integrate the first equation (7.36) w.r.t.  $v$ ; we get

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \left( \frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} \right) dv = \frac{\partial}{\partial t} \int_{\mathbb{R}} f(x, v, t) dv + \frac{\partial}{\partial x} \int_{\mathbb{R}} v f(x, v, t) dv \\ &= \frac{\partial \rho}{\partial t} + \frac{\partial(\rho u)}{\partial x}. \end{aligned}$$

Thus, if the initial conditions are exact, so is  $\rho$  at any time  $t$ . The other quantities are only first-order approximations. Multiplying (7.36) by  $v$  and integrating w.r.t.  $v$ , we obtain

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \left( \frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} \right) v dv \\ &= \frac{\partial}{\partial t} \int_{\mathbb{R}} f(x, v, t) v dv + \frac{\partial}{\partial x} \int_{\mathbb{R}} v^2 f(x, v, t) dv \\ &= \frac{\partial}{\partial t} (\rho u) + \frac{\partial}{\partial x} 2 \left( \rho e - \int_{\mathbb{R}} g(x, v, t) dv \right). \end{aligned}$$

As in (7.32), we have exactly at time  $t_n$

$$\int_{\mathbb{R}} g(x, v, t_n) dv = \rho_n R T_n \delta,$$

and hence

$$2 \left( \rho e - \int_{\mathbb{R}} g(x, v, t_n) dv \right) = (\rho u^2 + p)(t_n),$$

which is only satisfied at the order  $\mathcal{O}(\Delta t)$  for  $t \in ]t_n, t_{n+1}]$ . For instance, we write for simplicity the first step on  $[0, \Delta t]$ , using a Taylor expansion, the definition (7.35), and Eqs. (7.34), (7.32), and (7.18):

$$\begin{aligned} \rho u(x, t) &= \rho u(x, 0) + \Delta t \frac{\partial}{\partial t} (\rho u)(x, 0) + \mathcal{O}(\Delta t^2) \\ &= \rho u(x, 0) + \Delta t \frac{\partial}{\partial t} \int_{\mathbb{R}} f(x, v, t) v dv + \mathcal{O}(\Delta t^2) \\ &= \rho u(x, 0) + \Delta t \int_{\mathbb{R}} v \frac{\partial}{\partial t} f(x, v, t) dv + \mathcal{O}(\Delta t^2) \\ &= \rho u(x, 0) - \Delta t \int_{\mathbb{R}} v^2 \frac{\partial}{\partial t} f(x, v, t) dv + \mathcal{O}(\Delta t^2), \end{aligned}$$

where the derivatives are taken at time  $t = 0$ , from which we obtain the result since

$$\begin{aligned}
\rho u(x, t) &= \rho u(x, 0) - 2\Delta t \frac{\partial}{\partial x} \left( \rho_0 e_0 - \int_{\mathbb{R}} g_0(x, v) dv \right) + \mathcal{O}(\Delta t^2) \\
&= \rho u(x, 0) - 2\Delta t \frac{\partial}{\partial x} (\rho_0 e_0 - RT_0 \delta) \int_{\mathbb{R}} f_0(x, v) dv + \mathcal{O}(\Delta t^2) \\
&= \rho u(x, 0) - 2\Delta t \frac{\partial}{\partial x} \rho_0 (e_0 - RT_0 \delta) + \mathcal{O}(\Delta t^2) \\
&= \rho u(x, 0) - \Delta t \frac{\partial}{\partial x} (\rho u^2 + p)(x, 0) + \mathcal{O}(\Delta t^2),
\end{aligned}$$

gives the exact solution  $\rho u$  of the Euler equations within  $\mathcal{O}(\Delta t^2)$ . The proof for  $\rho e$  is similar.  $\square$

For numerical purposes, it is interesting to replace  $\int_{\mathbb{R}_+} M n(\theta) d\theta$  by an expression that is easier to handle. In practice, Perthame [941] has considered a numerical scheme associated with a function  $\chi$  satisfying the following properties:

$$\begin{cases} \chi \geq 0, \\ \int_{\mathbb{R}} \chi(w) (1, w^2) dw = (1, 1), \\ \chi(-w) = \chi(w) \left( \Rightarrow \int_{\mathbb{R}} \chi(w) w dw = 0 \right). \end{cases} \quad (7.38)$$

We then introduce the functions

$$\begin{cases} f_0(x, v) = \rho_0(x) (T_0(x))^{-\frac{1}{2}} \chi \left( \frac{v - u_0(x)}{T_0^{\frac{1}{2}}(x)} \right), \\ g_0(x, v) = \delta T_0(x) f_0(x, v) = \delta \rho_0(x) (T_0(x))^{\frac{1}{2}} \chi \left( \frac{v - u_0(x)}{T_0^{\frac{1}{2}}(x)} \right), \end{cases} \quad (7.39)$$

where  $\delta$  is defined by (7.18b). In fact, we can take for  $f$  and  $g$  two functions  $\chi$  and  $\zeta$ . This will be detailed below in Sect. 7.3.3.

*Remark 7.5.* We observe that the function  $\chi$  is constructed in order to take the place of

$$(2\pi)^{-\frac{1}{2}} \exp\left(-\frac{v^2}{2}\right) = T_0^{\frac{1}{2}} \rho_0^{-1} \int_{\mathbb{R}_+} M(v; \rho_0, \varepsilon_0, v, \theta) n(\theta) d\theta,$$

where we have incorporated the constant  $R$  in the temperature. We have used the identity defining  $\theta_0$ ,

$$\int_{\mathbb{R}_+} \exp\left(-\frac{\theta}{T}\right) n(\theta) d\theta = \theta_0,$$

which implies

$$\int_{\mathbb{R}^+} M(v; \rho_0, \varepsilon_0, v, \theta) d\theta = \frac{\rho_0}{(2\pi T_0)^{\frac{1}{2}}} \exp\left(-\frac{v^2}{2T}\right).$$

Indeed, on the one hand,

$$(2\pi)^{-\frac{1}{2}} \int_{\mathbb{R}} \exp\left(-\frac{w^2}{2}\right) (1, w, w^2)^T dw = (1, 0, 1)^T$$

and, on the other hand, the last two properties of  $\chi$  imply

$$\int_{\mathbb{R}} \chi(w) (1, w, w^2)^T dw = (1, 0, 1)^T.$$

With this in mind, we see that the definition of  $f_0$  and  $g_0$  mimics (7.31) and (7.32).  $\square$

It is easy to check that (7.33) still holds, i.e.,

$$\begin{aligned} (\rho_0, \rho_0 u_0)^T &= \int_{\mathbb{R}} f_0(x, v) (1, v)^T dv, \\ \rho_0 e_0 &= \rho_0 \left( \varepsilon_0 + \frac{u_0^2}{2} \right) = \int_{\mathbb{R}} \left\{ \frac{|v|^2}{2} f_0 + g_0 \right\} (x, v) dv. \end{aligned}$$

We then solve the linear transport equation (7.34) with  $f_0, g_0$  given by (7.39) and define the updated quantities  $\mathbf{U}(x, t) = (\rho, \rho u, \rho e)(x, t)$  by (7.35). As in Lemma 7.3, we can prove the following lemma.

*Lemma 7.4*

Assume that  $f$  and  $g$  are solutions of (7.34), where  $f_0, g_0$  are given by (7.39). Then, the quantities  $(\rho, \rho u, \rho e)(x, t)$  defined for  $0 \leq t \leq \Delta t$  by (7.35) are first-order approximations (in time) of the solution of the Euler equation (7.14).

The description of a general time step follows the same lines.

### 7.3.2 The Space Discretization

For the *space discretization*, we follow the ideas of Godunov's scheme: we assume that the functions  $\rho_0, \rho_0 u_0, E_0$  are piecewise constant over the cells  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ ,

$$\mathbf{U}(x, 0) = \mathbf{U}_0(x) = \mathbf{U}_j^0 = (\rho_j^0, \rho u_j^0, \rho e_j^0), \quad x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}),$$

(which means that we have taken the cell average of the initial data). As above, we solve (7.34) and obtain  $\mathbf{U}(x, t)$  by (7.35). Lastly, we project on the grid

$$\mathbf{U}_j^1 = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \mathbf{U}(x, \Delta t) dx. \quad (7.40)$$

Thus, the scheme results in three steps: if  $\mathbf{U}_n(x) = (\mathbf{U}_j^n) = (\rho_j^n, \rho u_j^n, \rho e_j^n)$  constant over the cell  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$  is known, we compute  $\mathbf{U}_{n+1}(x) = (\mathbf{U}_j^{n+1})$  by:

(i) Reconstruction of  $f_n, g_n$  by (7.39)

$$\begin{cases} f_n(x, v) = \rho_n(T_n)^{-\frac{1}{2}} \chi \left( \frac{v - u_n}{T_n^{\frac{1}{2}}} \right), \\ g_n(x, v) = \delta T_n f_n(x, v) = \delta \rho_n T_n^{\frac{1}{2}} \chi \left( \frac{v - u_n}{T_n^{\frac{1}{2}}} \right). \end{cases} \quad (7.41a)$$

(ii) Evolution by (7.34)

$$\begin{cases} \frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} = 0, & t \in ]t_n, t_{n+1}], x \in \mathbb{R}, \\ f(x, t_n; v) = f_n(x, v), \end{cases} \quad (7.41b_1)$$

$$\begin{cases} \frac{\partial g}{\partial t} + v \cdot \frac{\partial g}{\partial x} = 0, & t \in ]t_n, t_{n+1}], x \in \mathbb{R}, \\ g(x, t_n; v) = g_n(x, v), \end{cases} \quad (7.41b_2)$$

(iii) Projection: defining  $\mathbf{U}(\cdot, t_{n+1})$  by (7.35), where  $f, g(\cdot, t_{n+1})$  are solutions of (7.41b), set

$$\mathbf{U}_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \mathbf{U}(x, t_{n+1}) dx. \quad (7.41c)$$

We assume from now on that  $\chi$  is compactly supported,

$$\chi(w) = 0, \quad \text{for } |w| > w_M. \quad (7.42)$$

As was the case for Godunov's or van Leer's scheme, we can derive a very simple formula for  $\mathbf{U}_j^n$ .

*Proposition 7.3*

Let  $\chi$  satisfy (7.38) and (7.42). Under the CFL condition

$$\lambda \max\{|u_j^n| + w_M(T_j^n)^{\frac{1}{2}}\} \leq 1, \quad \lambda = \frac{\Delta t}{\Delta x}, \quad (7.43)$$

the difference scheme (7.41) can be written as

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \lambda \{ \mathbf{F}(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n) - \mathbf{F}(\mathbf{U}_{j-1}^n, \mathbf{U}_j^n) \}, \quad (7.44a)$$

where the numerical flux  $\mathbf{F}(\mathbf{U}, \mathbf{V})$  is given by

$$\mathbf{F}(\mathbf{U}, \mathbf{V}) = \mathbf{F}^+(\mathbf{U}) + \mathbf{F}^-(\mathbf{V}) \quad (7.44b)$$

and

$$\mathbf{F}^+(\mathbf{U}) = \int_{w \geq -\frac{u}{\sqrt{T}}} \begin{pmatrix} w\sqrt{T} + u \\ (w\sqrt{T} + u)^2 \\ \frac{1}{2}(w\sqrt{T} + u)^3 + \delta T(w\sqrt{T} + u) \end{pmatrix} \rho \chi(w) dw, \quad (7.44c)$$

$$\mathbf{F}^-(\mathbf{U}) = \int_{w \leq -\frac{u}{\sqrt{T}}} \begin{pmatrix} w\sqrt{T} + u \\ (w\sqrt{T} + u)^2 \\ \frac{1}{2}(w\sqrt{T} + u)^3 + \delta T(w\sqrt{T} + u) \end{pmatrix} \rho \chi(w) dw. \quad (7.44d)$$

*Proof.* Let us integrate (7.34) w.r.t.  $(v, x, t)$  over  $\mathbb{R} \times (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times (0, \Delta t)$ :

$$\begin{aligned} 0 &= \int_{\mathbb{R} \times (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times (0, \Delta t)} \left( \frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} \right) dv dx dt \\ &= \int_{\mathbb{R} \times (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})} \{f(x, v, \Delta t) - f(x, v, 0)\} dv dx \\ &\quad + \int_{\mathbb{R} \times (0, \Delta t)} v \{f(x_{j+\frac{1}{2}}, v, t) - f(x_{j-\frac{1}{2}}, v, t)\} dv dt. \end{aligned}$$

By (7.35) and (7.40), the first integral on the right-hand side is equal to

$$\int_{(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})} \{\rho(x, \Delta t) - \rho(x, 0)\} dx = \Delta x (\rho_j^1 - \rho_j^0).$$

The second integral is

$$\begin{aligned} &\int_{\mathbb{R} \times (0, \Delta t)} v \left\{ f_0(x_{j+\frac{1}{2}} - vt, v) - f_0(x_{j-\frac{1}{2}} - vt, v) \right\} dv dt \\ &= \int_{\mathbb{R}} \left\{ \int_{(0, \Delta t)} v \{f_0(x_{j+\frac{1}{2}} - vt, v) - f_0(x_{j-\frac{1}{2}} - vt, v)\} dv \right\} dt, \end{aligned}$$

with

$$f_0(x, v) = \rho_0(T_0(x))^{-\frac{1}{2}} \chi \left( \frac{v - u_0(x)}{T_0^{\frac{1}{2}}(x)} \right).$$

Setting

$$w = \frac{v - u_0}{\sqrt{T_0}},$$

the condition  $|w| \leq w_M$ , where  $[-w_M, w_M]$  is the support of  $\chi$ , means  $|\frac{(v-u_0)}{\sqrt{T_0}}| \leq w_M$  and implies that

$$|v| \leq |u_0| + w_M \sqrt{T_0}.$$

Now, by the CFL condition (7.43),

$$\Delta t(|u_0| + \sqrt{T_0}w_M) \leq \Delta x,$$

the integrals are limited to the  $v$  such that  $|v| \leq \frac{\Delta x}{\Delta t}$ . This yields that the quantity  $x_{j+\frac{1}{2}} - vt$  remains for  $t \in (0, \Delta t)$ , and  $v > 0$ , in the cell  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ , where  $\mathbf{U}_0(x)$  is constant and given by  $\mathbf{U}_j^0 = (\rho_j^0, \rho u_j^0, \rho e_j^0)$  (resp. for  $v < 0$  in the cell  $(x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}})$ , where  $\mathbf{U}_0(x) = \mathbf{U}_{j+1} = (\rho_{j+1}^0, \rho u_{j+1}^0, \rho e_{j+1}^0)$ ). We can thus write

$$\begin{aligned} & \int_{\mathbb{R}} \left\{ \int_{(0, \Delta t)} v f_0(x_{j+\frac{1}{2}} - vt, v) dt \right\} dv \\ &= \int_{v>0} \left\{ \int_{(0, \Delta t)} v f_0(x_{j+\frac{1}{2}} - vt, v) dt \right\} dv \\ &\quad + \int_{v \leq 0} \left\{ \int_{(0, \Delta t)} v f_0(x_{j+\frac{1}{2}} - vt, v) dt \right\} dv \\ &= \Delta t \left\{ \int_{v>0} \rho_j^0 (T_j^0)^{-\frac{1}{2}} \chi \left( \frac{v - u_j^0}{\sqrt{T_j^0}} \right) dv \right. \\ &\quad \left. + \int_{v \leq 0} \rho_{j+1}^0 (T_{j+1}^0)^{-\frac{1}{2}} \chi \left( \frac{v - u_{j+1}^0}{\sqrt{T_j^0}} \right) dv \right\}. \end{aligned}$$

After a change of variable  $w = \frac{(v-u)}{\sqrt{T}}$ , we get by (7.44c) the first component of  $\mathbf{F}^+(\mathbf{U}_j^0) + \mathbf{F}^-(\mathbf{U}_{j+1}^0)$ . Similarly,

$$\begin{aligned}
& \int_{\mathbb{R}} \left\{ \int_{(0, \Delta t)} v f_0(x_{j-\frac{1}{2}} - vt, v) dt \right\} dv \\
&= \int_{v>0} \left\{ \int_{(0, \Delta t)} v f_0(x_{j-\frac{1}{2}} - vt, v) dt \right\} dv \\
&\quad + \int_{v \leq 0} \left\{ \int_{(0, \Delta t)} v f_0(x_{j-\frac{1}{2}} - vt, v) dt \right\} dv \\
&= \Delta t \left\{ \int_{v>0} \rho_{j-1}^0 (T_{j-1}^0)^{-\frac{1}{2}} \chi \left( \frac{v - u_{j-1}^0}{\sqrt{T_j^0}} \right) dv \right. \\
&\quad \left. + \int_{v \leq 0} \rho_j^0 (T_j^0)^{-\frac{1}{2}} \chi \left( \frac{v - u_j^0}{\sqrt{T_j^0}} \right) dv \right\},
\end{aligned}$$

which gives the first component of  $\mathbf{F}^+(\mathbf{U}_{j-1}^0) + \mathbf{F}^-(\mathbf{U}_j^0)$ .

The other components are computed in the same way. This yields

$$\Delta x (\mathbf{U}_j^1 - \mathbf{U}_j^0) + \Delta t \{ \mathbf{F}^+(\mathbf{U}_j^0) + \mathbf{F}^-(\mathbf{U}_{j+1}^0) - \mathbf{F}^+(\mathbf{U}_{j-1}^0) - \mathbf{F}^-(\mathbf{U}_j^0) \} = 0,$$

and proves that the scheme is indeed given by (7.44).

The formulas for the fluxes can be written in the more compact form

$$\mathbf{F}^\pm(\mathbf{U}) = \rho \int_{\mathbb{R}} v_\pm \left( 1, v, \frac{|v|^2}{2} + \delta T \right)^T \chi \left( \frac{v - u}{\sqrt{T}} \right) \frac{dv}{\sqrt{T}}, \quad (7.45)$$

which gives a flux vector splitting scheme whose numerical flux is clearly consistent with the flux of the Euler equations.  $\square$

An important feature of this scheme is its positivity-preserving property.

*Theorem 7.2*

Assume the hypotheses of Proposition 7.3. Then, the difference scheme (7.44) satisfies

$$\rho_j^0 \geq 0, T_j^0 \geq 0, \forall j \Rightarrow \rho_j^n, T_j^n \geq 0, \quad \forall j, \forall n > 0, \quad (7.46)$$

and is  $L^1$ -stable.

*Proof.* Assume

$$\rho_j^0 \geq 0, T_j^0 \geq 0, \quad \forall j,$$

and consider the first step. The functions  $f_0$  and  $g_0$  in (7.39) are thus non-negative, and  $f, g$  remain such after the transport by (7.34). Now consider the moments; by (7.41c)

$$\rho_j^1 = \frac{1}{\Delta x} \int_{\mathbb{R} \times (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})} f_0(x - v\Delta t, v) dv dx \geq 0,$$

and

$$E_j^1 = \frac{1}{\Delta x} \int_{\mathbb{R} \times (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})} \left\{ \frac{v^2}{2} f_0(x - v\Delta t, v) + g_0(x - v\Delta t, v) \right\} dv dx.$$

We can write

$$\begin{aligned} \Delta x E_j^1 &\geq \int_{\mathbb{R} \times (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})} \frac{1}{2} \{ (u_j^1)^2 + (v - u_j^1)^2 \\ &\quad + 2u_j^1(v - u_j^1) \} f_0(x - v\Delta t, v) dv dx \\ &\geq (u_j^1)^2 \frac{\rho_j^1}{2} + u_j^1 \int_{\mathbb{R} \times (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})} (v - u_j^1) f_0(x - v\Delta t, v) dv dx, \end{aligned}$$

and the last integral vanishes. Since

$$E_j^1 = \rho_j^1 \left\{ \varepsilon_j^1 + \frac{(u_j^1)^2}{2} \right\},$$

this yields

$$RT_j^1 = (\gamma - 1)\varepsilon_j^1 \geq 0.$$

Consider next the  $L^1$  stability. Since the quantities are nonnegative and since the scheme (7.44) is obviously conservative, we have

$$\|\rho_1\|_{\mathbf{L}^1(\mathbb{R})} = \sum_{j \in \mathbb{Z}} \rho_j^1 = \sum_{j \in \mathbb{Z}} \rho_j^0 = \|\rho_0\|_{\mathbf{L}^1(\mathbb{R})},$$

and similarly

$$\sum_{j \in \mathbb{Z}} E_j^1 = \|E_0\|_{\mathbf{L}^1(\mathbb{R})}.$$

Then

$$\begin{aligned} \sum_{j \in \mathbb{Z}} \rho_j^1 |u_j^1| &= \|\rho_1 u_1\|_{\mathbf{L}^1(\mathbb{R})} \\ &\leq (\|\rho_1\|_{\mathbf{L}^1(\mathbb{R})} \|\rho_1(u_1)^2\|_{\mathbf{L}^1(\mathbb{R})})^{\frac{1}{2}} \leq (2\|\rho_0\|_{\mathbf{L}^1(\mathbb{R})} \|E_1\|_{\mathbf{L}^1(\mathbb{R})})^{\frac{1}{2}} \\ &\leq \|\rho_0\|_{\mathbf{L}^1(\mathbb{R})} + \|E_0\|_{\mathbf{L}^1(\mathbb{R})}, \end{aligned}$$

which ends the proof.  $\square$

*Example 7.2.* Let us take for  $\chi$  the function

$$\chi(w) = \frac{1}{2\sqrt{3}} \mathbf{1}_{\{|w| \leq \sqrt{3}\}},$$

which is the unique step function satisfying (7.38). The integrals (7.44c) are then very easily computed, and in the case  $\delta = 0, \gamma = 3$ , we recover van Leer's flux splitting formulas (Sect. 5.3). Indeed, from (7.45), we get

$$\begin{aligned}\mathbf{F}^+(\mathbf{U}) &= \left(\frac{\rho}{2\sqrt{3T}}\right) \int_0^{u+\sqrt{3T}} \left(v, v^2, \frac{v^3}{2}\right)^T dv \\ &= \left(\frac{\rho}{2\sqrt{3T}}\right) (u + \sqrt{3T})^2 \left(\frac{1}{2}, \frac{(u + \sqrt{3T})}{3}, \frac{(u + \sqrt{3T})^2}{8}\right)^T.\end{aligned}$$

Now, from the identity

$$p = \frac{\rho c^2}{\gamma},$$

we obtain

$$c^2 = \gamma T = 3T,$$

so that formulas (5.11)–(5.13) of Sect. 5, giving the components of van Leer's split flux, yield

$$\begin{aligned}F_1+ &= \frac{\rho}{4c}(u+c) = \frac{\rho}{4\sqrt{3T}}(u + \sqrt{3T})^2, \\ F_2+ &= \frac{2}{3}(u+c)F_1+ = \frac{\rho}{6\sqrt{3T}}(u + \sqrt{3T})^3, \\ F_3+ &= \frac{1}{4}(u+c)^2F_1+ = \frac{\rho}{4\sqrt{3T}}(u + \sqrt{3T})^4,\end{aligned}$$

which coincide with the components of  $\mathbf{F}^+(\mathbf{U})$ . We have similar formulas for  $F_i-$ .

We know that the corresponding B.G.K. model is the limit of entropy satisfying models. Indeed, with the notation of Example 7.1, we have  $b = 3T/2$  and  $a = \rho/2\sqrt{3T}$ , and the function  $f_0(x, v) = a1_{\{|v-u|^2 \leq 2bT\}}$  satisfies the constraints (7.26) since

$$\frac{\rho}{2\sqrt{3T}} \int_{\mathbb{R}} 1_{\{|v|^2 \leq 3T\}} \left(1, \frac{|v|^2}{2}\right)^T dv = \left(\rho, \frac{\rho T}{2}\right)^T = (\rho, \rho\varepsilon)^T$$

(see Remark 7.3; the case  $\delta = 0, \gamma = 3$ , gives  $T = 2\varepsilon$ ).

In the general case ( $\gamma$  and  $\delta \neq 0$  satisfying (7.18b)), we would like the functions  $f_0, g_0$ , associated with this  $\chi$  by (7.39), to be the integrals w.r.t.  $\theta$  of an "equilibrium" function  $N$ , i.e.,

$$f_0(x, v) = \frac{\rho}{\sqrt{T}} \chi \left( \frac{v-u}{\sqrt{T}} \right) = \int_{\mathbb{R}^+} N(v-u, \theta) n(\theta) d\theta$$

and

$$g_0(x, v) = \delta T f_0(x, v) = \delta \rho \sqrt{T} \chi \left( \frac{v-u}{\sqrt{T}} \right) = \int_{\mathbb{R}^+} N(v, \theta) \theta n(\theta) d\theta$$

(see also Remark 7.5). For instance, the above identities hold with the function

$$N(v, \theta) = \frac{\rho}{2\theta_0\sqrt{3T}} \mathbf{1}_{\{\frac{|v|^2}{2} \leq \frac{3T}{2}\}} \exp\left(-\frac{\theta}{T}\right)$$

(we can also replace  $\frac{1}{\theta_0} \exp(-\frac{\theta}{T})$  by a characteristic function  $c\mathbf{1}_{\{\theta \leq d\}}$  such that  $\int_{\mathbb{R}^+} \mathbf{1}_{\{\theta \leq d\}} n(\theta) d\theta = 1$ ,  $\int_{\mathbb{R}^+} \mathbf{1}_{\{\theta \leq d\}} \theta n(\theta) d\theta = \delta T$ ). However, it is not clear that an entropy inequality exists in that case, and this motivates another choice for  $\chi$  which we shall now present.  $\square$

### 7.3.3 A Maximum Principle on the Entropy

In order to satisfy the entropy inequality and a maximum principle on an entropy function, we are led to modify the above technique and for  $f$  and  $g$  deal with two functions  $\chi$  and  $\zeta$  rather than with two kinetic variables  $v$  and  $\theta$ . More precisely, one chooses ( $\delta$  is defined by (7.8) with  $d = 1$ )

$$\chi(w) = a\left(1 - \frac{w^2}{b}\right)_+^\delta, \quad (7.47a)$$

$$\zeta(w) = \tilde{a}\left(1 - \frac{w^2}{b}\right)_+^{\delta+1} = \tilde{b}\left(\chi(w)\right)^{\frac{(\gamma+1)}{(3-\gamma)}}, \quad (7.47b)$$

and set

$$f(v) = f(v; \rho, u, T) = \frac{\rho}{\sqrt{T}} \chi\left(\frac{v-u}{\sqrt{T}}\right), \quad (7.48a)$$

$$g(v) = g(v; \rho, u, T) = \rho\sqrt{T} \zeta\left(\frac{(v-u)}{\sqrt{T}}\right), \quad (7.48b)$$

where the constants  $a, b$  are given by the constraints (7.38) and  $\tilde{a}, \tilde{b}$  by

$$\int_{\mathbb{R}} \zeta(w) dw = \delta.$$

Thus

$$\int_{\mathbb{R}} f(1, v)^T dv = (\rho, 0) \quad (7.49)$$

and

$$\int_{\mathbb{R}} g dv = \rho T \delta,$$

so that, as previously,

$$\int_{\mathbb{R}} \left\{ \frac{|v|^2}{2} f + g \right\} dv = \rho \frac{(T+u^2)}{2} = \rho T \delta = \frac{\rho T}{(\gamma-1)} + \frac{\rho u^2}{2} = \rho e.$$

The associated numerical split flux is naturally

$$\left\{ \begin{array}{l} \mathbf{F}^\pm(\mathbf{U}) = \rho \int_{\mathbb{R}} v_\pm \left\{ \left( 1, v, \frac{|v|^2}{2} \right)^T \chi \left( \frac{v-u}{\sqrt{T}} \right) \right. \\ \quad \left. + (0, 0, T)^T \zeta \left( \frac{(v-u)}{\sqrt{T}} \right) \right\} \frac{dv}{\sqrt{T}} \end{array} \right. \quad (7.50a)$$

or

$$\mathbf{F}^\pm(\mathbf{U}) = \int_{\mathbb{R}} v_\pm \left( f, vf, \frac{|v|^2}{2} f + g \right)^T dv \quad (7.50b)$$

which is the analog of (7.45).

We choose for an entropy function (in the sense of Lax)

$$S = \rho T^{\frac{-1}{(\gamma-1)}}; \quad (7.51)$$

since  $p = \rho T$  and  $\log \frac{p}{\rho^\gamma} = \log \left( \frac{T}{\rho^{(\gamma-1)}} \right)$ , we check easily that  $S$  is indeed of the form  $-\varphi(\log(\frac{p}{\rho^\gamma}))$  where  $\varphi$  is a nonpositive, nondecreasing concave function, which implies (see Harten [590] and Tadmor [1090]) that  $U = \rho S$  satisfies  $\frac{\partial(\rho S)}{\partial t} + \frac{\partial(\rho u S)}{\partial x} \leq 0$ . Tadmor has proven a maximum principle for the entropy

$$S(x, t + \Delta t) \leq \max \{ S(y, t); |y - x| \leq \Delta t \|u\|_{\mathbf{L}^\infty} \}$$

and the analogous property for numerical schemes such as Lax–Friedrich’s and Godunov’s. Here, we can derive a discrete entropy inequality and a maximum principle on the above entropy  $S$  for the kinetic scheme (Khobalatte and Perthame [695]).

*Theorem 7.3*

Consider the difference scheme (7.44ab)

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \lambda \{ \mathbf{F}(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n) - \mathbf{F}(\mathbf{U}_{j-1}^n, \mathbf{U}_j^n) \}$$

with

$$\mathbf{F}(\mathbf{U}, \mathbf{V}) = \mathbf{F}^+(\mathbf{U}) + \mathbf{F}^-(\mathbf{V}),$$

where  $\mathbf{F}^\pm$  are defined by (7.50). Under the CFL condition  $\lambda \max(|u_i^n| + \sqrt{T_i^n/b}) \leq 1$ , the scheme satisfies the property (7.46),

$$\rho_j^0 \geq 0, T_j^0 \geq 0, \forall j \Rightarrow \rho_j^n \geq 0, T_j^n \geq 0, \quad \forall j, \forall n > 0,$$

together with the entropy inequality

$$(\rho \pi_v)_i^{n+1} - (\rho \pi_v)_i^n + \lambda \{ G_{i+\frac{1}{2}}^n - G_{i-\frac{1}{2}}^n \} \leq 0, \quad \forall v > 0, \quad (7.52)$$

and a maximum principle on the entropy (7.51)

$$S_i^{n+1} = \frac{\rho_i^{n+1}}{(T_i^{n+1})^{\frac{1}{(\gamma-1)}}} \leq \max(S_{i-1}^n, S_i^n, S_{i+1}^n). \quad (7.53)$$

In (7.52),  $\pi_v$  is a one-parameter family of degenerated entropy functions,

$$\pi_v = \begin{cases} 0 & \text{if } \frac{\rho^{\gamma-1}}{T} < v, \\ 1 & \text{if } \frac{\rho^{\gamma-1}}{T} = v, \\ +\infty & \text{otherwise.} \end{cases} \quad (7.54)$$

Note that  $\rho\pi_v$  is obtained as the limit as  $p$  tends to  $+\infty$  of the convex entropies

$$\rho\left(\frac{S^{\gamma-1}}{v}\right)^p = \rho\left(\frac{\rho^{\gamma-1}}{(Tv)}\right)^p.$$

The numerical entropy flux  $G$  associated with  $\rho\pi_v$ ,

$$G_{i+\frac{1}{2}}^n = G(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n),$$

can also be split

$$G(\mathbf{U}, \mathbf{V}) = G^+(\mathbf{U}) + G^-(\mathbf{V}),$$

with  $G^\pm$  such that

$$G^\pm(\mathbf{U}) = F_\rho^\pm(\mathbf{U})\pi_v(\mathbf{U}),$$

where  $F_\rho^\pm(\mathbf{U})$  is the numerical mass flux (first component of  $\mathbf{F}^\pm(\mathbf{U})$ ). Thus,  $G$  is consistent:

$$G(\mathbf{U}, \mathbf{U}) = \rho u \pi_v.$$

*Proof of Theorem 7.3.* Let us first set

$$f_i^n(v) = f(v; \rho_i^n, u_i^n, T_i^n), \quad g_i^n(v) = g(v; \rho_i^n, u_i^n, T_i^n)$$

and

$$\begin{aligned} \bar{f}_i(v) &= f_i^n(v) - \lambda\{v_+ f_i^n(v) - v_- f_{i+1}^n(v) - v_+ f_{i-1}^n(v) + v_- f_i^n(v)\} & (7.55) \\ &= f_i^n(v) - \lambda\{v_+(f_i^n(v) - f_{i-1}^n(v)) - v_-(f_{i+1}^n(v) - f_i^n(v))\}, \\ \bar{g}_i(v) &= g_i^n(v) - \lambda\{v_+ g_i^n(v) - v_- g_{i+1}^n(v) - v_+ g_{i-1}^n(v) + v_- g_i^n(v)\}, \end{aligned}$$

so that

$$\begin{aligned} \mathbf{U}_i^{n+1} &= \int_{\mathbb{R}} \left( \bar{f}_i(v), v \bar{f}_i(v), \frac{|v|^2}{2} \bar{f}_i(v) + \bar{g}_i(v) \right)^T dv, \\ \mathbf{F}^\pm(\mathbf{U}_i^n) &= \pm \int_{\mathbb{R}} v_\pm \left( f_i^n, v f_i^n(v), \frac{|v|^2}{2} f_i^n + g_i^n \right)^T dv. \end{aligned}$$

The first property (7.46) follows as in Theorem 7.2.

Let us sketch the proof of the inequality (7.53). We consider the function

$$h = h(f, g) = (f^{\gamma+1} g^{\gamma-3})^{\frac{1}{(\gamma-1)}}. \quad (7.56)$$

The function  $h$  is a convex function of  $f$  and  $g$ , and the exponents have been chosen in order that the power of  $(1 - w^2/b)_+$  vanishes. Thus, we get

$$h_i^n = (S_i^n)^2 \tilde{b}^{\frac{(\gamma-3)}{(\gamma-1)}} 1_{\{|v-u_i^n|^2 \leq bT\}} = c(S_i^n)^2 1_{\{|v-u_i^n|^2 \leq bT\}},$$

where  $c = \tilde{b}^{\frac{(\gamma-3)}{(\gamma-1)}}$ . We obtain by (7.55)

$$\bar{h}_i = h(\bar{f}_i, \bar{g}_i) \leq h_{i-1}^n(\lambda v_+) + h_i^n(1 - \lambda v_+ - \lambda v_-) + h_{i+1}^n(\lambda v_-), \quad (7.57)$$

which yields

$$\bar{h}_i \leq c \max\{S_{i-1}^n, S_i^n, S_{i+1}^n\}^2. \quad (7.58)$$

We then consider the problem of minimization,

$$j(\rho, \Sigma) = \min \int_{\mathbb{R}} \left\{ \frac{|v|^2}{2} f + g \right\} dv, \quad (7.59)$$

where the minimum is taken over all functions  $f \geq 0, g \geq 0$  satisfying the constraints (7.49) and such that  $h(f, g) \leq \Sigma$  (for given  $\Sigma$  and  $\delta$ ), and set  $j(\rho, \Sigma) = \frac{\rho\tau}{(\gamma-1)}$ , which defines  $\tau$  in terms of  $j$ .

### Lemma 7.5

*Problem (7.59) admits the following solution: the minimum of  $j$  is obtained for the functions  $f(v; \rho, 0, \tau) = (\frac{\rho}{\sqrt{\tau}})\chi(\frac{v}{\sqrt{\tau}})$ ,  $g(v; \rho, 0, \tau) = \rho\sqrt{\tau}\zeta(\frac{v}{\sqrt{\tau}})$  where  $\chi$  are  $\zeta$  defined by (7.47) and  $j(\rho, \Sigma) = \frac{\rho}{\gamma-1}(\frac{\rho}{\sqrt{\Sigma/c}})^{\gamma-1}$ .*

We apply the results of Lemma 7.5 with

$$\Sigma = c \ max(S_{i-1}^n, S_i^n, S_{i+1}^n)^2.$$

Since by (7.58),  $\bar{h}_i = h(\bar{f}_i, \bar{g}_i) \leq \Sigma$ ,  $\bar{f}_i(v + u_i^n)$ , and  $\bar{g}_i(v + u_i^n)$  satisfy the constraints, we can write

$$\begin{aligned} \frac{\rho_i^{n+1} T_i^{n+1}}{(\gamma-1)} &= \int_{\mathbb{R}} \left\{ \frac{|v|^2}{2} \bar{f}_i(v + u_i^{n+1}) + \bar{g}_i(v + u_i^{n+1}) \right\} dv \\ &\geq j(\rho_i^{n+1}, \Sigma) = \frac{\rho_i^{n+1}}{(\gamma-1)} \left( \frac{\rho_i^{n+1}}{\sqrt{\Sigma/c}} \right)^{\gamma-1}, \end{aligned}$$

and we get by (7.51)

$$S_i^{n+1} \leq \left( \frac{\Sigma}{c} \right)^{\frac{1}{2}} = \max(S_{i-1}^n, S_i^n, S_{i+1}^n)$$

as expected.

The proof of (7.52) relies on a similar argument: one considers the function

$$k(f, g) = f \left( \frac{h(f, g)}{v^2} \right)^p = f \left( \frac{(f^{\gamma+1} g^{\gamma-3})^{\frac{1}{(\gamma-1)}}}{v^2} \right)^p,$$

for which an inequality of the form (7.57) is derived, and uses another minimization problem (analogous to (7.25)),

$$H(\rho, T) = \min \left\{ \int_{\mathbb{R}} f(h(f, g))^p dv \right\},$$

where the minimum is taken over all  $f, g \geq 0$  satisfying the constraints (7.49) and

$$\int_{\mathbb{R}} \left\{ f(v) \frac{|v|^2}{2} + g(v) \right\} dv = \frac{\rho T}{(\gamma - 1)}.$$

The minimum is obtained for some functions  $f_p$  and  $g_p$ ,

$$f_p = \frac{\rho a_p}{\sqrt{T}} \left( 1 - \frac{w^2}{b_p T} \right)_+^{\delta + \frac{1}{2p}},$$

$$g_p = c_p T F_p \left( 1 - \frac{w^2}{b_p T} \right)_+ = c_p \rho a_p \sqrt{T} \left( 1 - \frac{w^2}{b_p T} \right)_+^{\delta + 1 + \frac{1}{2p}},$$

where the constants  $a_p, b_p, c_p$  are determined by the constraints. Then

$$\int_{\mathbb{R}} k(f_p, g_p) dv \leq \int_{\mathbb{R}} \bar{k}_i dv \leq \int_{\mathbb{R}} \{ \lambda v_+ h_{i-1}^n + (1 - \lambda v_+ - \lambda v_-) h_i^n + \lambda v_- h_{i+1}^n \} dv,$$

and (7.52) follows by letting  $p$  tend to  $+\infty$ . We skip the details of the proof.  $\square$

*Remark 7.6.* We are able to prove an entropy inequality for these functions  $f_p$  and  $g_p$  with entropy functions that are not degenerate, but no maximum principle on the entropy; this is obtained for the limit functions as  $p$  goes to  $+\infty$ . This should be compared with Example 7.2. This remark and the results of this section are taken from Khobalatte–Perthame [695].  $\square$

*Proof of Lemma 7.5.* We just sketch the proof, which follows from classical arguments of the calculus of variations after writing the Euler–Lagrange equations of the minimization problem with constraints. The minimum is achieved for  $f(v; \rho, 0, \tau) = (\frac{\rho}{\sqrt{\tau}}) \chi(\frac{v}{\sqrt{\tau}})$ ,  $g(v; \rho, 0, \tau) = (\rho \sqrt{\tau}) \zeta(\frac{v}{\sqrt{\tau}})$ , defined by (7.47). Let us just note that, as we have already observed, the exponents in formulas (7.47) are such that

$$\delta(\gamma + 1) + (\gamma - 3)(\delta + 1) = 0,$$

which implies that

$$h(f, g) = (f^{\gamma+1}g^{\gamma-3})^{\frac{1}{(\gamma-1)}} = c \left( \frac{\rho}{\tau^{\frac{1}{(\gamma-1)}}} \right)^2 1_{\{|v|^2 \leq b\tau\}}.$$

The constraint inequality becomes an equality for the functions  $f$  and  $g$ , and we obtain the identity

$$c \left( \frac{\rho}{\tau^{\frac{1}{(\gamma-1)}}} \right)^2 = \Sigma, \text{ or } \tau = \left( \frac{\rho}{(\frac{\Sigma}{c})^{\frac{1}{2}}} \right)^{(\gamma-1)},$$

which in turn gives  $j$ . □

### 7.3.4 Second-Order Accurate Schemes

The extension of these schemes to obtain second-order accuracy can be done in two ways. The first uses a Chapman–Enskog-type analysis for the time discretization and takes initial conditions  $f_1, g_1$ , whose Taylor expansions are higher-order approximations of the exact solutions (see the proof of Lemma 7.3) (we refer to Deshpande [411, 845] and Perthame [941, 942]). Then, following the MUSCL (monotonic upstream-centered scheme for conservation laws) approach of van Leer, one takes piecewise linear functions for  $\mathbf{U}(x, 0) = \mathbf{U}_0(x)$ .

The other method is simpler and can be extended easily to two-dimensional systems: one divides the cell  $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$  into two parts and takes functions that are constant on a half cell:  $\mathbf{U}_i^-$  on  $(x_{j-\frac{1}{2}}, x_j)$  and  $\mathbf{U}_i^+$  on  $(x_j, x_{j+\frac{1}{2}})$ . In one dimension, it is equivalent to introducing slopes; we can develop a MUSCL-type scheme, and it is easy to prove second-order accuracy. Indeed, we define

$$\begin{cases} \rho_i^\pm = \rho_i \pm D\rho_i, \\ u_i^\pm = \bar{u}_i \pm Du_i, \\ s_i^\pm = \bar{s}_i \pm Ds_i, \end{cases} \quad (7.60)$$

where  $\bar{u}_i, \bar{s}_i$  are second order close to  $u_i, s_i$  and will be chosen to guarantee conservation of momentum and energy; the increments or (loosely speaking) the slopes need to be limited. The limited variables are ( $\rho, u, s = \frac{\rho^{\gamma-1}}{T} = \frac{\rho^\gamma}{p}$ );  $u$  and  $s$  are not the conservative variables but are chosen for the limitation procedure with the aim of getting a maximum principle on the entropy. The slopes are obtained starting from a centered prediction that is slightly corrected, for instance,

$$\begin{cases} D\rho_i = \operatorname{sgn}(\rho_{i+1} - \rho_{i-1}) \min\left(\frac{|\rho_{i+1} - \rho_{i-1}|}{4}, \rho_i\right), \\ Du_i = \operatorname{sgn}(u_{i+1} - u_{i-1}) \min\left(\frac{|u_{i+1} - u_{i-1}|}{4}, \sqrt{T_i/\gamma - 1}\right), \\ Ds_i = \operatorname{sgn}(s_{i+1} - s_{i-1}) \min\left(\frac{|s_{i+1} - s_{i-1}|}{4}, \frac{s_i}{4}\right). \end{cases} \quad (7.61)$$

The values  $\bar{u}_i$  and  $\bar{s}_i$  are second-order modifications of  $u_i, s_i$  computed in order that  $\mathbf{U}_i^\pm = (\rho_i^\pm, \rho_i^\pm u_i^\pm, \rho_i^\pm e_i^\pm = \rho_i^\pm \frac{u_i^{\pm 2}}{2} + \frac{\rho_i^{\pm \gamma}}{s_i^{\pm}(\gamma-1)})$  satisfy

$$\mathbf{U}_i = \frac{1}{2}(\mathbf{U}_i^+ + \mathbf{U}_i^-).$$

Thus

$$\begin{cases} \rho_i^+ u_i^+ + \rho_i^- u_i^- = 2\rho_i u_i, \\ \frac{\rho_i^+ u_i^{+2}}{2} + \frac{\rho_i^- u_i^{-2}}{2} + \frac{(\rho_i^+)^{\gamma}}{s_i^+} + \frac{(\rho_i^-)^{\gamma}}{s_i^-} = 2E_i. \end{cases} \quad (7.62)$$

Substituting the expressions (7.60) for  $\rho_i^\pm, u_i^\pm$ , the first Eq. (7.62) yields

$$\bar{u}_j = u_j - Du_j \frac{D\rho_j}{\rho_j},$$

and the second equation gives in turn  $\bar{s}_i$  as the positive root of a quadratic polynomial.

Thus, starting from  $\mathbf{U}_i = (\rho_i, \rho_i u_i, \rho_i e_i)$ , we have defined  $\mathbf{U}_i^\pm$ , which we use to derive the numerical flux. In fact, if we use a particular (second-order) Runge-Kutta two-stage scheme in time, one can prove that the resulting scheme is second order in space and time and preserves the positivity of  $\rho$  and  $T$ . Thus, we consider the scheme

$$\begin{cases} \mathbf{U}_i^{n,1} = \mathbf{U}_i^n - \lambda \{\mathbf{F}_{i+\frac{1}{2}}^n - \mathbf{F}_{i-\frac{1}{2}}^n\}, \\ \mathbf{U}_i^{n,2} = \mathbf{U}_i^{n,1} - \lambda \{\mathbf{F}_{i+\frac{1}{2}}^{n,1} - \mathbf{F}_{i-\frac{1}{2}}^{n,1}\}, \\ \mathbf{U}_i^{n+1} = \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_i^{n,2}), \end{cases}$$

where we have set

$$\begin{aligned} \mathbf{F}_{i+\frac{1}{2}} &= \mathbf{F}^+(\mathbf{U}_i^+) + \mathbf{F}^-(\mathbf{U}_{i+1}^-), \\ \mathbf{F}_{i+\frac{1}{2}}^1 &= \mathbf{F}^+(\mathbf{U}_i^{1+}) + \mathbf{F}^-(\mathbf{U}_{i+1}^{1-}), \end{aligned}$$

and  $\mathbf{F}^\pm$  are defined in (7.50) and  $\mathbf{U}_i^{1\pm}$  are computed from  $\mathbf{U}_i^1$  through the formulas (7.60)–(7.62). With a further limitation on the only slope  $\Delta s$ ,

$$|\Delta s_i| \leq \max\{s_{i-1}, s_i, s_{i+1}\} - s_i,$$

(and no limiter on  $\rho$  and  $u$ ), one obtains, moreover, a maximum principle on  $s_{i\pm\frac{1}{2}}$  up to second-order terms, which yields a very satisfactory damping of the oscillations (see Khobalatte and Perthame [695] and Perthame and Qiu [946]).

*Remark 7.7.* A first drawback concerning the use of kinetic schemes, which was already mentioned in Sect. 5, Remark 5.1, for all flux splitting schemes, is the poor resolution of contact discontinuities and slip surfaces that are heavily smeared. Thus, the application to viscous external flows, where a precise capture of boundary layers is important, requires some modifications (see De Vuyst [397]).

Another limitation seems to be the lack of generality since, as we have presented them, they rely heavily on the underlying kinetic theory. However, they have been extended to more general equations of states (Coron and Perthame [363]), nonequilibrium flow (De Vuyst [396], Villedieu (1994)), two-phase (Allaire and Zelmanse [28]), Coquel et al. [345]), MHD (Khanfir [376]), and the shallow water equations, which we will mention again below.  $\square$

## 7.4 Some Extensions of the Kinetic Approach

The kinetic formalism which links the Boltzmann equation and the Euler system has been extended to scalar equations and to other systems of equations. We only give the main lines of a powerful theory which has interesting outcomes in the derivation of numerical schemes; we refer to B. Perthame's textbook [944] for a thorough treatment of the subject. In particular, efficient kinetic schemes have been derived recently for the shallow water equations [64, 949].

### 7.4.1 Kinetic Representation

Starting from a macroscopic description, the idea of the kinetic approach is to associate a kinetic description, for theoretical or numerical purposes, mimicking the link between the Euler system and the Boltzmann equation. For scalar nonlinear conservation laws,

$$\frac{\partial u}{\partial t} + \sum_{i=1}^d \frac{\partial}{\partial x_i} F_i(u) = 0,$$

where in this section, we use the notation  $F_i$  for the flux, in order to distinguish it from the density  $f$ , there exists a complete theory due to Lions, Perthame, and Tadmor [812], with a *kinetic formulation*, which means that

it provides an equivalent description of the conservation law. It allows to represent  $u$  as the moment of some “equilibrium” density  $\chi$  and more generally any function of  $u$  as

$$S(u) - S(0) = \int_{\mathbb{R}} S'(\xi) \chi(\xi; u) d\xi,$$

where  $\chi$  is defined by

$$\chi(\xi; u) = \begin{cases} +1, & 0 < \xi < u, \\ -1, & u < \xi < 0, \\ 0, & \text{otherwise.} \end{cases} \quad (7.63)$$

We associate with  $\chi(\xi; u(\mathbf{x}, t))$  a *linear* transport equation

$$\frac{\partial}{\partial t} \chi(\xi; u(t, \mathbf{x})) + \mathbf{a}(\xi) \cdot \nabla_{\mathbf{x}} \chi(\xi; u(t, \mathbf{x})) = \frac{\partial}{\partial \xi} m(t, \mathbf{x}, \xi), \quad (7.64)$$

where  $\mathbf{a} = (F'_1, \dots, F'_d)$  and  $m$  is some nonnegative bounded measure;  $m$  vanishes when  $u$  is smooth and is linked somehow to the entropy dissipation otherwise. The equivalence is proved in details in the textbook of Perthame [944], Chapter 3 ([950] for a pioneering paper).

*Remark 7.8.* Let us consider the transport equation with a BGK-like kernel

$$\partial_t f_\varepsilon + \mathbf{a}(\xi) \cdot \nabla_{\mathbf{x}} f_\varepsilon = \frac{1}{\varepsilon} (\chi(\xi; u_\varepsilon(\mathbf{x}, t)) - f_\varepsilon)$$

where  $u_\varepsilon(\mathbf{x}, t) = \int_{\mathbb{R}} f_\varepsilon(t, \mathbf{x}, \xi) d\xi$ , for a well-chosen initial condition  $f^0(\mathbf{x}, \xi)$  (with  $u_0(\mathbf{x}) = \int_{\mathbb{R}} f^0(\mathbf{x}, \xi) d\xi$ ). It is possible to characterize rigorously the corresponding limits,  $f_\varepsilon \rightarrow \chi$  as  $\varepsilon \rightarrow 0$ , and the equilibrium is indeed  $\chi$  given by (7.63), and  $u_\varepsilon \rightarrow u$  where  $u$  is a solution to the kinetic formulation (7.64) so that the limit of the right-hand side takes the form of a measure  $m$  (the *kinetic entropy defect measure*). The *kinetic formulation* is the limit of this kinetic approximation equation of the conservation law. This *formulation* provides a very powerful tool to obtain theoretical results (existence, uniqueness, regularizing effects, etc.); we refer to the textbook of B. Perthame [944].

□

For systems to have a *kinetic formulation*, for which the kinetic and macroscopic formulations are equivalent, supposes that they admit a sufficient number of entropies. In particular, for the system of elasticity in 1d, there exists a kinetic formulation and, for the isentropic gas dynamics system, a quite complete (semi-kinetic) theory [944]. There are, however, few examples of such systems for which rigorous proofs can be obtained. Then, there exists another level, the *kinetic representation* (a kinetic representation is weaker than a kinetic formulation), which is used for numerical purposes, for which

the BGK model above can serve as an introduction. Here we just sketch the great lines for the isentropic gas dynamics system in 1D. We assume  $p = \kappa \varrho^\gamma$ , with  $1 < \gamma \leq 3$ . We introduce the function

$$\mathcal{M}(\varrho, \xi) = \varrho^{\frac{3-\gamma}{2}} \chi\left(\frac{\xi}{\varrho^{\frac{\gamma-1}{2}}}\right),$$

where we use the same notation  $\chi$ , though the function differs from (7.63); it is now an even, nonnegative function, satisfying

$$\int_{\mathbb{R}} \chi(s) ds = 1, \quad \int_{\mathbb{R}} s^2 \chi(s) ds = \kappa,$$

(as in (7.38), but for the normalization factor  $\kappa$  instead of 1), and we introduce the moments of  $\mathcal{M}$  (but for 1/2 in the third component so as to get directly the momentum flux which involves twice the kinetic energy)

$$(\varrho, \varrho u, \varrho u^2 + \kappa \varrho^\gamma)^T = \int_{\mathbb{R}} \mathcal{M}(\varrho, u - \xi)(1, \xi, \xi^2)^T d\xi. \quad (7.65)$$

Let us define  $Q$  by

$$\partial_t \mathcal{M}(\varrho, \xi - u) + \xi \partial_x \mathcal{M}(\varrho, \xi - u) = Q(t, x, \xi). \quad (7.66)$$

One can see that if the moments solve the isentropic gas dynamics system,  $Q$  satisfies  $\int_{\mathbb{R}} Q d\xi = 0$ ,  $\int_{\mathbb{R}} \xi Q d\xi = 0$ , but  $Q$  does not vanish, contrary to the Boltzmann kernel which vanishes for Maxwellians.

Then the kinetic equation, which is the “analogous” of the Boltzmann equation, writes with  $f = f(t, x, \xi)$  representing a density of particles and a BGK-type kernel for the collision term

$$\partial_t f_\varepsilon + \xi \partial_x f_\varepsilon = \frac{1}{\varepsilon} (\mathcal{M}(\varrho, u - \xi) - f_\varepsilon), \quad (7.67)$$

where  $\varrho, \varrho u$  (depending on  $\varepsilon$ ) are the moments of  $f_\varepsilon(t, x, \xi)$ . Formally (this is still an open problem to give a rigorous proof) as  $\varepsilon \rightarrow 0$ ,  $f_\varepsilon \rightarrow \mathcal{M}$ , the *equilibrium* or Maxwellian, and  $\frac{1}{\varepsilon} (\mathcal{M}(\varrho, u - \xi) - f_\varepsilon) \rightarrow Q$  ( $Q$  defined in (7.66)), and the limit moments, given by (7.65), satisfy the isentropic gas dynamics system.

A minimum entropy principle holds with a particular choice of function  $\chi$ , i.e., the corresponding Maxwellian, say  $\mathcal{M}_E(\varrho, u - \xi)$ , minimizes the microscopic energy. The function  $\chi$  is defined by

$$\chi(s) = \alpha \left(1 - \frac{s^2}{\beta}\right)_+^\lambda$$

with  $\lambda = \frac{3-\gamma}{2(\gamma-1)}$ , where  $\alpha, \beta$  are constants such that the constraints defining the moments of  $\chi$  are satisfied.

Let us detail the result for  $\gamma = 2$  which corresponds to the shallow water equations since  $p(h) = gh^2/2$  and  $\kappa = g/2$ ; then

$$\mathcal{M}(h, \xi) = \sqrt{h}\chi\left(\frac{\xi}{\sqrt{h}}\right),$$

and the moments are given by  $h = \int_{\mathbb{R}} \mathcal{M}(h, \xi - u) d\xi$ ,  $hu = \int_{\mathbb{R}} \xi \mathcal{M}(h, \xi - u) d\xi$ , and  $hu^2 + gh^2/2 = \int_{\mathbb{R}} \xi^2 \mathcal{M}(h, \xi - u) d\xi$ . To take into account the bottom, one needs to add a term  $-gZ' \partial_{\xi} f$  to Eq. (7.66). Then a pair of functions  $(h, hu)$  is a strong solution of the shallow water system (see (3.2) in the Chap. I) if and only if  $\mathcal{M}(h, \xi - u)$  solves

$$\partial_t \mathcal{M} + \xi \partial_x \mathcal{M} - gZ' \partial_{\xi} \mathcal{M} = Q(t, x, \xi),$$

for some  $Q$  satisfying  $\int_{\mathbb{R}} Q d\xi = 0$ ,  $\int_{\mathbb{R}} \xi Q d\xi = 0$ . The minimum energy principle is obtained for

$$\chi(s) = \frac{\sqrt{2}}{\pi \sqrt{g}} \left(1 - \frac{s^2}{2g}\right)_+^{1/2}$$

where the energy defined by

$$\mathcal{E}(f) = \int_{\mathbb{R}} \left( \frac{\xi^2}{2} f(\xi) + \frac{\pi^2 g^2}{6} f^3(\xi) + gZ f(\xi) \right) d\xi$$

is minimized over the set of functions  $f \geq 0$  with moments  $(h, hu)$ . Then at the minimum  $f = \mathcal{M}(h, \xi - u)$  (*Gibbs equilibrium*), the value of  $\mathcal{E}$  is

$$\mathcal{E}(\mathcal{M}(h, \xi - u)) = hu^2/2 + gh^2/2 + ghZ$$

which is the mathematical entropy of the Saint-Venant system.

Based on this kinetic representation, we can construct numerical schemes by using the upwind scheme for the transport equation on the density  $f$ , which from  $f^n = \mathcal{M}(h^n, \xi - u^n)$  updates  $f^n \rightarrow f^{n+1}$ , and this evolution step is followed by instantaneous relaxation toward equilibrium, defining  $\mathbf{U}_i^{n+1}$  by the moments of  $f_i^{n+1}$ . Details for the Saint-Venant system can be found in [949], and more references will be given in the last chapter concerning source terms.

#### 7.4.2 Discrete Kinetic Approximations

Discrete kinetic approximations of a system of  $p$  conservation laws

$$\partial_t \mathbf{u} + \sum_j \partial_{x_j} \mathbf{F}_j(\mathbf{u}) = \mathbf{0}$$

are derived from the above kinetic approach, replacing the infinite family  $f(\xi), \xi \in \mathbb{R}$  by a finite one  $f = (f_i), i \in \{1, 2, \dots, N\}$  and assuming a finite number of transport velocities. This leads to a finite dimensional semilinear system with a BGK-type kernel

$$\partial_t f + \sum_{j=1}^d \Lambda_j \partial_{x_j} f = \frac{1}{\varepsilon} (\mathcal{M}(Pf) - f),$$

where  $f(\mathbf{x}, t) \in \mathbb{R}^N$ ,  $\Lambda_j$  are constant  $N \times N$  diagonal matrices,  $\mathcal{M}(\cdot) \in \mathbb{R}^N$  is a “Maxwellian distribution” and  $P$  some operator. We assume, as for (7.13), that  $\mathcal{M}(\mathbf{u})$  may be characterized by its “moments,” noted  $\mathbf{u} \in \mathbb{R}^p$ . These assumptions correspond to the introduction of a linear operator  $P : \mathbb{R}^N \rightarrow \mathbb{R}^p$ , i.e., a constant  $p \times N$  matrix so that the following identities hold:

$$P\mathcal{M}(\mathbf{u}) = \mathbf{u}, \quad P\Lambda_j \mathcal{M}(\mathbf{u}) = \mathbf{F}_j(\mathbf{u}),$$

where the  $\mathbf{F}_j$ ’s denote the fluxes of the system. Let us consider, for example, the 1D case ( $d = 1$ ) and a scalar conservation law ( $p = 1$ )

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

One can approximate the equation by the finite dimensional system

$$\partial_t f + \Lambda \partial_x f = \frac{1}{\varepsilon} (\mathcal{M}(u) - f), \quad (7.68)$$

where  $f = (f_1, \dots, f_N)^T$ ,  $\mathcal{M} = (\mathcal{M}_1, \dots, \mathcal{M}_N)^T$ , and

–  $\Lambda$  is a constant diagonal  $N \times N$  matrix  $\Lambda = \text{diag}(\lambda_i)$ .

– The projection operator  $P$ , which corresponds to integration ( $f \mapsto \int f d\xi$ ) for the full kinetic representation, is a linear form on  $\mathbb{R}^N$ , thus characterized by the row vector with only one entry ( $P = (1, \dots, 1)$ ), which gives a discrete sum  $Pf = \sum_{i=1}^N f_i$ .

–  $\mathcal{M}$ , the local Maxwellian, satisfies the consistency relations (analogous to discrete moments)

$$P\mathcal{M}(u) \equiv \sum_{i=1}^N \mathcal{M}_i(u) = u, \quad P\Lambda \mathcal{M}(u) \equiv \sum_{i=1}^N \lambda_i \mathcal{M}_i(u) = F(u),$$

for all  $u$  in the set of states. If we emphasize the dependence on  $\varepsilon$  and set  $f \equiv f_\varepsilon$  and define the moments of  $f_\varepsilon$ ,  $u_\varepsilon = Pf_\varepsilon = \sum_{i=1}^N f_{\varepsilon i}$ , and  $v_\varepsilon = P\Lambda f_\varepsilon = \sum_{i=1}^N \lambda_i f_{\varepsilon i}$ , then, if  $f_\varepsilon$  is solution of (7.68), we deduce that  $u_\varepsilon, v_\varepsilon$  satisfy

$$\begin{cases} \partial_t u_\varepsilon + \partial_x v_\varepsilon = 0, \\ \partial_t v_\varepsilon + P\Lambda^2 \partial_x f_\varepsilon = \frac{1}{\varepsilon} (F(u_\varepsilon) - v_\varepsilon). \end{cases} \quad (7.69)$$

As  $\varepsilon \rightarrow 0$ , at least formally,  $f_\varepsilon$  tends to a Maxwellian  $\mathcal{M}(u)$ , with discrete moments  $u, F(u)$ ;  $u_\varepsilon = Pf^\varepsilon \rightarrow u$ ,  $v_\varepsilon \rightarrow F(u)$ , and  $u$  is solution of the conservation law  $\partial_t u + \partial_x F(u) = 0$ .

Let us give a model example (which corresponds to the Jin–Xin relaxation (which we will detail in Sect. 8.2.2), with  $N = 2$ ,  $f = (f_1, f_2)$ , and  $\Lambda = \text{diag}(-a, a)$ , for some constant  $a > 0$ ,  $P = (1, 1)$ , and for which the two components of the Maxwellian are given by

$$\mathcal{M}(u) = (\mathcal{M}_1(u), \mathcal{M}_2(u)) \equiv \left( \frac{1}{2}(u - F(u)/a), \frac{1}{2}(u + F(u)/a) \right).$$

Then from  $u = f_1 + f_2$  and  $v = -af_1 + f_2$ , we get

$$f_1 \equiv \frac{1}{2}\left(u - \frac{v}{a}\right) = -\frac{1}{2a}w, \quad f_2 \equiv \frac{1}{2}\left(u + \frac{v}{a}\right) = \frac{1}{2a}z,$$

and the discrete kinetic formulation is equivalent to

$$\begin{cases} \partial_t u_\varepsilon + \partial_x v_\varepsilon = 0 \\ \partial_t v_\varepsilon + a^2 \partial_x u_\varepsilon = \frac{1}{\varepsilon}(F(u_\varepsilon) - v_\varepsilon). \end{cases} \quad (7.70)$$

Note that  $f_1, f_2$  are the Riemann invariants of the homogeneous part of (7.70), and  $f_1$  is associated with  $\lambda_2 = a$  (resp.  $f_2$  to  $\lambda_1 = -a$ ). We have introduced the notations  $w, z$ , only to make a link with the next section where the same system will be encountered in the Jin–Xin relaxation scheme (see (8.13)). As will be explained in the next section, the constant  $a$  must be chosen in order that the limit propagation speed  $f'(u)$  is bounded by these two velocities  $-a < f'(u) < a$ ; this stability condition is called the sub-characteristic or Whitham's condition.

Note that this approach is possible for systems of  $p$  equations, and then  $f_i \in \mathbb{R}^p, i = 1, 2, u \in \mathbb{R}^p, N = 2p, \Lambda = \text{diag}(\Lambda_-, \Lambda_+), \Lambda_\pm p \times p$  diagonal matrix  $= \text{diag}(\pm a)$ , and so on. We refer to Serre [1041] for convergence results (and references therein).

Then a simple numerical scheme for the limit conservation equation consists in the upwind scheme for the homogeneous part of the linear diagonal kinetic system, followed by an instantaneous projection step on the equilibrium manifold. The scheme we get is written in terms of  $u = Pf$ ; it is quite simple and robust. This works in a similar way for a system, and the convergence to a solution of the equilibrium system is proved in [741].

We refer to [55, 162, 894] for more details and general discrete kinetic approximations.

The discrete kinetic approach is a direct introduction to relaxation approximation which we now consider.

## 8 Relaxation Schemes

This section introduces to a different approach used in the theoretical and numerical analysis of hyperbolic systems, the relaxation approach, which is inherited from physics too and which, from a mathematical point of view, has many connections with the kinetic formulation.

Many physical systems contain rapid relaxation processes, so that one is interested by studying the limit of the (fine) relaxation model as the relaxation time tends to zero; the (coarser) limit is called the *equilibrium* model (see some classical examples in [602, 835, 1218]).

The limit solution is supposed to lie in some equilibrium manifold. An equilibrium is attained for some state with minimum entropy, at least if the relaxation process is stable; this can be proved in some formal framework.

Now, the Euler system of compressible flows has been derived from the (formal) limit of the Boltzmann equation for dilute gases as the mean free path goes to zero; the mean free path can be considered as a relaxation parameter. The Boltzmann equation describes the “relaxation” toward a state of minimum for some entropy (see Sect. 7.1), and the equilibria are called Maxwellian distributions. Thus, in analogy, it is frequent to call also Maxwellians the equilibria in the relaxation context.

The relaxation schemes we will introduce in this section are derived by involving the relaxation formalism in their construction; they are eventually written in the general form of finite volume schemes. It is essential for a better understanding of their properties to have first a quick look at the underlying relaxation theory, as we did above for the kinetic formalism.

### 8.1 Introduction to Relaxation

#### 8.1.1 The General Relaxation System

In a pioneering work [291], Chen, Levermore, and Liu (1994) consider hyperbolic systems of conservation laws of  $n$  equations with a general relaxation term

$$\frac{\partial \mathbf{w}}{\partial t} + \frac{\partial}{\partial x} \mathbf{h}(\mathbf{w}) = \mu \mathcal{R}(\mathbf{w}), \quad t > 0, \quad (8.1)$$

where  $\mu > 0$  is a parameter,  $\tau = \frac{1}{\mu}$  is the “relaxation time,” the matrix  $\mathbf{h}'(\mathbf{w})$  is diagonalizable with real eigenvalues, and the set of states is  $\Omega \subset \mathbb{R}^n$ . For simplicity, we have restricted ourselves to the 1D case, but the formalism is defined quite naturally in any space dimension; we also take a somehow simplified source term without much loss of generality. Indeed, we assume that the relaxation term  $\mathcal{R}(\mathbf{w})$  concerns only the  $n - p$  last equations; the  $p$  first components of  $\mathcal{R}$  are 0

$$\mathcal{R}(\mathbf{w}) = (0, \dots, 0, \mathcal{R}_{p+1}, \dots, \mathcal{R}_n)^T.$$

We note  $\mathbf{u} = \mathbb{P}(\mathbf{w})$  the projection of  $\mathbb{R}^n$  onto  $\mathbb{R}^p$

$$\mathbb{P} : \mathbf{w} = (w_1, \dots, w_n)^T \mapsto \mathbf{u} = (w_1, \dots, w_p)^T.$$

Moreover, we assume that  $\mathcal{R}(\mathbf{w}) = \mathbf{0}$  defines an equilibrium manifold  $M$

$$\mathcal{R}(\mathbf{w}) = \mathbf{0} \Leftrightarrow \mathbf{w} \in M$$

and also that an equilibrium state, say  $\mathbf{w}_{eq}$ , is completely determined by its  $p$  first components  $\mathbf{u} = \mathbb{P}(\mathbf{w}_{eq})$ , and we note  $\mathbf{w}_{eq} = \mathcal{M}(\mathbf{u})$  which defines a mapping,  $\mathcal{M} : \mathbb{P}\Omega \rightarrow M$ , that may be written in the following way:

$$\mathcal{M}(\mathbf{u}) = (\mathbf{u}, \mathbf{e}(\mathbf{u}))$$

where  $\mathbf{v} = \mathbf{e}(\mathbf{u})$  has  $n - p$  components  $e_i(\mathbf{u})$ . We may, for instance, think of a simple case where  $\mathcal{R}$  has a BGK-type form  $\mathbf{e}(\mathbf{u}) - \mathbf{v}$ . Hence, for  $\mu \rightarrow \infty$ , the  $n - p$  last differential equations degenerate in “algebraic” identities  $v_i = e_i(\mathbf{u})$ ,  $i = p + 1, \dots, n$ , while the  $p$  first ones yield that  $\mathbf{u}$  is (at least formally) a solution of a system of  $p$  equations

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{0}, \quad (8.2)$$

where  $\mathbf{f}(\mathbf{u}) = \mathbb{P}(\mathbf{h}(\mathcal{M}(\mathbf{u})))$ .

An important issue is to prove the convergence (we do not precise in which sense, but the problem is nonlinear, and the proofs are not easy), and the convergence results need some stability conditions. Different stability criteria are exhibited which are sufficient conditions; some of them are easy to handle, if not easily satisfied, and have an immediate counterpart at the numerical level. The first one introduced in [291] is the existence of an *entropy extension*. Starting from system (8.2), we assume it is endowed with an entropy pair  $(\eta, q)$ , and then we define

*Definition 8.1*

An entropy extension of  $(\eta, q)$  for the relaxation system (8.1) is a convex function  $\mathcal{E}(\mathbf{w})$ , such that there exists an entropy flux  $\mathcal{G}(\mathbf{w})$  satisfying  $\mathcal{G}'(\mathbf{w}) = \mathcal{E}'(\mathbf{w})\mathbf{h}'(\mathbf{w})$  and such that:

- The pair  $(\mathcal{E}, \mathcal{G})$  is an extension of the entropy–entropy flux pair  $(\eta, q)$  for (8.2) which means that, up to an additive constant, the following identities hold:  $\mathcal{E}(\mathcal{M}(\mathbf{u})) = \eta(\mathbf{u})$ ,  $\mathcal{G}(\mathcal{M}(\mathbf{u})) = q(\mathbf{u})$ .
- The minimization principle holds:  $\mathcal{E}(\mathcal{M}(\mathbf{u})) \leq \mathcal{E}(\mathbf{w})$  whenever  $\mathbf{u} = \mathbb{P}(\mathbf{w})$ .
- The source term is dissipative:  $\mathcal{E}'(\mathbf{w})\mathcal{R}(\mathbf{w}) \leq 0$ .

From the definition, we get that any smooth solution of (8.1) satisfies

$$\frac{\partial}{\partial t} \mathcal{E}(\mathbf{w}) + \frac{\partial}{\partial x} \mathcal{G}(\mathbf{w}) = \mu \mathcal{E}'(\mathbf{w}) \mathcal{R}(\mathbf{w}) \leq 0,$$

and as  $\mu \rightarrow \infty$ , it yields (formally) the entropy inequality for (8.2)

$$\partial_t \eta(\mathbf{u}) + \partial_x q(\mathbf{u}) \leq 0.$$

We do not define all the other stability conditions in detail; we only mention the fact that the different stability criteria are not equivalent, and we refer to the work of F. Bouchut [164, 165] where the link between the different conditions is clearly established. In particular, the existence of an entropy extension implies that the characteristic speeds of the reduced system (8.2) are interlaced with the characteristic speeds of the relaxation system (8.1). We will illustrate below this property, called the *interlacing sub-characteristic condition* (also referred to as Whitham's condition), on simple examples. This condition is often easier to handle (see [486] for a practical example of relaxation for two-phase flow models).

In the same way, we will detail the Chapman–Enskog expansion on examples; here we only introduce the great lines. Let  $\mathbf{w} = \mathbf{w}_\mu$  be a solution of the relaxation system (8.1), and then it is expected that  $\mathbf{u}_\mu = \mathbb{P}\mathbf{w}_\mu$  is “nearly” a solution of (8.2) when  $\mu$  is large enough. This can be made more precise by writing  $\mathbf{w}_\mu$  as a power expansion in  $\frac{1}{\mu}$ :

$$\mathbf{w}_\mu = \mathbf{w}_0 + \frac{1}{\mu} \mathbf{w}_1 + \mathcal{O}\left(\frac{1}{\mu^2}\right),$$

and plugging the expansion in (8.1). Then the first term  $\mathbf{w}_0 = \mathcal{M}(\mathbf{u}_0)$  is an equilibrium (a “Maxwellian”), and  $\mathbf{u}_0$  solves (8.2). The *corrector*  $\mathbf{w}_1$  can be determined so that, neglecting terms of order 2 in  $\frac{1}{\mu}$ ,  $\mathbf{u}_\mu$  appears as a solution of a dissipative approximation of (8.2), of order 1 in  $\frac{1}{\mu}$  of the form

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \frac{1}{\mu} \partial_x (D(\mathbf{u}) \partial_x \mathbf{u}).$$

Another stability criterion consists precisely in the fact that this approximation is indeed dissipative ( $D(\mathbf{u})$  nonnegative), and it is again ensured by the existence of an entropy extension.

With the entropy extension, a convergence result can be proved for a system of  $n = 2$  equations [291]. More general results have been proved since by W.-A. Yong [1210, 1211] who introduces for his proofs some related *structural stability* conditions, linked to the well-known Shizuta–Kawashima condition ([1053]; see [854] and references therein). The results in [1210] concern existence of classical solutions of the relaxation system near a smooth equilibrium initial data and convergence as  $\mu \rightarrow \infty$ ; [1211] contains results on the existence of a global in time smooth solution of the relaxation system under initial data close to a *constant* equilibrium state, with a continuity result; other results can be found in [97, 582, 1218].

### 8.1.2 Relaxation and Approximation

Mimicking a physical process, relaxation is used as a tool to approximate a given “equilibrium” system (8.2), from either a theoretical or a numerical point of view. Starting from some hyperbolic system (8.2), it may be possible to introduce a larger but somehow simpler relaxation system (simpler in the sense that it may be easier to study), which writes in the form (8.1), on which interesting results can be proved. Then, from the results concerning the large system, one can deduce results for the equilibrium system. These results are either:

- existence results (see, for instance, [291, 582, 1211]) and some application to the  $p$ -system or to viscoelasticity in [233, 789] and to the Euler system, [234, 268, 801],

- or nice numerical schemes, such as introduced by Jin and Xin [668], Suliciu (see [1076], [164]), or Coquel and Perthame [360] and then [874] for real gas computation or in the context of two-phase flow [92, 949] for the Saint-Venant system.

What one hopes is that the relaxation process is stable, which is obviously necessary when one studies the limit of the solution of the relaxation system as  $\mu \rightarrow \infty$ , all the most since numerical solutions are involved. We mention theoretical results concerning different stability criteria, namely, results taken out of [163]. An interesting survey is given by R. Natalini [895], after [893].

Now, we just sketch the great lines involved in the derivation of a *relaxation scheme* which involves a time-splitting technique. There are three steps:

(i) reconstruction, (ii) evolution, and (iii) projection on the equilibrium manifold. In the first step, from data given at time  $t_n$  for the hyperbolic system (8.2), one reconstructs data at equilibrium which serve as data for the relaxation system (8.1); in the evolution step, on a time interval of length  $\Delta t$ , the homogeneous part of the relaxation system (corresponding to  $\mu = 0$ ) is solved exactly, and it is supposed to be possible in practice because the system is *simple*; last, in step (iii), the solution obtained after step (ii) is projected back on the equilibrium manifold.

The resulting scheme is a Godunov-type scheme, associated with a simple Riemann solver where the speeds of the waves are computed thanks to the relaxation system.

*Remark 8.1.* Let us emphasize the link between the kinetic approach and the general relaxation formalism: Boltzmann equation can be seen as a *kinetic relaxation* system for the Euler system. Consider for (8.2) the Euler equations (7.14) in the one-dimensional case  $d = 1$ . First, we need to make a correspondence with the notations used for Boltzmann equation (7.1): in Sect. 7.1, we noted  $v \in \mathbb{R}$  the velocity of the particle which we denote by  $\xi \in \mathbb{R}$  in the following lines. The (kinetic) relaxation system is semilinear diagonal in *infinite dimension* ( $n = \infty$  in the sense that it is indexed by  $\xi \in \mathbb{R}$ ), and then the connection between the two approaches means more precisely:

- $\mathbf{w} \in \mathbb{R}^n$  in (8.1) corresponds in the kinetic approach to an (infinite) set of functions  $f(., \xi), \xi \in \mathbb{R}$ , and thus  $\mathbf{w}(x, t)$  corresponds to  $f(t, x, \xi), \xi \in \mathbb{R}$ .
- the linear projection operator  $\mathbf{w} \mapsto \mathbb{P}(\mathbf{w})$  corresponds to the operator  $f \mapsto \int Kf(., \xi)d\xi$ ,
- equilibria  $\mathcal{M}(\mathbf{u})$  correspond to the Maxwellian distributions (7.13),
- consistency relations for the equilibria correspond to relations between  $\mathcal{M}$  and its moments.

Also, for the BGK relaxation approximation, with the notations of the general formalism,  $\mathcal{R}(\mathbf{w}) = (\mathbf{0}, \mathbf{e}(\mathbf{u}) - \mathbf{v})^T$ .

Let us mention that in this chapter, many ideas come from expansions with respect to the small parameter  $\tau = 1/\mu$  involved in the relaxation process: Hilbert expansions or Chapman–Enskog expansions. These expansions will be mostly formal, since the justification is often out of scope.

The subject of relaxation/kinetic approximation is an important field of research. The questions are: justify the relaxation toward an equilibrium solution, derive the existence of solutions, and study their properties, for instance, derive solutions for the Euler or Navier–Stokes equations from the study of the Boltzmann equation or from another relaxation approximation system; these are still mainly open questions (see [1163] for the presentation of some recent results on hydrodynamic limits; see also [127, 128, 1063]).

Following the above relaxation/kinetic framework, the numerical schemes for approximating the solution  $\mathbf{u}$  of a system of conservation laws for which there exists a relaxation/kinetic formalism follow three steps (as already said). The states  $\mathbf{u}_i^n$  one wants to compute are equilibrium/*Maxwellian* of the relaxation/kinetic model:

1. reconstruction: from  $\mathbf{u}^n$ , construct an equilibrium state  $\mathbf{w}^n$ /a Maxwellian, i.e.,  $\mathbf{w}^n = \mathcal{M}(\mathbf{u}^n)/f^n(\xi) = M(\xi, \mathbf{u}^n)$ ,
2. time evolution: thanks to (a scheme for) the relaxation/kinetic model, update  $\mathbf{w}^n \rightarrow \mathbf{w}^{n+1-}/f^n(\xi) \rightarrow f^{n+1-}(\xi)$ ,
3. projection on the equilibrium manifold gives  $\mathbf{u}^{n+1} = \mathbb{P}(\mathbf{w}^{n+1-})$ , respectively,  $\int Kf^{n+1-}(\xi)d\xi$  in the kinetic framework.

In the evolution step, only the transport part is solved, the reason why we note with a minus sign. Note that  $\mathbf{v}^{n+1} = \mathbf{e}(\mathbf{u}^{n+1})$ .

Exploiting the way these schemes are built, one can prove that they have good stability properties (such as entropy satisfying) because they mimic a relaxation/kinetic process which has this property. As already seen in the case of kinetic schemes, in the end, the numerical scheme is just a finite volume scheme with explicit numerical flux, and the implementation need not go into the details of each step.

For what concerns relaxation schemes, they are written as HLL-type solvers and can also be studied with little reference to relaxation, as noted in Remark 4.9 (see the textbook of F. Bouchut [163]). We also mention [73, 779] and [933] (this last reference concerns shallow water models) where the link with Roe-type schemes is interesting and well described.  $\square$

## 8.2 Model Examples

We study on a model example how a larger relaxation system can be introduced to approach a given hyperbolic system and how some particular scheme designed for approximating the large system will induce a *simple* scheme for the equilibrium conservation law.

### 8.2.1 Relaxation to a Scalar Conservation Law

Consider the *relaxation p*-system written as

$$\begin{cases} \partial_t u + \partial_x v = 0 \\ \partial_t v + \partial_x p(u) = \mu(f(u) - v); \end{cases} \quad (8.3)$$

the system is hyperbolic if  $p'(u) > 0$  (in the Jin–Xin model (7.70),  $p(u) = a^2 u$ ). As  $\mu \rightarrow \infty$ , we expect the *equilibrium* relation  $v = f(u)$ , which once substituted in the first equation results in the *equilibrium* conservation law

$$\partial_t u + \partial_x f(u) = 0. \quad (8.4)$$

The *equilibrium manifold*  $M$  is the set of states  $(u, v)$  satisfying  $v = f(u)$ .

*Remark 8.2.* The homogeneous system extracted from (8.3) (setting  $\mu = 0$ ) is very close to (but the notations differ from those of) the usual *p*-system (barotropic gas dynamics in Lagrangian coordinates) introduced before

$$\begin{cases} \partial_t v - \partial_x u = 0 \\ \partial_t u + \partial_x p(v) = 0, \end{cases}$$

where  $v > 0$  denotes the specific volume  $1/\varrho$  and  $u$  is the velocity; the pressure is an increasing function of  $\varrho$  and thus a decreasing function of  $v$ . In other applications such as isothermal elasticity (see [789, 1144]), a similar system is involved which is rather written as

$$\begin{cases} \partial_t u - \partial_x v = 0 \\ \partial_t v - \partial_x \sigma(u) = 0, \end{cases}$$

where  $\sigma$  stands for the stress, an increasing function of the deformation  $u \in \mathbb{R}$ , and  $v$  for the velocity.  $\square$

The first question is how can we justify the relaxation limit  $u^\mu \rightarrow u$ , as  $\mu \rightarrow \infty$ , where  $u^\mu, v^\mu$  (resp.  $u$ ) are solution of the *relaxation* system (8.3) (resp. of the *equilibrium* equation (8.4))?

For  $\mu$  large enough, we expect to have  $v^\mu \sim f(u^\mu)$ , and we may try to find a first-order correction term. This is the aim of the *Chapman–Enskog* expansion in inverse powers of  $\mu$ : assume  $u^\mu, v^\mu$  is solution of (8.3), and write

$$v^\mu = f(u^\mu) + \mu^{-1} v_1^\mu + \mathcal{O}(\mu^{-2});$$

the term  $v_1^\mu$  is a *corrector*, which we want to evaluate (at the order  $\mathcal{O}(\mu^{-1})$ ). We write  $v_1^\mu = \mu(v^\mu - f(u^\mu)) + \mathcal{O}(\mu^{-1})$ , and then substituting the expansions in the two equations (as  $\mu \rightarrow \infty$ ), we get

$$\partial_t u^\mu + \partial_x f(u^\mu) = -\mu^{-1} \partial_x v_1^\mu + \mathcal{O}(\mu^{-2}) \quad (8.5)$$

and

$$\partial_t f(u^\mu) + \partial_x p(u^\mu) = -v_1^\mu + \mathcal{O}(\mu^{-1}). \quad (8.6)$$

Expressing the time derivative of  $f(u^\mu)$  up to  $\mathcal{O}(\mu^{-1})$  and plugging in (8.5) gives

$$\begin{aligned} \partial_t(f(u^\mu)) &= f'(u^\mu) \partial_t u^\mu = -f'(u^\mu) \partial_x f(u^\mu) + \mathcal{O}(\mu^{-1}) \\ &= -(f'(u^\mu))^2 \partial_x u^\mu + \mathcal{O}(\mu^{-1}) \end{aligned}$$

and thus together with (8.6), we get

$$(p'(u^\mu) - f'(u^\mu)^2) \partial_x u^\mu = -v_1^\mu + \mathcal{O}(\mu^{-1}).$$

The simplest way to satisfy the above equation is obtained by setting

$$v_1^\mu = -(p'(u^\mu) - f'(u^\mu)^2) \partial_x u^\mu \quad (8.7)$$

thus

$$v^\mu = f(u^\mu) - \mu^{-1} (p'(u^\mu) - f'(u^\mu)^2) \partial_x u^\mu + \mathcal{O}(\mu^{-2})$$

and this precises the correction term  $\mathcal{O}(\mu^{-1})$  in the first Eq. (8.5). Thus, from (8.5), we see that  $u^\mu$  solution of (8.3) satisfies at order  $\mathcal{O}(\mu^{-2})$  Eq. (8.4) with a second-order derivative term

$$\partial_t u + \partial_x f(u) = \frac{1}{\mu} \partial_x ((p'(u) - f'(u)^2) \partial_x u), \quad (8.8)$$

where we have dropped the index  $\mu$ ; we may write  $u \sim u^\mu$ ,  $v \sim v^\mu$  (we have not equality) in the sense that if  $(u^\mu, v^\mu)$  solves exactly (8.3), we may expect that  $u^\mu$  solves approximately (8.8), which is a second-order (in the sense of second-order partial derivative terms) perturbation of (8.4). Stability obviously requires Eq. (8.8) to be dissipative which means

$$p'(u) - f'(u)^2 > 0,$$

equivalently

$$-\sqrt{p'(u)} < f'(u) < \sqrt{p'(u)}, \quad (8.9)$$

which is also called the *sub-characteristic condition* (condition given by Whitham; see [1188]). Note that  $\pm\sqrt{p'(u)}$  are the two characteristic speeds of system (8.3), while  $f'(u)$  is the characteristic speed of (8.4). For stability

reasons, waves in the limit (coarser description) cannot travel faster than in the approximating sequence (finer description): this is a natural causality principle.

For this simple example, the convergence of the relaxation system toward the conservation law is proved; we refer to [293].

*Remark 8.3.* The linear example  $f(u) = Au$  is thoroughly treated in [895] with explicit computations (one can exhibit an elementary solution  $(u_\mu, v_\mu)$ ), and the condition under which a sequence is bounded (so that one can extract a convergent subsequence) is naturally  $a^2 \geq A^2$ ; see also [785].  $\square$

*Remark 8.4.* Let us make a link with another type of asymptotic behavior on a simple example. Consider the system (8.3) with a specific right-hand side

$$\begin{cases} \partial_t u + \partial_x v = 0 \\ \partial_t v + \partial_x p(u) = -\mu v, \end{cases} \quad (8.10)$$

where  $\mu$  is a *friction* (damping) coefficient. Comparing with (8.3) gives formally  $f(u) = 0$ , and (8.5) (8.7) become

$$\begin{cases} \partial_t u = \frac{1}{\mu} \partial_x (p'(u) \partial_x u), \\ v = -\frac{1}{\mu} \partial_x p(u); \end{cases} \quad (8.11)$$

thus, at first order as  $\mu \rightarrow \infty$ , the system (8.10) exhibits a parabolic-type asymptotic behavior with the velocity given by the second equation in (8.11), which is known as Darcy's law. Let us perform the scaling  $t = \mu s$ ,  $w = \mu v$ , so that the long time behavior for large friction of solutions of (8.10) leads to consider the system

$$\begin{cases} \partial_s u - \partial_{xx} (p(u)) \\ w = -\partial_x p(u). \end{cases} \quad (8.12)$$

It is an example of diffusive relaxation. There are also convergence results for this example (see Serre-Xiao [1045] [635], also [743], [634]). A similar example is given by the isothermal Euler equations [370] with convergence of the density to the heat equation; [801] concerns a more general barotropic law.

It is important to design schemes which are consistent with the asymptotic limit: the so-called AP (asymptotic preserving) schemes. We will come again on the subject in the chapter concerning source terms.  $\square$

### 8.2.2 Introduction to the Jin–Xin Relaxation Scheme

Choosing  $p(v) = a^2 v$  in (8.3) gives

$$\begin{cases} \partial_t u + \partial_x v = 0 \\ \partial_t v + a^2 \partial_x u = \mu(f(u) - v), \end{cases} \quad (8.13)$$

where  $a > 0$  is a constant, satisfying Whitham's stability condition

$$-a < f'(u) < a. \quad (8.14)$$

We have already encountered this model system in (7.70). The main idea of a relaxation scheme, which we illustrate on this simple example, is that an appropriate discretization of (8.13) for  $\mu$  large enough will provide an approximation of the solution  $u$  of the conservation law (8.4)  $\partial_t u + \partial_x f(u) = 0$ . For instance, let us consider for (8.13) the upwind scheme in characteristic variables. We need to diagonalize the homogenous system (8.13) with Riemann invariants (see Chap. II, Sect. 7.1)

$$w = v - au, \quad z = v + au,$$

each propagating with one characteristic speed  $\pm a$ , which writes

$$\begin{cases} \partial_t w - a\partial_x w = 0, \\ \partial_t z + a\partial_x z = 0. \end{cases} \quad (8.15)$$

The flux of (8.13) is  $(v, a^2 u)^T$ ; let us write similarly the numerical flux in the form  $g_{j+1/2} = (v_{j+1/2}, a^2 u_{j+1/2})^T$ . The upwind scheme for the linear homogeneous part once diagonalized (8.15) gives the fluxes  $w_{j+1/2} = w_{j+1}$  since  $-a < 0$  and similarly  $z_{j+1/2} = z_j$ ; hence

$$\begin{cases} w_{j+1/2} = (v - au)_{j+1/2} = v_{j+1} - au_{j+1}, \\ z_{j+1/2} = (v + au)_{j+1/2} = v_j + au_j, \end{cases}$$

and the flux  $g_{j+1/2}$  writes

$$\begin{cases} u_{j+1/2} = \frac{1}{2}(u_j + u_{j+1}) - \frac{1}{2a}(v_{j+1} - v_j), \\ v_{j+1/2} = \frac{1}{2}(v_j + v_{j+1}) - \frac{1}{2}(u_{j+1} - u_j). \end{cases}$$

Inverting the relations  $(u, v) \mapsto (w, z)$  yields

$$u = (z - w)/2a, \quad v = (w + z)/2,$$

and for the fully discrete first-order scheme for approximating (8.13), we take the source term into account in a second step, and this gives for  $u, v$  (with  $\lambda = \Delta t / \Delta x$ )

$$\begin{cases} u_j^{n+1} = u_j^n - \frac{\lambda}{2}(v_{j+1}^n - v_{j-1}^n) + \frac{\lambda a}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n), \\ v_j^{n+1} = v_j^n - \frac{\lambda a^2}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{\lambda a}{2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n) \\ \quad + \mu \Delta t(f(u_j^{n+1}) - v_j^{n+1}). \end{cases} \quad (8.16)$$

Though, to get (8.16), we have treated the source term with an implicit method, it is in fact an explicit scheme, since the first equation gives  $u_j^{n+1}$ , and in the second, one can easily express  $v_j^{n+1}$  in terms of known values. In variable  $(w, z)$ , it writes

$$\begin{cases} w_j^{n+1} = w_j^n + \lambda a(w_{j+1}^n - w_j^n) + \mu \Delta t(f(u_j^{n+1}) - v_j^{n+1}), \\ z_j^{n+1} = z_j^n - \lambda a(z_{j+1}^n - z_{j-1}^n) + \mu \Delta t(f(u_j^{n+1}) - v_j^{n+1}), \end{cases}$$

which is not so obviously explicit! One can prove that for fixed  $\mu$ , the sequence of solutions  $(u, v)_{\Delta, \mu}$  of the scheme (at least a subsequence) converges to a solution  $(u, v)_\mu$  of the relaxation system as  $\Delta t, \Delta x \rightarrow 0$ , while, when moreover  $\mu \rightarrow \infty$ ,  $u_{\Delta, \mu}$  (a subsequence) converges to a solution of the scalar equation.

Since the system is linear, the upwind scheme coincides with Godunov's scheme for the first-order system (8.15) in variables  $(w, z)$  and thus corresponds to the exact Riemann solver in characteristic variables  $(w, z)$

$$W_R(x/t; (w_l, z_l), (w_r, z_r)) = \begin{cases} (w_l, z_l) & \frac{x}{t} < -a, \\ (w_r, z_l) & -a < \frac{x}{t} < a, \\ (w_r, z_r) & \frac{x}{t} > a, \end{cases} \quad (8.17)$$

and (8.17) may then be written in conservative variables  $(u, v)$ . Starting from initial data at equilibrium,  $(u_j^0, v_j^0 = f(u_j^0))$  in (8.16), and adding the projection on the equilibrium manifold which means replacing  $v_j^{n+1}$  by  $f(u_j^{n+1})$  gives a Lax–Friedrichs-type scheme for (8.4) which writes

$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}(f(u_{j+1}^n) - f(u_{j-1}^n)) + \frac{\lambda a}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n). \quad (8.18)$$

This scheme is thus associated with the exact Riemann solver for the homogeneous part of the relaxation system (8.3) followed by projection on the equilibrium manifold  $(u, v = f(u))$  (this last step is also called instantaneous relaxation). It results in a Godunov-type scheme for the scalar equation associated with an approximate Riemann solver

$$w(x/t; u_l, u_r) = \begin{cases} u_l, & \frac{x}{t} < -a \\ \frac{1}{2}(u_l + u_r) - \frac{1}{2a}(f(u_r) - f(u_l)), & -a < \frac{x}{t} < a, \\ u_r, & \frac{x}{t} > a. \end{cases}$$

This formula is directly obtained from (8.17) by combining the change of variables  $(w, z) \mapsto (u, v)$  which gives for the first component

$$w_R(x/t; u_l, u_r) = \begin{cases} u_l, & \frac{x}{t} < -a, \\ \frac{1}{2}(u_l + u_r) - \frac{1}{2a}(v_r - v_l), & -a < \frac{x}{t} < a, \\ u_r, & \frac{x}{t} > a, \end{cases}$$

with a simultaneous projection on the equilibrium manifold  $(u, v = f(u))$ . The numerical flux is given by  $v_{j+1/2} = \frac{1}{2}(v_j + v_{j+1}) - \frac{a}{2}(u_{j+1} - u_j)$  which thus writes

$$g(u_l, u_r) = \frac{1}{2}(f(u_l) + f(u_r)) - \frac{a}{2}(u_r - u_l).$$

The constant  $a$  needs to be chosen *large* enough, but the numerical diffusion increases with  $a$ ; for accuracy reasons,  $a$  should be chosen as *small* as possible. The Rusanov scheme optimizes  $a$ , under the constraint (8.14)

$$a = \sup_{(u_l, u_r)} |f'(u)|.$$

This scheme is monotone under CFL 1.

Note that we have developed the complete scheme (8.16) for the full relaxation system (8.13) because of its simplicity. However, if we are only interested in a numerical scheme for the conservation law (8.4), we do not need to take the whole source term into account: the reconstruction step means  $u_j^0 \mapsto (u_j^0, f(u_j^0))$ , and the general evolution step of the relaxation scheme writes as (8.16) with  $\mu = 0$ , i.e., the upwind scheme for the homogeneous part of the linear relaxation system (8.13). This step is followed by an instantaneous relaxation, i.e., a projection step on the equilibrium manifold  $\{(u, v); v = f(u)\}$ , and keeping the first component allows to update  $u_j^n$ . Thus we obtain a simple and robust numerical scheme for the scalar equation which is a Godunov-type scheme with an approximate Riemann solver.

### 8.2.3 Generalization

For a system of  $p$  equations, a straightforward extension gives a relaxation system of  $n = 2p$  equations

$$\begin{cases} \partial_t \mathbf{u} + \partial_x \mathbf{v} = \mathbf{0}, \\ \partial_t \mathbf{v} + \mathbf{A} \partial_x \mathbf{u} = \mu(\mathbf{f}(\mathbf{u}) - \mathbf{v}), \end{cases} \quad (8.19)$$

where  $\mathbf{A}$  is now a constant diagonal matrix with positive entries. Similarly, one can introduce a relaxation scheme and prove convergence to a solution of the equilibrium system (see [741]). The choice of  $a$  in the resulting Rusanov scheme will be

$$a = \sup_{\mathbf{u}} \sup_j |\lambda_j(\mathbf{u})|.$$

The above Jin–Xin relaxation scheme works whatever the hyperbolic system under consideration; see [668]. However, for a given (equilibrium) system, it is not always necessary to double the variables, and it may even lead to a scheme which is too diffusive. If some equations are already linear or so simple that they can be kept in the relaxation system, we may try to mimic the above numerical approach for approximating only the specific nonlinear terms. We now illustrate this approach on the simple example of the  $p$ -system, for which only the second equation involves a nonlinear function.

### 8.2.4 Selective Relaxation for the $p$ -System

Consider the  $p$ -system (barotropic Euler system in Lagrangian variables), given by

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x p = 0, \end{cases} \quad (8.20)$$

with  $\tau = 1/\varrho$  the specific volume and the pressure  $p = p(\tau)$  satisfying  $p' < 0$  for hyperbolicity.

For the nonlinear scalar conservation law  $\partial_t u + \partial_x f(u) = 0$ , we can understand the computations in Sect. 8.2.1 (with the choice  $p(v) = a^2 v$  in (8.3) as a “linearization” of the nonlinear flux  $f(u)$  by  $au$ ). This cannot be done globally, without any other modification; otherwise the approximation would be poor; hence we have added a new variable  $v \sim f(u)$  (in the sense that  $v$  is a kind of extension of  $f(u)$ , expected to be close to  $u$  for  $\mu$  large enough); thus  $\partial_t u + \partial_x v \sim 0$ , and we have looked for a law satisfied by  $v$ , starting from the equation satisfied by  $f(u)$ . Since  $\partial_t f(u) + f'(u)^2 \partial_x u = 0$ , if  $(f'(u))^2 \sim a^2$ , this has led us to write the second equation  $\partial_t v + a^2 \partial_x u = \mu(f(u) - v)$ .

The  $p$ -system is nonlinear only because of the pressure  $p$  (the first equation is linear, and we can keep it). If we mimic the scalar case, we consider the larger linear system

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x \Pi = 0, \\ \partial_t \Pi + c^2 \partial_x u = \mu(p - \Pi) \end{cases} \quad (8.21)$$

with  $\Pi \sim p$ . The equation for  $\Pi$  comes exactly as explained above in the scalar case: from multiplying the first equation by  $p'(\tau)$ , which gives  $\partial_t p(\tau) - p'(\tau) \partial_x u = 0$ , and linearizing  $p$  by replacing  $p'$  by  $-c^2$ , in accordance with  $p' < 0$ . The homogeneous system has eigenvalues  $\pm c, 0$ ; it is thus linearly degenerate. Whitham’s stability condition writes  $c^2 > -p'(\tau)$  (for all  $\tau$  under consideration) which implies that the eigenvalues are interlaced

$$-c < -\sqrt{-p'(\tau)} < 0 < \sqrt{-p'(\tau)} < c.$$

We can try to perform a Chapman–Enskog expansion as for the scalar case. Looking for a first-order corrector, say  $\Pi^1$  so that

$$\Pi = p + \mu^{-1} \Pi^1 + \mathcal{O}(\mu^{-2})$$

(we skip all the superscripts  $\mu$  for simplicity), we have  $\Pi^1 = \mu(\Pi - p) = -(\partial_t \Pi + c^2 \partial_x u) \approx -(\partial_t p + c^2 \partial_x u)$  at the order  $\mathcal{O}(\mu^{-1})$ , and then we write  $\partial_t p + c^2 \partial_x u = p' \partial_t \tau + c^2 \partial_x u = (p' + c^2) \partial_x u$  so that, from the second equation, we write (neglecting higher-order terms  $\mathcal{O}(\mu^{-2})$ )

$$\partial_t u + \partial_x p = -\mu^{-1} \partial_x \Pi^1 = \mu^{-1} \partial_x ((p' + c^2) \partial_x u),$$

and the right-hand side is a diffusive term under Whitham's condition.

Instead of system (8.21), it may be more convenient to consider the relaxation system

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x \Pi = 0, \\ \partial_t \mathcal{T} = \mu(\tau - \mathcal{T}), \end{cases} \quad (8.22)$$

where  $\Pi$  is now

$$\Pi(\tau, \mathcal{T}) = p(\mathcal{T}) + c^2(\mathcal{T} - \tau). \quad (8.23)$$

Thus the nonlinear pressure law  $p(\tau)$  has been “linearized” at the expense of adding a new variable  $\mathcal{T}$  (we write  $\mathcal{T} \sim \tau$ , in the sense that it kind of extends the specific volume), and  $\Pi(\tau, \mathcal{T}) = p(\mathcal{T}) + c^2 \mathcal{T} - c^2 \tau$  is now linear wrt.  $\tau$ , with  $\partial_\tau \Pi = -c^2$  (a negative constant since  $p' < 0$ ). Since this new pressure law  $\Pi = \Pi(\tau, \mathcal{T})$  should coincide with  $p$  at equilibria, and the relaxation system should relax to the  $p$ -system, we have  $\Pi = -c^2 \tau + h(\mathcal{T}) \rightarrow p$  when the system “relaxes.” We have thus chosen the simplest way to achieve all these constraints, introducing as new variable  $\mathcal{T}$ , a kind of extended volume fraction, and  $\mathcal{T} \rightarrow \tau$  as  $\mu \rightarrow \infty$  so that  $h(\tau) = p(\tau) + c^2 \tau$ ; moreover  $\mathcal{T}$  satisfies the simplest possible PDE with a relaxation right-hand side.

The eigenvalues of the homogeneous part of (8.22) are again 0 and  $\pm c$ ; we assume that  $c$  satisfies Whitham's (or (sub-characteristic) condition  $c^2 > \max(-p'(s))$ ). Note that under this condition, the mapping  $h : \xi \mapsto p(\xi) + c^2 \xi$  defined on  $]0, \infty[$  is invertible so that  $\mathcal{T} = h^{-1}(\Pi + c^2 \tau)$ . This proves that the mapping  $(\tau, u, \Pi) \mapsto (\tau, u, \mathcal{T})$  induces an admissible change of variables, so that the homogenous systems (8.21) and (8.22) are equivalent, and the relaxation systems too, provided  $\mu$  in (8.21) is replaced by  $\mu(p' + c^2)/c^2$ .

This relaxation system was introduced by Suliciu [1076], as an approximation for the equations of isothermal viscoelasticity, and it has been studied in several recent papers; see, for instance, [163, 268, 582, 789, 1144, 1145], and [742]; both [233, 268] use Yong's results [1210]. In fact, in [233], the authors work with variables  $(\tau, u, \Pi)$ , and in the annex of [233], the authors prove, under some stronger assumption,  $c^2 > \Gamma_M \geq \max(-\tilde{p}'(s)) \geq \Gamma_m > 0$  and for smooth initial data, the convergence (as  $\mu \rightarrow \infty$ ) of the global solution of the relaxation system to a local in time solution of the  $p$ -system

(8.20); [268] also addresses the full system with energy (written in Eulerian coordinates) for which the authors have an existence result for the relaxation system (near a data at equilibrium) and convergence to a solution of the Euler equations. Moreover in [268], a numerical procedure is introduced; the results are extended to fluid systems [349]. In [234], which investigates more generally totally linearly degenerate systems with relaxation, the Suliciu system is proved to be *rich* [1041].

### 8.3 A Relaxation Scheme for the Euler System

For the sake of completeness, we first detail the relaxation system for Euler system in Eulerian coordinates, though the computations are simpler in the Lagrangian frame, and we could use the correspondence between the two frames, Lagrangian and Eulerian, already mentioned in Chap. II (Sect. 2.1). Indeed all properties (linearly degenerate fields, for instance) can be proved by using a general equivalence result, without new proofs. However, we detail some of the computations since the Eulerian frame is frequently used in the applications. Among the applications, one may think of the shallow water model which corresponds to a particular barotropic Euler system, and in case of nonconstant topography, the Lagrangian frame is not adequate.

#### 8.3.1 Suliciu Relaxation for the Isentropic Euler System

We consider the system

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + p) = 0, \end{cases} \quad (8.24)$$

with (nonlinear) pressure law  $p = p(\varrho, s_0)$ ; the entropy  $s = s_0$  is constant since we are in the isentropic case, and we will just note  $p = p(\varrho)$  (barotropic law), and we assume  $p'(\varrho) \equiv \partial_\varrho p(\varrho, s) > 0$ .

We introduce a larger  $3 \times 3$  system with a relaxation term and a relaxation parameter  $\mu$ . To obtain the larger system, with the idea of “relaxing” the only nonlinear term which is the pressure, we first derive an equation for  $p$ ; multiplying the first developed equation by  $p'(\varrho)$  gives

$$\partial_t p(\varrho) + u \partial_x p(\varrho) + \varrho p'(\varrho) \partial_x u = 0,$$

and together with (again) the density equation, we can write

$$\partial_t \varrho p(\varrho) + \partial_x(\varrho p(\varrho) u) + \varrho^2 p'(\varrho) \partial_x u = 0.$$

We introduce a new variable  $\Pi$  in place of  $p(\varrho)$  and a constant Lagrangian sound velocity  $c$  and put  $c^2$  in place of  $\varrho^2 p'(\varrho)$  and exploit the fact that

$$\partial_t(\varrho\Pi) + \partial_x(\varrho\Pi u + c^2 u) \text{ should be } 0 \text{ if } p = \Pi.$$

Thus, we consider the system

$$\begin{cases} \partial_t\varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + \Pi) = 0, \\ \partial_t(\varrho\Pi) + \partial_x(\varrho\Pi u + c^2 u) = \mu\varrho(p - \Pi). \end{cases} \quad (8.25)$$

*Remark 8.5.* There is another way of “relaxing” the nonlinear momentum equation. In [933], another relaxation system is introduced in the context of shallow water, which means for  $p(\varrho) = \kappa\varrho^\gamma$  with  $\gamma = 2$  then extended to a general  $\gamma \geq 1$ . For  $\gamma = 2$ , a variable  $\omega$  is introduced for the relaxation of the momentum  $m = \varrho u$ , and the momentum equation is linearized about an average state (with density and velocity  $\tilde{\varrho}, \tilde{u}$  adequately chosen), giving the relaxation system

$$\begin{cases} \partial_t\varrho + \partial_x(\varrho u) = 0 \\ \partial_t m + \partial_x(\frac{\omega^2}{\varrho} + p(\varrho)) = 0 \\ \partial_t\omega + (-\tilde{u}^2 + p'(\tilde{\varrho}))\partial_x\varrho + 2\tilde{u}\partial_x\omega = \mu(m - \omega). \end{cases}$$

Still in the context of shallow water, another relaxation system is introduced in [141] with the same aim of deriving a scheme with good properties; however, no theoretical approximation result is known at present.  $\square$

The choice of  $c$  in (8.25) is such that Whitham’s condition is satisfied

$$\varrho^2 p'(\varrho) \leq c^2$$

for all  $\varrho$  in some interval  $\mathcal{I}$  included in  $\varrho > 0$ , and this will ensure that the associated Riemann solver:

- preserves the positivity of  $\varrho$
- is entropy satisfying.

*Remark 8.6.* From the third equation in (8.25), we get

$$\partial_t\Pi + u\partial_x\Pi + \frac{c^2}{\varrho}\partial_xu = \mu(p - \Pi).$$

Note the similarity with the equation (see (5.8), Chap. II) for the pressure satisfied by the exact solution of Euler system

$$\partial_tp + u\partial_xp + \varrho c_{\mathcal{E}}^2\partial_xu = 0,$$

where  $c_{\mathcal{E}}^2$  is the square of the sound velocity in Eulerian coordinates, given by  $p'(\varrho)$  in the barotropic case. If  $c^2$  denotes the square of the velocity in the Lagrangian system, it becomes after transformation in Eulerian coordinates  $c^2\tau^2$ . Indeed, in the barotropic case, denoting  $p(\varrho) = \tilde{p}(\tau)$  to distinguish the two mappings, we have  $p'(\varrho) = -\tau^2\tilde{p}'(\tau)$ , so that in fact the term  $\varrho c_{\mathcal{E}}^2 = \varrho^2 c_{\mathcal{E}}^2 / \varrho$  does correspond to the term  $c^2/\varrho$ .  $\square$

The homogenous system (8.25) has three real eigenvalues  $u, u \pm c/\varrho$  and only linearly degenerate fields, and the resolution of the Riemann problem is made easy. The system can be diagonalized

$$\begin{cases} \partial_t(\Pi + cu) + (u + c/\varrho)\partial_x(\Pi + cu) = 0, \\ \partial_t(\Pi - cu) + (u - c/\varrho)\partial_x(\Pi - cu) = 0, \\ \partial_t(1/\varrho + \Pi/c^2) + u\partial_x(1/\varrho + \Pi/c^2) = 0. \end{cases} \quad (8.26)$$

As we did above for the system in Lagrangian coordinates, we consider another formulation (introducing the new variable  $\mathcal{T}$ )

$$\begin{cases} \partial_t\varrho + \partial_x(\varrho u) = 0 \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + \Pi) = 0 \\ \partial_t(\varrho\mathcal{T}) + \partial_x(\varrho\mathcal{T}u) = \mu\varrho(\tau - \mathcal{T}). \end{cases} \quad (8.27)$$

This new system needs a *closure relation* for  $\Pi$  which we take in the same form as (8.23), emphasizing the notations for the different mappings

$$\Pi = \Pi(\varrho, \varrho\mathcal{T}) = \tilde{\Pi}(\mathcal{T}, \tau) \equiv \tilde{p}(\mathcal{T}) + c^2(\mathcal{T} - \tau), \quad (8.28)$$

with again  $\tau = 1/\varrho$  and  $\tilde{p}(\tau) = p(\varrho)$ . Formally as  $\mu \rightarrow \infty$ ,  $\mathcal{T} - \tau \rightarrow 0$ , so that  $\varrho\mathcal{T} \rightarrow 1$ , and  $\tilde{\Pi}(\mathcal{T}, \tau) \rightarrow \tilde{p}(\tau) = p$ , and we recover the Euler system (8.24) at equilibrium.

Setting  $\mathbf{U} = (\varrho, \varrho u, \varrho\mathcal{T})^T$ , the Jacobian matrix of system (8.27) is given by

$$\mathbf{A}(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 \\ -u^2 + \Pi_{\varrho} & 2u & \Pi_{\varrho}\mathcal{T} \\ -u\mathcal{T} & \mathcal{T} & u \end{pmatrix}.$$

We check again that the eigenvalues are  $\lambda_1(\mathbf{U}) = u - c\tau < \lambda_2(\mathbf{U}) = u < \lambda_3(\mathbf{U}) = u + c\tau$ . The corresponding eigenvectors can be taken as

$$\mathbf{r}_1(\mathbf{U}) = (1, u - c\tau, \mathcal{T})^T, \quad \mathbf{r}_2(\mathbf{U}) = (1, u - \Pi_{\varrho}/\Pi_{\varrho}\mathcal{T})^T, \quad \mathbf{r}_3(\mathbf{U}) = (1, u + c\tau, \mathcal{T})^T.$$

Moreover, the three characteristic fields are linearly degenerate. Let us characterize the contact wave curves associated with each field.

*Lemma 8.1*

Assume that  $c$  satisfies

$$c^2 > \max(-\tilde{p}'(\xi)), \quad (8.29)$$

and let  $\mathbf{U}_\ell$  be a given state. Then the wave curves  $\mathcal{C}_i(\mathbf{U}_\ell)$ ,  $i = 1, 2, 3$ , can be characterized as follows:

- The curve  $\mathcal{C}_1(\mathbf{U}_\ell)$  is given by  $\mathcal{C}_1(\mathbf{U}_\ell) = \{\mathbf{U} \in \Omega; u = u_\ell - \frac{1}{c}(\Pi - \Pi_\ell), \mathcal{T} = \mathcal{T}_\ell\}$ .
- The curve  $\mathcal{C}_2(\mathbf{U}_\ell)$  is given by  $\mathcal{C}_2(\mathbf{U}_\ell) = \{\mathbf{U} \in \Omega; u = u_\ell, \Pi = \Pi_\ell\}$ .
- The curve  $\mathcal{C}_3(\mathbf{U}_\ell)$  is given by  $\mathcal{C}_3(\mathbf{U}_\ell) = \{\mathbf{U} \in \Omega; u = u_\ell + \frac{1}{c}(\Pi - \Pi_\ell), \mathcal{T} = \mathcal{T}_\ell\}$ .

In (8.29), the supremum is taken over all possible values of the specific volume occurring in the solution of a Riemann problem.

*Proof.* The result for  $\mathcal{C}_2(\mathbf{U}_\ell)$  comes directly from the Rankine–Hugoniot jump condition. For  $\mathcal{C}_1(\mathbf{U}_\ell)$ , the Rankine–Hugoniot jump condition gives

$$[\varrho u^2 + \Pi] = \lambda_1 [\varrho u] = (u - c\tau) \varrho u - (u_\ell - c\tau_\ell) \varrho_\ell u_\ell,$$

which yields that  $\Pi + au$  is invariant across a 1-wave:

$$\Pi + cu = \Pi_\ell + cu_\ell.$$

This last relation gives in turn that  $\tilde{p}(\mathcal{T}) + c^2\mathcal{T}$  is also a 1-invariant. Indeed

$$\Pi + cu = \tilde{p}(\mathcal{T}) + c^2(\mathcal{T} - \tau) + cu = \tilde{p}(\mathcal{T}) + c^2\mathcal{T} + c\lambda_1;$$

hence  $\tilde{p}(\mathcal{T}) + c^2\mathcal{T}$  is the difference of two invariants. Now the assumption on  $c$  implies that the mapping  $h : s \mapsto \tilde{p}(s) + c^2s$  is strictly increasing and thus invertible; this yields that  $\mathcal{T}$  is also a 1-invariant.

The results for the 3-invariants follow similarly.  $\square$

### Proposition 8.1

Given two constant states  $\mathbf{U}_\ell, \mathbf{U}_r$ , the solution of the Riemann problem for the homogeneous system (8.27) consists of three contact discontinuities, each propagating with a characteristic speed  $\lambda_i$ ,  $i = 1, 2, 3$ , separating  $\mathbf{U}_\ell$ , two intermediate states  $\mathbf{U}_\ell^*, \mathbf{U}_r^*$ , and  $\mathbf{U}_r$ . The states  $\mathbf{U}_\ell^*, \mathbf{U}_r^*$  are, respectively, characterized by  $(u^*, \Pi^*; \mathcal{T}_\ell, \tau_\ell^*)$  and  $(u^*, \Pi^*; \mathcal{T}_r, \tau_r^*)$  with

$$\begin{cases} u^* = \frac{u_\ell + u_r}{2} - \frac{\Pi_r - \Pi_\ell}{2c}, \\ \Pi^* = \frac{\Pi_\ell + \Pi_r}{2} - \frac{c}{2}(u_r - u_\ell), \\ \tau_\ell^* = \tau_\ell - \frac{1}{2c^2}(\Pi_r - \Pi_\ell - c(u_r - u_\ell)), \\ \tau_r^* = \tau_r + \frac{1}{2c^2}(\Pi_r - \Pi_\ell + c(u_r - u_\ell)). \end{cases} \quad (8.30)$$

*Proof.* The projections of  $\mathcal{C}_1(\mathbf{U}_\ell)$  and  $\mathcal{C}_3(\mathbf{U}_r)$  on the  $(u, \Pi)$ –plane are given by Lemma 8.1

$$u = u_\ell - \frac{1}{c}(\Pi - \Pi_\ell), \quad u_r = u + \frac{1}{c}(\Pi_r - \Pi).$$

Their intersection determines the common value  $(u^*, \Pi^*)$  of the intermediate states. Now, using the property that the eigenvalues are Riemann invariants,  $\lambda_1(\mathbf{U}_\ell) = \lambda_1(\mathbf{U}_\ell^*)$  and  $\lambda_3(\mathbf{U}_r^*) = \lambda_3(\mathbf{U}_r)$ , gives the other formulas. The fact that the intermediate states belong to  $\Omega$ , i.e.,  $\tau_\ell^* > 0, \tau_r^* > 0$ , is equivalent to the ordering of the three waves in the solution of the Riemann problem

$$\lambda_1(\mathbf{U}_\ell) < u^* < \lambda_3(\mathbf{U}_r),$$

since

$$u^* - \lambda_1(\mathbf{U}_\ell) = c\tau_\ell^*, \quad \lambda_3(\mathbf{U}_r) - u^* = c\tau_r^*.$$

We assume with Whitham's condition that  $c$  is large enough so that all intermediate states in the solution of the Riemann problem have a finite density ( $\tau > 0$ ).  $\square$

*Remark 8.7.* We will not go into details concerning the treatment of vacuum which needs a particular care. Indeed, if one of the two densities  $\varrho_l$  or  $\varrho_r$  tends to 0, while the other remains finite, the extreme eigenvalues  $u_l - c/\varrho_l, u_r + c/\varrho_r$  will tend to infinity, and, unless  $c$  goes to 0, the CFL condition enforces a too restrictive time step. We refer to the monograph [163], Section 2.4.5, which allows to modify  $c$  and introduces well-chosen values  $c_l, c_r$ .  $\square$

### 8.3.2 Entropy and the Relaxation Procedure

It is easy to check that the isentropic Euler system (8.24) has a convex (mathematical) entropy  $\eta$  (which is the energy) and associated entropy flux noted  $\theta$  in this section, respectively, given by

$$\eta(\mathbf{u}) = \varrho\varepsilon(\varrho) + \frac{1}{2}\varrho u^2 = \varrho e, \quad \theta(\mathbf{u}) = (\eta + p)u, \quad (8.31)$$

which are functions of the conservative variable  $\mathbf{u} = (\rho, \varrho u)$  or simply of  $(\varrho, u)$ , where the internal  $\varepsilon$  (resp. total) energy (resp.  $e$ ) is defined by

$$\varepsilon'(\varrho) = \frac{p(\varrho)}{\varrho^2}, \quad \text{and} \quad e = \varepsilon + \frac{1}{2}u^2.$$

Since the homogeneous system (8.27) is linearly degenerate, we do not need a strictly convex entropy to select its admissible solutions. However, in order to justify the relaxation procedure, it is important to exhibit a convex extension of  $\eta$  and an inequality when we take into account the relaxation term in the last equation of (8.27) as we will see a little further on. Moreover, the result we obtain proves the entropy character of the associated scheme.

Thus we want to prove that the relaxation system (8.27) is endowed with an *entropy extension* relative to the entropy  $\eta$  of (8.24), i.e., that coincides with  $\eta$  on the equilibrium manifold. If we can indeed exhibit an entropy extension, which we note  $\mathcal{E}$ , and associated entropy flux  $\mathcal{G} = (\mathcal{E} + \Pi)u$ , this extension is not strictly convex on the whole domain, and we cannot apply the results of Section 2.4 in [163]. However, it is convex on the equilibrium manifold, and some minimum entropy principle holds, as results from the following propositions (we use again the notation  $\tilde{\varphi}(\tau) = \varphi(\frac{1}{\tau})$ ).

*Proposition 8.2*

Let

$$\begin{cases} \Sigma(\tau, u, \mathcal{T}) = \tilde{\varepsilon}(\mathcal{T}) + \frac{1}{2}u^2 + \frac{1}{2c^2}(\tilde{\Pi}^2(\tau, \mathcal{T}) - \tilde{p}^2(\mathcal{T})), \\ \mathcal{E} = \varrho\Sigma, \text{ and } \mathcal{H} = \mathcal{E} + \Pi, \end{cases} \quad (8.32)$$

where  $\tilde{\Pi}$  is given by (8.28). Then, for equilibrium states, we have

$$\mathcal{E}(\tau, u, \tau) = \eta(\mathbf{u});$$

smooth solutions of (8.27) satisfy

$$\partial_t \mathcal{E} + \partial_x \mathcal{H}u = -\mu\varrho(\tilde{p}'(\mathcal{T}) + c^2)(\tau - \mathcal{T})^2, \quad (8.33)$$

and the right-hand side of (8.33) is negative if condition (8.29) holds.

We also have the formula

$$\mathcal{E}(\varrho, u, \Pi) = \varrho u^2/2 + \varrho\varphi(\Pi + c^2/\varrho) + \varrho\Pi^2/2c^2,$$

(here we have not used different notations for the mappings, and  $\mathcal{E}$  can be written as a function of  $(\tau, u, \mathcal{T})$  or  $(\varrho, u, \Pi)$  according to the context) where the function  $\varphi$  satisfies

$$\varphi(p(\varrho) + c^2/\varrho) = \varepsilon(\varrho) - p(\varrho)^2/2c^2.$$

This relation defines  $\varphi$  since  $h : s \mapsto \tilde{p}(s) + c^2s$  is strictly increasing and is thus invertible under condition (8.29).

*Proof.* We do not go into every details of the proof, which needs some computations. We do have a natural definition of the energy for the system (8.27) and the function

$$\mathcal{H}_1(\mathbf{U}) = \varrho\epsilon(\varrho, \mathcal{T}) + \frac{1}{2}\varrho u^2,$$

where

$$\frac{\partial\epsilon}{\partial\varrho}(\varrho, \mathcal{T}) = \frac{\Pi}{\varrho^2}$$

is a (nonconvex) entropy for system (8.27) with entropy flux

$$\mathcal{G}_1(\mathbf{U}) = u(\mathcal{H}_1(\mathbf{U}) + \Pi).$$

Thus we look for another (convex) extension, and for that aim, we write other conservation laws derived from the system (8.27). Combining the equations of (8.27) multiplied by  $\partial_\tau \Pi(\tau, \mathcal{T})$  or  $\partial_{\mathcal{T}} \Pi(\tau, \mathcal{T})$ , it is not difficult to check that smooth solutions of (8.27) satisfy

$$\partial_t \left( \frac{\varrho u^2}{2} + \frac{\varrho \Pi^2}{2c^2} \right) + \partial_x \left( \left( \frac{\varrho u^2}{2} + \frac{\varrho \Pi^2}{2c^2} + \Pi \right) u \right) = \nu \varrho \frac{\Pi}{c^2} (p'(\mathcal{T}) + c^2)(\tau - \mathcal{T}).$$

The sign of the right-hand side is not known. Hence, we will construct an extension of the function  $\eta$  in the form

$$\varrho \Sigma = \frac{\varrho u^2}{2} + \frac{\varrho \Pi^2}{2c^2} + \varrho \Phi(\mathcal{T}),$$

and we look for a function  $\Phi(\mathcal{T})$  such that, as a first condition, the function  $\varrho \Sigma$  coincides with  $\eta$  at equilibrium states. When adding to the preceding equality the relation

$$\partial_t \varrho \Phi(\mathcal{T}) + \partial_x \varrho \Phi(\mathcal{T}) u = \mu \varrho \Phi'(\mathcal{T})(\tau - \mathcal{T}),$$

the right-hand side  $\mu \varrho \left( \frac{\Pi}{c^2} (p'(\mathcal{T}) + c^2) + \Phi'(\mathcal{T}) \right) (\tau - \mathcal{T})$  should have a definite sign. Setting  $\tilde{\varepsilon}(\tau) = \varepsilon(\varrho)$  which satisfies  $\tilde{\varepsilon}'(\tau) = -\tilde{p}(\tau)$ , the first condition gives

$$\Phi(\tau) = \tilde{\varepsilon}(\tau) - \tilde{p}^2(\tau)/2c^2.$$

Then with this definition of  $\Phi$ , we have  $\Phi'(\mathcal{T}) = -\tilde{p}(\mathcal{T}) - \tilde{p}\tilde{p}'(\mathcal{T})/c^2$ , and, with (8.28), the resulting right-hand side term writes

$$\left( \frac{\Pi}{c^2} (p'(\mathcal{T}) + c^2) + \Phi'(\mathcal{T}) \right) (\tau - \mathcal{T}) = -(p'(\mathcal{T}) + c^2)(\tau - \mathcal{T})^2$$

which is indeed negative under (8.30), proving the result. The expression in terms of  $(\varrho, u, \Pi)$  follows with  $\varphi \equiv \Phi \circ h^{-1}$ , with the previously defined function  $h(s) = \tilde{p}(s) + c^2 s$ .  $\square$

At equilibrium, we have  $\Pi = p$  and  $\mathcal{E}(\varrho, u, p) = \eta$ , and we now prove that the following minimization principle holds:

$$\eta(\varrho, u) \leq \mathcal{E}(\varrho, u, \Pi).$$

We first study the convexity of  $\mathcal{E} = \varrho \Sigma$  wrt. to  $\mathbf{U} = (\varrho, \varrho u, \varrho \mathcal{T})$  or equivalently (as results from the preliminaries in Chap. III) the convexity of the mapping  $(\tau, u, \mathcal{T}) \mapsto \Sigma = \tilde{\varepsilon}(\mathcal{T}) + \frac{u^2}{2} + \frac{1}{2c^2} (\tilde{\Pi}^2(\tau, \mathcal{T}) - \tilde{p}^2(\mathcal{T}))$ . We have

$$\partial_{\tau\tau}(\Sigma) = c^2, \quad \partial_{\mathcal{T}\mathcal{T}}(\Sigma) = (\tilde{p}' + c^2) + \tilde{p}''(\mathcal{T} - \tau), \quad \partial_{\tau\mathcal{T}}(\Sigma) = -\tilde{p}' - c^2,$$

and the determinant of the Hessian matrix is  $-\tilde{p}'(\tilde{p}' + c^2) + c^2\tilde{p}''(\mathcal{T} - \tau)$  which may become negative if  $\tau - \mathcal{T}$  increases. Hence  $\Sigma$  is not convex on the whole domain; however, for the approximation results, we are interested in the behavior in a neighborhood of the equilibrium manifold, and we have a minimization principle which will be important for obtaining entropy properties of our numerical scheme: the maximum dissipation of entropy is attained for equilibrium states.

*Proposition 8.3*

For a given  $\mathbf{U} = (\varrho, \varrho u, \varrho \mathcal{T})$ , we note  $\mathbf{U}^{eq} = (\varrho, \varrho u, 1) = (\mathbf{u}, 1)$ . We have

$$\eta(\mathbf{u}) = \mathcal{E}(\mathbf{U}^{eq}) = \min_{\mathcal{T} \in K} \mathcal{E}(\mathbf{U}). \quad (8.34)$$

Here,  $K$  is a compact interval so that  $\varrho^{-1} = \tau \in K$ , and we consider a finite value of  $c$  satisfying  $c^2 > \max_{\tau \in K}(-\tilde{p}'(\tau))$ .

*Proof.* For fixed  $(\varrho, \varrho u)$ , we take the partial derivative wrt.  $\mathcal{T}$  of  $\mathcal{E} = \varrho \Sigma(\mathbf{U})$  given by (8.32)

$$\varrho \Sigma = \varrho \tilde{\varepsilon}(\mathcal{T}) + \varrho \frac{u^2}{2} + \varrho \frac{1}{2c^2}(\tilde{\Pi}^2(\tau, \mathcal{T}) - \tilde{p}^2(\mathcal{T})).$$

Since  $\varrho$  is fixed, we compute equivalently  $\partial_{\mathcal{T}} \Sigma(\tau, u, \mathcal{T})$

$$\partial_{\mathcal{T}} \Sigma(\tau, u, \mathcal{T}) = -\tilde{p}(\mathcal{T}) + \frac{1}{c^2}(\tilde{\Pi} \partial_{\mathcal{T}} \tilde{\Pi}(\tau, \mathcal{T}) - \tilde{p} \tilde{p}'(\mathcal{T})),$$

and from the definition of  $\tilde{\Pi}$ , we get easily

$$\partial_{\mathcal{T}} \Sigma(\tau, u, \mathcal{T}) = \frac{1}{c^2}(\tilde{p}'(\mathcal{T}) + c^2)(\tilde{\Pi} - \tilde{p}) = (\tilde{p}'(\mathcal{T}) + c^2)(\mathcal{T} - \tau),$$

and the right-hand side vanishes only for  $\mathcal{T} = \tau$ , i.e.,  $\Sigma$  has its only possible extremum at the equilibrium state. We compute as above  $\partial_{\mathcal{T}\mathcal{T}} \Sigma = \tilde{p}''(\mathcal{T})(\mathcal{T} - \tau) + \tilde{p}'(\mathcal{T}) + c^2$  and at this equilibrium state  $\partial_{\mathcal{T}\mathcal{T}} \Sigma(\mathbf{U}^{eq}) = \tilde{p}'(\tau) + c^2 > 0$  so that it is indeed a minimum.  $\square$

Let us give a complementary remark on the stability of the relaxation process. When  $\mu \rightarrow \infty$ , we expect the equilibrium relation  $\mathcal{T} = \tau$  and thus  $\tilde{\Pi} = \tilde{p}(\tau)$ , relation which, once substituted in the second equation of (8.27), results in the *equilibrium* momentum conservation law

$$\frac{\partial}{\partial t} \varrho u + \frac{\partial}{\partial x}(\varrho u^2 + p) = 0.$$

Let us consider a Chapman–Enskog-type expansion and in that aim introduce a first-order correction term for  $\mathcal{T}$  in inverse powers of  $\mu$ :

$$\mathcal{T}^{(\mu)} = \tau^{(\mu)} + \mu^{-1} \tau_1^{(\mu)} + \mathcal{O}(\mu^{-2}),$$

where we now emphasize the dependence on  $\mu$  of the solution of (8.27). The second equation of (8.27) yields, replacing  $\Pi$  by its expression (8.28) and  $\tilde{p}(\mathcal{T})$  by the expansion  $\tilde{p}(\mathcal{T}) = \tilde{p}(\tau) + \mu^{-1}\tau_1\tilde{p}'(\tau) + \mathcal{O}(\mu^{-2})$ ,

$$\partial_t \varrho^{(\mu)} u^{(\mu)} + \partial_x (\varrho^{(\mu)} (u^{(\mu)})^2 + \tilde{p}(\tau^{(\mu)})) = -\mu^{-1} \partial_x ((\tilde{p}'(\tau^{(\mu)}) + c^2) \tau_1^{(\mu)}) + \mathcal{O}(\mu^{-2})$$

and the third equation of (8.27) gives

$$\partial_t (\varrho^{(\mu)} \tau^{(\mu)}) + \partial_x (\varrho^{(\mu)} \tau^{(\mu)} u^{(\mu)}) = -\varrho^{(\mu)} \tau_1^{(\mu)} + \mathcal{O}(\mu^{-1}),$$

and since  $\varrho^{(\mu)} \tau^{(\mu)} = 1$ , we get, neglecting the  $\mathcal{O}(\mu^{-2})$  term and dropping the superscript  $\mu$ ,

$$\partial_t \varrho u + \partial_x (\varrho u^2 + \tilde{p}(\tau)) = \mu^{-1} \partial_x ((\tilde{p}'(\tau) + c^2) \tau \partial_x u).$$

This last equation is a dissipative approximation of the momentum conservation equation if the stability or Whitham criterion (8.29) holds.

### 8.3.3 The Relaxation Solver for the Isentropic Euler System

The exact Riemann solver for the relaxation system (8.27) provides a numerical scheme for the Euler system (8.24), which is a relaxation scheme. We detail the presentation of the scheme with the relaxation system; it is, however, important to keep in mind that the resulting scheme does not need any computation of the auxiliary variables introduced in the relaxation system (as already said). In particular, there is in fact no need in the algorithm of the relaxation source term since the data are set at equilibrium. These elements are kept in order to understand the underlying interpretation which makes the scheme a stable and consistent scheme for approximating the Euler system.

We sketch the main lines of the procedure to advance the solution in time from  $t_n$  to  $t_{n+1} = t_n + \Delta t$ . We follow the three steps (already mentioned) needed in the relaxation solver: reconstruction, evolution, and projection; the last two steps correspond to solve the relaxation system (8.27) by an operator splitting method. Starting from  $\mathbf{u}_0(x) = (\varrho_0, \varrho_0 u_0)^T(x)$ , an initial datum for system (8.24), we first associate the usual piecewise constant function  $\mathbf{u}_\Delta^0$  equal to  $\mathbf{u}_j^0 = \frac{1}{\Delta x} \int_{C_j} \mathbf{u}_0(x) dx$  on cell  $C_j = (x_j - \Delta x/2, x_j + \Delta x/2)$ . We describe the updating procedure for  $n = 0$ , since it is similar at any other time.

1. Let  $(\mathbf{u}_j^0) = (\varrho_j^0, \varrho_j^0 u_j^0)^T$  be an initial discretized datum. Define the extended initial datum at equilibrium  $\mathbf{U}_j^0 = (\varrho_j^0, \varrho_j^0 u_j^0, \varrho_j^0 \mathcal{T}_j^0)^T$  for system (8.27) with  $\mathcal{T}_0 \equiv 1/\varrho_0$  and the associated piecewise constant function  $\mathbf{U}_\Delta^0$ .
2. With the initial data  $\mathbf{U}_\Delta^0$ , solve exactly the Cauchy problem

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) = 0, \\ \partial_t(\rho u) + \partial_x(\rho u^2 + \Pi) = 0, \quad t \in (0, \Delta t] \\ \partial_t(\rho \mathcal{T}) + \partial_x(\rho \mathcal{T} u) = 0, \end{cases} \quad (8.35)$$

with (8.28):  $\Pi = \Pi(\varrho, \rho \mathcal{T}) = \tilde{\Pi}(\tau, \mathcal{T})$ , we obtain  $\mathbf{U}_1^-(x) = \mathbf{U}(x, \Delta t)$ .

3. Project  $\mathbf{U}_1^- = (\varrho_1^-, \varrho_1 u_1^-, \varrho_1 \mathcal{T}_1^-)^T$  on the equilibrium manifold of system (8.27) to get  $\mathbf{U}_1 = (\varrho_1, \varrho_1 u_1, 1)^T$ , and define  $\mathbf{u}_1(x) = (\varrho_1, \varrho_1 u_1)^T(x)$  and  $\mathbf{u}_j^1 = \frac{1}{\Delta x} \int_{C_j} \mathbf{u}_1(x) dx$ .

System (8.27) is a system with source term, say  $\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = \mu \mathcal{R}(\mathbf{U})$ . In step 2, we have to solve a juxtaposition of Riemann problems, which means that we use Godunov's method for (8.35), the homogenous part of (8.27)  $\partial_t \mathbf{U} + \partial_x \mathbf{F}(\mathbf{U}) = \mathbf{0}$ , while step 3 corresponds to solve the source term, say  $\partial_t \mathbf{U} = \mu \mathcal{R}(\mathbf{U})$ , with  $\mu \rightarrow \infty$  which gives  $\mathcal{R}(\mathbf{U}) = \mathbf{0}$  (this step may also be called instantaneous relaxation); this explains the fact that we have spoken of using a splitting method to solve (8.27).

*Proposition 8.4*

The resulting scheme can be written as

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda \left( \mathbf{g}_{j+1/2}^n - \mathbf{g}_{j-1/2}^n \right), \quad j \in \mathbb{Z}, n \geq 0, \quad (8.36)$$

with the two components of the flux  $\mathbf{g}_{j+1/2}^n$  given by  $(\varrho u, \varrho u^2 + \Pi)(\mathbf{W}_{j+1/2}^n)$  i.e., evaluated on state  $\mathbf{W}_{j+1/2}^n = \mathbf{W}_R(0+; \mathbf{U}_j^n, \mathbf{U}_{j+1}^n)$  solution of the Riemann problem for (8.35). Moreover, the relaxation solver satisfies a discrete entropy inequality

$$\eta(\mathbf{u}_j^{n+1}) \leq \eta(\mathbf{u}_j^n) - \lambda((\mathcal{H}u)_{j+1/2}^n - (\mathcal{H}u)_{j-1/2}^n), \quad j \in \mathbb{Z}, n \geq 0, \quad (8.37)$$

with  $\eta, \mathcal{H}$  given in (8.31) (8.32) and  $\mathcal{H}_{j+1/2}^n = \mathcal{H}(\mathbf{W}_{j+1/2}^n)$ .

Recall that  $\mathbf{W}_R(.; \mathbf{U}_L, \mathbf{U}_R)$  may be discontinuous at the interface, but the flux is continuous.

*Proof.* The first statement comes from the conservation form of Godunov's scheme for (8.35), from which we just keep the two first components. Now, the entropy solution of the Riemann problem for (8.35) satisfies an entropy equality relative to the extended entropy  $\mathcal{E}$ , because all the fields are linearly degenerate. Then, we know that for Godunov's scheme, the formula for the updated value can be obtained by integrating this equation on a mesh cell  $C_j \times (t_n, t_{n+1})$  under CFL 1/2. This leads to

$$\mathcal{E}(\mathbf{U}_j^{n+1}) = \mathcal{E}(\mathbf{U}_j^n) - \lambda((\mathcal{H}u)_{j+1/2}^n - (\mathcal{H}u)_{j-1/2}^n). \quad (8.38)$$

Then  $\mathcal{E}(\mathbf{U}_j^n) = \eta(\mathbf{u}_j^n)$  because the state is at equilibrium. Eventually we get an inequality in the projection step, because of the minimization principle (Proposition 8.3) which yields  $\eta(\mathbf{u}_j^{n+1}) \leq \mathcal{E}(\mathbf{U}_j^{n+1})$ .  $\square$

### 8.3.4 The Relaxation Solver for the Euler System

For the full Euler system with energy, a relaxation system can be introduced similarly, and a relaxation scheme can be derived from this approach. There is, however, a trick which consists in working first with the entropy equation, instead of the energy one. If we come back to the Lagrangian frame, which leads to simpler computations, completing (8.20) with an energy equation, the equilibrium system writes

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x p = 0, \\ \partial_t e + \partial_x p u = 0, \end{cases} \quad (8.39)$$

with  $p = p(\tau, \varepsilon)$ ,  $e = \frac{1}{2}u^2 + \varepsilon$ ; we set  $\mathbf{u} = (\tau, u, e)^T$ . For smooth solutions, the entropy equation writes  $\partial_t s = 0$ , and (8.39) is equivalent to

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x p = 0, \\ \partial_t s = 0. \end{cases} \quad (8.40)$$

where we consider now the pressure  $p$  as  $p = \tilde{p}(\tau, s)$  and similarly  $\varepsilon = \varepsilon(\tau, s)$ ; recall that  $\partial_\tau \varepsilon = -p$ . The relaxation system is given by (8.22) to which we add the entropy equation

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x \Pi = 0, \\ \partial_t s = 0, \\ \partial_t \mathcal{T} = \mu(\tau - \mathcal{T}), \end{cases} \quad (8.41)$$

with the naturally extended closure law

$$\Pi = \Pi(\tau, s, \mathcal{T}) = \tilde{p}(\mathcal{T}, s) + c^2(\mathcal{T} - \tau). \quad (8.42)$$

We now set  $\mathbf{U} = (\tau, u, s, \mathcal{T})^T$ . Since  $s$  remains constant in time, the only wave across which  $s$  can be discontinuous is stationary, and all the previous approach concerning the isentropic case is easily extended. There is an “entropy” which is simply the augmented energy (8.32)

$$\Sigma(\tau, u, \mathcal{T}, s) = \varepsilon(\mathcal{T}, s) + \frac{u^2}{2} + \frac{1}{2c^2}(\Pi^2(\tau, s, \mathcal{T}) - \tilde{p}^2(\mathcal{T}, s)), \quad (8.43)$$

with the entropy (energy) flux which is  $\Pi u$  since we are in a Lagrangian frame. Note that on equilibria,  $\mathcal{T} = \tau$ ,  $\Pi = p$ , and thus  $\Sigma$  coincides with the usual energy on the equilibrium manifold  $\Sigma = e$ ; moreover, we can prove that a minimization principle holds. The Riemann problem is also easily solved in this set of variables just adding to the solution, in the isentropic case, a “parameter,” corresponding to the entropy kept constant, on each

side of the interface,  $s_L$ , for  $x < 0$ , and  $s_R$  for  $x > 0$ . Working with the entropy equation is not a problem when one is interested in approximation results which concern essentially smooth solutions. However, for a numerical scheme, since one wants to compute discontinuities satisfying the right jump condition, hence traveling at the right speed, it is important to derive a scheme which is conservative wrt. the energy conservation equation and is consistent with an entropy dissipation inequality. This can be done in a rather straightforward way. The three steps of the numerical scheme described in the isentropic case are essentially the same and must be completed by a step which turns back to a conservation equation on the energy and in such a way that a discrete entropy inequality holds.

*Proposition 8.5*

Under CFL  $1/2$ , the resulting global relaxation scheme can be written as

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda \left( \mathbf{g}_{j+1/2}^n - \mathbf{g}_{j-1/2}^n \right), \quad j \in \mathbb{Z}, n \geq 0, \quad (8.44)$$

with the three components of the flux  $\mathbf{g}_{j+1/2}^n$  given by  $(-u, \Pi, \Pi u)(\mathbf{W}_{j+1/2}^n)$  (i.e., evaluated on state  $\mathbf{W}_{j+1/2}^n$ ) where

$$\mathbf{W}_{j+1/2}^n = \mathbf{W}_R(0+; \mathbf{U}_j^n, \mathbf{U}_{j+1}^n),$$

and where  $\mathbf{W}_R$  denotes the solution of the Riemann problem for the homogeneous part of (8.41). Moreover, the global relaxation solver satisfies a discrete entropy inequality

$$S_j^{n+1} \leq S_j^n, \quad j \in \mathbb{Z}, n \geq 0, \quad (8.45)$$

where  $S_j^{n+1} = -s(\mathbf{u}_j^{n+1})$ .

Note that in (8.45),  $S_j^{n+1} = -s(\mathbf{u}_j^{n+1})$  is the mathematical entropy ( $s$  is the physical entropy) of the updated state.

*Proof.* Let us denote by  $\mathbf{U}(x, t_{n+1}-)$  the solution obtained at the end of the evolution step; it concerns the variable  $\mathbf{U} = (\tau, u, s, \mathcal{T})^T$ , which means we have two equations in  $(\tau, u)$ , which remain unchanged in the projection step and give a scheme in conservation form for  $(\tau_j^{n+1}, u_j^{n+1})$ , with (consistent) numerical flux denoted, respectively, by  $(g_{\tau, j+1/2}^n, g_{u, j+1/2}^n)$ ; these fluxes given as expected by the exact flux  $-u, \Pi$  evaluated on  $\mathbf{W}_R(0+; \mathbf{U}_j^n, \mathbf{U}_{j+1}^n)$ . Then we have a third equation for the entropy, with numerical flux denoted by  $g_{S, j+1/2}^n$  (which happens to be  $= 0$  in Lagrangian coordinates; in Eulerian coordinates, it is not null)

$$S_j^{n+1-} = S_j^n - \lambda(g_{S, j+1/2}^n - g_{S, j-1/2}^n). \quad (8.46)$$

Now, we have to define the updated energy. Since all the fields are linearly degenerate, we have also an energy conservation equation satisfied by the solution of the Riemann problem

$$\partial_t \Sigma + \partial_x \Pi u = 0,$$

and integrating this equation on a mesh cell  $C_j \times (t_n, t_{n+1})$ , under CFL 1/2, and setting  $\Sigma_j^{n+1-} \equiv \frac{1}{\Delta x} \int_{C_j} \Sigma(\mathbf{U}(x, t_{n+1}-)) dx$ , we have

$$\Sigma_j^{n+1-} = \Sigma_j^n - \lambda(\Pi u_{j+1/2}^n - \Pi u_{j-1/2}^n), \quad (8.47)$$

where  $\Pi u_{j+1/2}$  is given by

$$(\Pi u)_{j+1/2}^n \equiv (\Pi u)(\mathbf{W}(0\pm; \mathbf{U}_j^n, \mathbf{U}_{j+1}^n)). \quad (8.48)$$

We define the final updated energy, i.e.,  $e_j^{n+1}$ , by  $e_j^{n+1} \equiv \Sigma_j^{n+1-}$  equivalently the right-hand side of (8.47) which writes (since the data  $\mathbf{U}_j^n$  is at equilibrium)

$$e_j^{n+1} = e_j^n - \lambda(\Pi u_{j+1/2}^n - \Pi u_{j-1/2}^n), \quad (8.49)$$

and we note that the discrete energy flux is consistent with the continuous one. We have thus written a scheme in conservation form for the variable  $\mathbf{u} = (\tau, u, e)^T$ , with respective numerical fluxes  $(g_{\tau,j+1/2}^n, g_{u,j+1/2}^n, \Pi u_{j+1/2}^n)^T$ . Now, we want to prove some discrete entropy inequality. For the equilibrium projection step, assume for a while that the projection on the equilibrium manifold is done pointwise, i.e., define  $\mathbf{U}^{eq}(x, t_{n+1}^-) = (\tau, u, s, \mathcal{T})^T(x, t_{n+1}^-)$  by

$$\begin{cases} (\tau, u, s)(x, t_{n+1}^-) = (\tau, u, s)(x, t_{n+1}-), \\ \mathcal{T}(x, t_{n+1}^-) = \tau(x, t_{n+1}-), \end{cases}$$

where, as previously, the value  $\varphi(., t_{n+1}-)$  of a quantity  $\varphi$  denotes the value at the end of the evolution step. We use the minimization principle which says that the maximal dissipation of entropy is attained for equilibrium states

$$\Sigma(\mathbf{U}^{eq}) = \min_{\mathcal{T} \in K} \Sigma(\mathbf{U}), \quad \text{if } \mathbf{U}^{eq} = (\tau, u, s, \mathcal{T})^T, \mathbf{U} = (\tau, u, s, \mathcal{T})^T$$

( $K$  is a compact set such that it contains all possible values of  $\mathcal{T}$  obtained at the evolution step). After the pointwise relaxation, we thus deduce

$$\frac{1}{\Delta x} \int_{C_j} \Sigma(\mathbf{U}(x, t_{n+1}^-)) dx \leq \frac{1}{\Delta x} \int_{C_j} \Sigma(\mathbf{U}(x, t_{n+1}-)) dx = e_j^{n+1}, \quad (8.50)$$

the last equation coming from the definition (8.48) of  $e_j^{n+1}$ . Now, we also have for a state at equilibrium

$$\Sigma(\mathbf{U}^{eq}(x, t_{n+1}^-)) = e(\mathbf{U}(x, t_{n+1}^-)). \quad (8.51)$$

We notice that the three first components  $(\tau, u, s)$  are kept unchanged during the equilibrium projection step, whether pointwise or globally defined. Now, since the function  $(u, \tau, s) \mapsto e$  is convex and since  $e(x, t_{n+1}-) = e((\tau, u, s)(x, t_{n+1}-))$ , we can apply Jensen's inequality and get, together with (8.50) and (8.51),

$$e_j^{n+1-} \equiv e(u_j^{n+1}, \tau_j^{n+1}, s_j^{n+1-}) \leq \frac{1}{\Delta x} \int_{C_j} e(x, t_{n+1}-) dx \leq e_j^{n+1}.$$

Recall that since  $e = \varepsilon(\tau, s) + u^2/2$ , and  $e_s = \partial_s \varepsilon = T > 0$ , we get that  $S = -s$  is decreasing wrt. the variable  $e$ . Then, let us define

$$S_j^{n+1} = S(\mathbf{u}_j^{n+1}) = -s(\tau_j^{n+1}, u_j^{n+1}, e_j^{n+1})$$

which is the definition of the (mathematical) entropy of state  $\mathbf{u}_j^{n+1}$ ; from  $e_j^{n+1-} \leq e_j^{n+1}$ , we deduce

$$S_j^{n+1} \leq S(\tau_j^{n+1}, u_j^{n+1}, e_j^{n+1-}) = S_j^{n+1-},$$

and with (8.46), we get

$$S_j^{n+1} \leq S_j^n - \lambda(g_{S,j+1/2}^n - g_{S,j-1/2}^n),$$

and recall that  $g_S = 0$  (in Lagrangian coordinates); thus we have indeed obtained (8.45). Note that in (8.45),  $S_j^{n+1} = -s(\mathbf{u}_j^{n+1})$  is indeed the (mathematical) entropy of the updated state and not the third component  $S_j^{n+1-}$  of the updated value after the projection.  $\square$

As already said, the resulting scheme can be written in an explicit way with formula involving the exact solution of the Riemann problem, as detailed in (8.30). We get a Godunov-type scheme associated with a simple Riemann solver. There is no need to understand the scheme as a relaxation scheme to prove its properties, and we first encountered this scheme as a simple scheme (see Sect. 4.5.2, Remark 4.9). However, we think that looking at the scheme from a relaxation point of view is interesting and provides a good understanding of why the stability properties hold. It can also lead to many generalizations; in particular, this approach has been extended to general fluid systems in Eulerian or Lagrangian coordinates in [349].

Let us also mention [175] where the scheme is slightly modified so as to satisfy a semi-discrete entropy inequality under a less restrictive condition which allows to ask other properties such as the exact resolution of stationary shocks or [265] for the exact resolution of isolated shocks.

Then, we have restricted ourselves to describe in detail the scheme for the Euler system, but relaxation schemes are developed for many applications: multicomponent flow [407], two-phase flow [92, 348], phase transition [266], and MHD [172, 173, 349].

## Notes

In this chapter, we have chosen to detail some of the finite volume schemes that are the most often mentioned and used in the applications to computational fluid dynamics, as well as some interesting variants or extensions of these methods. But we do not pretend to have an exhaustive survey of the field, and we precise that more references will be given in the next chapter, in particular concerning applications other than gas dynamics. Apart from the kinetic and relaxation schemes, they were already presented in the scalar case (and we refer to G.R., Chapters 3 and 4 [539]). We must nevertheless mention some pioneering papers which were not yet cited in the present volume: Harten [589, 590, 590–592]; LeVeque with Goodman [549, 778]; Osher [917, 918], with Chakravarthy [920], Sanders [921], Solomon [922], Sweby [923], and Tadmor [924]; Roe [979, 987]; Sanders [1005]; Shu [1054, 1055]; Tadmor [1089]; and van Leer [1148–1151, 1153–1156, 1158] and also [1159].

Then, there are many interesting papers in several dedicated proceedings of conferences or summer schools (for instance, [75, 232, 320, 430, 532], etc.).

However, we could not develop all of the previously mentioned ones such as modified Lax–Wendroff [842],  $S_\beta^\alpha$  schemes (Lerat [761], Peyret and Taylor [951], Hirsch [617]), F.C.T. (Kunhardt and Wu [721], Salari and Steinberg [1003], and the references therein; see [155] for the first paper and [724] for a more recent overview); quasi-monotone, PPM (Colella and Woodward [332]); ENO(essentially nonoscillatory) (Shu and Osher [1058, 1059], Yang [1200]); and staggered mesh (Nessyahu and Tadmor [897], Sanders and Weiser [1009], Huynh [643]). Neither do we explore more deeply discrete one-sided Lipschitz condition [194, 1088], numerical entropy for MUSCL-type methods [166], and the influence of data extrema [203]; nor introduce different methods such as local extremum diminishing (Kim and Jameson [696]), compact schemes (Cockburn and Shu [322]), and random choice methods (see Glimm [530], Chorin [303], and also [326, 621, 833]), in spite of their importance, especially for Glimm’s method; nor study the extension to implicit (Lerat [763], Yee [1206], Yee et al. [1209], Poinsot and Candel [955], Yee et al. [1207], Mulder and van Leer [879], Liang and Chan [799], Blunt and Rubin [152], Collins et al. [333]) and front- or shock-tracking methods (Chern et al. [298], Mao [847], Charrier and Tessieras [282], Henshaw [605], Davis [394], LeVeque and Shyue [780, 781], and the references therein), nonconservative schemes (Raviart and Sainsaulieu [968], choosing primitive formulations (Karni [679, 680], Harabetian and Pego [588], Kumbaro [720], Cauret et al. [252], Adamczewski et al. [21], Berger and Colombeau [117]), and the problem of convergence of such schemes [631, 657, 754] and approximation of non-strictly hyperbolic systems (Freistühler and Pitman [492], Tveito and Winther [1142]) or systems of mixed type (Stewart and Wendroff [1073], Shu [1056]). Except in Sects. 2 and 6, we have mainly considered the application to the equations of gas dynamics in Eulerian coordinates; we might mention some specific problems, accuracy at a contact discontinuity [314, 418, 433, 434],

spurious solutions [962], and slowly moving shocks [664, 977, 1074]; for a Lagrangian approach, see Munz [883] and for detonation waves and reactive flows LeVeque and Shyue [780], Hilditch and Colella [616], and [330]. Concerning systems with source terms (Roe [983], Leveque and Yee [786], Sweby [1080], Mulder and van Leer [879], Glaister [525], Mehlman [860], Böing et al. [154], Chalabi [259, 260], Griffiths et al. [561], Bermudez and Vasquez [119], Klingenstein [704], Greenberg and LeRoux [554, 560], Fey et al. [475], Schroll and Winther [1025], LeVeque [776]) or relaxation (S. Jin [658], Jin and Xin [668], Bereux and Sainsaulieu [115, 116, 1024]), they are now given a whole chapter (Chap. VII). For boundary conditions, a brief survey will be given in Chap. VI. We did not take up the study of theoretical aspects of schemes such as the validity of the modified equation [550], nonlinear stability of discrete shocks (Liu and Xin [820] and the references therein), viscous perturbations (see Harabetian [585–587]), or convergence and error estimates [766–768, 782, 814, 833, 898, 899, 1081, 1104]; in fact, concerning systems, few results are available (Tveito and Winther [1142], Chen and Liu [292], Chen and LeFloch [290], Isaacson and Temple [647]). We did not either go on to the effective implementation (Ajmani et al. [24], for instance).

We mention other main references for the kinetic theory, P.-L. Lions [809], Perthame and Pulvirenti [945], and the references therein for the B.G.K. model; Lions, Perthame, and Tadmor [813] for a kinetic formulation of the isentropic gas dynamics; and [1187] for its application on convergence proofs and for  $p$ -systems Bardos [84, 85] and Whitham Section 6.3 [1188]. We may also mention another example concerning the Broadwell model, a simplified discrete velocity model for the Boltzmann equation, and the corresponding relaxation scheme [224]. For (kinetic) schemes, see [190], Deshpande [411], Pullin [959], Kaniel [674], Harten et al. [595], and Mandal and Deshpande [845, 846]; see also Coron and Perthame [363], Reitz [972], Macrossan [837], Prendergast and Xu [958, 1199], Yang and Huang [1202], Xu et al. [1198], Croisille [374], De Vuyst [396], Villedieu and coauthors [463, 1171], and also [296]. Several recent papers provide a comparative study of some shock-capturing schemes (Sod [1068], Montagné et al. [873], Chargy et al. [280], Roberts [977], Yang and Przekwas [1201], Quirk (1992), Menne et al. [864], Hannapel et al. [581], Lin [802], and Rider [975]) for methods in Lagrangian coordinates. We have tried to present the underlying theory, computations, and results as simply as possible and to mention the main contributors to the subject. We shall not list all of the references here, especially since it is impossible to quote all the authors who have contributed to recent developments in the field. We just quote some important contributions that have not yet been given: first of all the very complete text of Hirsch [617] and also Roache [976] and many interesting papers such as Colella [327], Hall [577], LeVeque [773], and Toro [1124, 1126].

*Note Added in the Second Edition*

Since the first edition of this book, there has been a tremendous research effort in the domain, leading to a huge quantity of papers, several textbooks, and many proceedings of international conferences dedicated to the subject, and it is possible neither to take into account nor even to cite all the contributions. For instance, interesting results may be obtained by the (wave) front-tracking technique first proposed by Dafermos [378] and then generalized and used numerically in [622]; see [335], and the references therein, or the textbooks by Bressan [198] and the more recent [626] by Holden and Risebro.

Concerning kinetic and relaxation approximation, a few of them can be found in the related sections. We have now added a specific chapter for the treatment of source terms where relaxation will be again treated.

Other theoretical results concern discrete shock profiles or (in)stability [72, 201, 354, 365, 1037, 1042], reconstruction schemes [727], and convergence results [123, 124, 741]; we will give more references in Chap. V, also for error estimates.

Concerning schemes for ideal MHD, let us also mention a few more references, many of them linked to HLL-type, Roe-type, or Osher–Solomon scheme: [76, 77, 146, 235, 236, 242, 250, 376, 446, 497, 569, 649, 792, 957, 989, 1030, 1128], etc.

Other applications, such as traffic flow modeling, have received a lot of attention; see, for instance, [126, 273] and references therein. More references are given in Chap. VII.

A new topic has received some interest: a posteriori error estimate [555] (see a few other references on the subject at the end of Chap. V).

Then there are many added refinements, new approaches, etc.; it is out of scope to cite them all. Let us just mention some recent approaches aiming at taking into account smaller-scale viscosity or dissipation effects [462, 871]; in another direction, staggered grids have received recently more interest [125, 608] and then uncertainty propagation [1135, 1136] and uncertainty quantification [13, 148, 954].



# V

## The Case of Multidimensional Systems

### 1 Generalities on Multidimensional Hyperbolic Systems

#### 1.1 Definitions

We shall mainly consider in this section a two-dimensional  $p \times p$  hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) + \frac{\partial}{\partial y} \mathbf{g}(\mathbf{u}) = \mathbf{0}. \quad (1.1)$$

We shall not give much detail on the theoretical aspects of the problem, which are not fully understood and are more complex than in the scalar case  $p = 1$  or in the one-dimensional case. Let us just recall the definition of hyperbolicity (see the Introduction, Sect. 1) for a system in dimension  $d$ , which can be written as

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) = \mathbf{0}, \quad (1.2a)$$

where  $\mathbf{u} = (u_1, \dots, u_p)^T \in \Omega \subset \mathbb{R}^p$ ;  $\mathbf{f}_j(\mathbf{u}) = (f_{1j}(\mathbf{u}), \dots, f_{pj}(\mathbf{u}))^T \in \mathbb{R}^p$ ,  $j = 1, \dots, d$ ; and  $\mathbf{x} \in \mathbb{R}^d$ ,  $t \in \mathbb{R}^+$ . We shall also use the more compact form

$$\frac{\partial \mathbf{u}}{\partial t} + \operatorname{div} \mathbb{F}(\mathbf{u}) = \mathbf{0}, \quad \mathbb{F} = (\mathbf{f}_1, \dots, \mathbf{f}_d)^T. \quad (1.2b)$$

We define for  $\mathbf{u} = (u_1, \dots, u_p)^T \in \Omega$  and  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_d)^T \in \mathbb{R}^d$  the  $p \times p$  matrix  $\mathbf{A}(\mathbf{u}, \boldsymbol{\omega})$  “in the direction  $\boldsymbol{\omega}$ ” by

$$\mathbf{A}(\mathbf{u}, \boldsymbol{\omega}) = \sum_{j=1}^d \mathbf{f}'_j(\mathbf{u}) \omega_j, \quad (1.3)$$

where  $\mathbf{f}'_j(u) = (\partial f_{ij}/\partial u_k)_{1 \leq i, k \leq p}$  is the Jacobian matrix of  $\mathbf{f}_j(\mathbf{u})$ . With short-hand notation, we shall also write

$$\mathbf{A}(\mathbf{u}, \boldsymbol{\omega}) = \mathbb{F}'(\mathbf{u}) \cdot \boldsymbol{\omega}.$$

*Definition 1.1*

The system (1.2) is called *hyperbolic* if for any  $\mathbf{u} \in \Omega$  and any direction  $\boldsymbol{\omega} \in \mathbb{R}^d, |\boldsymbol{\omega}| = 1$ , the matrix  $\mathbf{A}(\mathbf{u}, \boldsymbol{\omega})$  defined by (1.3) has  $p$  real eigenvalues

$$\lambda_1(\mathbf{u}, \boldsymbol{\omega}) \leq \lambda_2(\mathbf{u}, \boldsymbol{\omega}) \leq \cdots \leq \lambda_p(\mathbf{u}, \boldsymbol{\omega}),$$

with a complete family of (right) eigenvectors  $\mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega})$ ,

$$\mathbf{A}(\mathbf{u}, \boldsymbol{\omega})\mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}) = \lambda_k(\mathbf{u}, \boldsymbol{\omega})\mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}).$$

Moreover, the system is *strictly hyperbolic* at a state  $\mathbf{u}$  if the eigenvalues are distinct.

The notions of *genuine nonlinearity* ( $\forall \mathbf{u}, D_{\mathbf{u}}\lambda_k(\mathbf{u}, \boldsymbol{\omega}) \cdot \mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}) \neq 0$ ) and *linear degeneracy* ( $\forall \mathbf{u}, D_{\mathbf{u}}\lambda_k(\mathbf{u}, \boldsymbol{\omega}) \cdot \mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}) = 0$ ) can then be defined for the  $k$ th field  $\lambda_k(\mathbf{u}, \boldsymbol{\omega})$  in the direction  $\boldsymbol{\omega}$ , as well as the  *$k$ -Riemann invariants* (i.e., a function  $w : \mathbb{R}^p \rightarrow \mathbb{R}$  such that  $Dw(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}) = 0$ ). In particular, if the  $k$ th field is linearly degenerate,  $\lambda_k(\cdot, \boldsymbol{\omega})$  is a  *$k$ -Riemann invariant*.

We shall also denote as usual the eigenvectors of  $\mathbf{A}(\mathbf{u}, \boldsymbol{\omega})^T$  by  $\mathbf{l}_k(\mathbf{u}, \boldsymbol{\omega})$  and normalize the vectors as in Chap. II, (2.53), (2.54) (in particular,  $D_{\mathbf{u}}\lambda_k(\mathbf{u}, \boldsymbol{\omega}) \cdot \mathbf{r}_k(\mathbf{u}, \boldsymbol{\omega}) = 1$  for a genuinely nonlinear field).

The system is symmetrizable if there exists a matrix  $\mathbf{S}(\mathbf{u})$ , smoothly varying with  $\mathbf{u}$ , such that the matrices  $\mathbf{S}(\mathbf{u})\mathbf{A}_j(\mathbf{u})$  where  $\mathbf{A}_j(\mathbf{u}) = \mathbf{f}'_j(u)$  are symmetric,  $j = 1, \dots, d$ . This property holds if we assume that the system admits a strictly convex entropy (Godunov and Mock's theorem, see the Chap. I, Theorem 5.1; see Harten [590] for the Euler equations, Godunov [543] for further examples).

*Example 1.1.* Ideal gas isentropic dynamics. Let us consider the system in dimension  $d = 2$

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial y}(\rho v^2 + p) = 0, \end{cases} \quad (1.4)$$

with  $p = p(\rho) > 0$  ( $p$  is the pressure). Setting  $\mathbf{U} = (\rho, u, v)^T$ , the system (1.4) is of the form (1.1). We find the eigenvalues by working with the nonconservative (primitive) variables  $\mathbf{V} = (p, u, v)^T$ ,

$$\begin{cases} \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + v \frac{\partial p}{\partial y} + \rho p' \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) = 0, \\ \rho \left( \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} \right) + \frac{\partial p}{\partial x} = 0, \\ \rho \left( \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} \right) + \frac{\partial p}{\partial y} = 0. \end{cases}$$

The matrix  $\mathbf{A}(\mathbf{V}, \boldsymbol{\omega})$  is given by

$$\mathbf{A}(\mathbf{V}, \boldsymbol{\omega}) = \begin{pmatrix} \mathbf{u} \cdot \boldsymbol{\omega} & \rho p' \omega_1 & \rho p' \omega_2 \\ \omega_1 / \rho & \mathbf{u} \cdot \boldsymbol{\omega} & 0 \\ \omega_2 / \rho & 0 & \mathbf{u} \cdot \boldsymbol{\omega} \end{pmatrix},$$

where we have set  $\mathbf{u} = (u, v)^T$ , and

$$\mathbf{u} \cdot \boldsymbol{\omega} = u\omega_1 + v\omega_2$$

denotes the scalar product of the two vectors (which we shall also write  $\mathbf{u}^T \boldsymbol{\omega}$ ). Assuming  $p'(\rho) > 0$ , the matrix  $\mathbf{A}(\mathbf{V}, \boldsymbol{\omega})$  has three real eigenvalues, which are given by

$$\lambda_0 = \mathbf{u} \cdot \boldsymbol{\omega}, \quad \lambda_{\pm} = \mathbf{u} \cdot \boldsymbol{\omega} \pm \sqrt{p'(\rho)} |\boldsymbol{\omega}| = \mathbf{u} \cdot \boldsymbol{\omega} \pm \sqrt{p'(\rho)}$$

for  $|\boldsymbol{\omega}| = 1$ , and (1.4) is thus strictly hyperbolic. We shall set

$$c^2 = p'(\rho),$$

so that  $\lambda_{\pm} = \mathbf{u} \cdot \boldsymbol{\omega} \pm c$ . Assuming more specifically  $p(\rho) = A\rho^\gamma$ ,  $\gamma > 1$ , we note that  $c^2 = p'(\rho) = \frac{\gamma p}{\rho}$ . The corresponding eigenvectors can then be chosen as

$$\mathbf{r}_0(\mathbf{V}, \boldsymbol{\omega}) = (0, -\omega_2, \omega_1)^T,$$

which is independent of  $\mathbf{u}$  and yields a linearly degenerate field since  $(-\omega_2, \omega_1)$  is orthogonal to  $\boldsymbol{\omega}$ , and

$$\mathbf{r}_{\pm}(\mathbf{V}, \boldsymbol{\omega}) = (c\rho, \pm\omega_1, \pm\omega_2)^T.$$

The associated fields are genuinely nonlinear provided  $\rho p'' + 2p' \neq 0$  since

$$D_{\mathbf{v}} \lambda_{\pm}(\mathbf{V}, \boldsymbol{\omega}) \cdot \mathbf{r}_{\pm}(\mathbf{V}, \boldsymbol{\omega}) = \pm \frac{(\rho p'' + 2p')}{2p'}$$

(which holds for a polytropic gas, in which case the expression is  $> 0$ ; see (2.41) in Chap. II).  $\square$

We shall be especially interested in the one-dimensional “projected equations” in the direction  $\boldsymbol{\omega}$

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial \zeta} (\mathbb{F} \cdot \boldsymbol{\omega})(\mathbf{u}) = \mathbf{0}, \quad (1.5a)$$

where

$$\zeta = \mathbf{x} \cdot \boldsymbol{\omega} = \sum_{j=1}^d x_j \omega_j, \quad \mathbb{F} \cdot \boldsymbol{\omega} = \sum_{j=1}^d \mathbf{f}_j(\mathbf{u}) \omega_j. \quad (1.5b)$$

Note that  $\zeta$  is the variable in the direction of  $\boldsymbol{\omega}$ , which appears, for instance, when we solve a Riemann problem where the initial data are given constant on each side of the set  $\{\zeta = 0\}$  (line in the two-dimensional case, plane in dimension  $d = 3$ ). The system (1.5a) is obtained when we look for a solution  $\mathbf{u}$  of (1.2a)–(1.2b) that depends only on  $\mathbf{x} \cdot \boldsymbol{\omega}$ ,

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(\mathbf{x} \cdot \boldsymbol{\omega}, t), \quad (1.5c)$$

where  $\mathbf{v}(\zeta, t) : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}^p$ . We find that the Jacobian is

$$(\mathbb{F} \cdot \boldsymbol{\omega})'(\mathbf{v}) = \mathbf{A}(\mathbf{v}, \boldsymbol{\omega}),$$

and the eigenvalues are again the  $\lambda_k(\cdot, \boldsymbol{\omega})$ . The system (1.2a)–(1.2b) is thus hyperbolic if the one-dimensional projected equations (1.5a)–(1.5b) are hyperbolic for any  $\boldsymbol{\omega} \in \mathbb{R}^d$ .

## 1.2 Characteristics

We first give the following definition (where  $\varphi : \mathbb{R}^d \times [0, \infty[ \rightarrow \mathbb{R}$ ).

*Definition 1.2*

A (hyper)surface  $\Sigma$  in  $\mathbb{R}^d \times [0, \infty[$  with equation  $\varphi(\mathbf{x}, t) = 0$  is characteristic for the system (1.2) at a point  $(\mathbf{x}_0, t_0)$  if the matrix

$$\frac{\partial \varphi}{\partial t} \mathbf{I} + \sum_{j=1}^d \mathbf{f}'_j(\mathbf{u}) \frac{\partial \varphi}{\partial x_j}$$

is singular at this point.

In dimension  $d = 1$ ,  $\Sigma$  is a curve; in dimension  $d = 2$ ,  $\Sigma$  is a surface. We shall say that  $\Sigma$  is characteristic if  $\Sigma$  is characteristic at each point. Setting

$$n_t = \frac{\partial \varphi}{\partial t}, \quad \boldsymbol{\nu} = \left( \frac{\partial \varphi}{\partial x_j} \right)^T \in \mathbb{R}^d, \quad \mathbf{n} = (n_t, \boldsymbol{\nu}) \in \mathbb{R}^{d+1},$$

we know that  $\mathbf{n}$  is a normal vector<sup>1</sup> to  $\Sigma$ , and we can suppose without loss of generality that the vector  $\boldsymbol{\nu}$  is a unit vector in  $\mathbb{R}^d$ . If  $\Sigma$  is characteristic, we obtain the following lemma (where  $\lambda_k(\mathbf{u}, \boldsymbol{\nu})$  denotes the  $k$ th eigenvalue of the matrix  $\mathbf{A}(\mathbf{u}, \boldsymbol{\nu})$  defined by (1.3)).

*Lemma 1.1*

Consider a characteristic surface  $\Sigma$  for system (1.2) and let  $\mathbf{n} = (n_t, \boldsymbol{\nu})$ , where  $\boldsymbol{\nu}$  is a unit vector in  $\mathbb{R}^d$ , denote an outward normal vector to  $\Sigma$ . There exists locally an index  $k$ ,  $1 \leq k \leq p$ , such that

$$n_t = -\lambda_k(\mathbf{u}, \boldsymbol{\nu}).$$

*Proof.* By definition, if  $\Sigma$  is characteristic at a point  $(\mathbf{x}, t)$ , the matrix

$$M(\mathbf{u}, \mathbf{n}) = n_t \mathbf{I} + \mathbf{F}'(\mathbf{u}) \cdot \boldsymbol{\nu} = n_t \mathbf{I} + \mathbf{A}(\mathbf{u}, \boldsymbol{\nu}) \quad (1.6)$$

is not invertible, which implies that  $-n_t$  is an eigenvalue of  $\mathbf{A}(\mathbf{u}, \boldsymbol{\nu})$ ,

$$n_t = -\lambda_k(\mathbf{u}, \boldsymbol{\nu}), \quad (1.7)$$

and

$$M(\mathbf{u}, \mathbf{n}) = \mathbf{A}(\mathbf{u}, \boldsymbol{\nu}) - \lambda_k(\mathbf{u}, \boldsymbol{\nu}) \mathbf{I}.$$

By a continuity argument, the index  $k$ , which a priori depends on the point  $(\mathbf{x}, t)$ , will be the same for all points in a neighborhood.  $\square$

Of course, in the one-dimensional case, Definition 1.2 coincides with the usual notions given in Chap. I or in Chap. II: a curve  $x = \varphi(t)$  is characteristic if  $\frac{\partial \varphi}{\partial t} - a(u(x, t)) = 0$  in the scalar case (resp.  $\frac{\partial \varphi}{\partial t} - \lambda_k(\mathbf{u}(\mathbf{x}, t)) = 0$  for a one-dimensional system).

*Remark 1.1.* For an initial boundary value problem in dimension  $d = 2$ , one often says that the boundary of a domain  $\mathcal{O}$  in  $\mathbb{R}^2$  is “non-characteristic” if  $\mathbf{A}(\mathbf{u}, \boldsymbol{\nu}) = \mathbf{f}'(\mathbf{u})\nu_1 + \mathbf{g}'(\mathbf{u})\nu_2$  is invertible, where  $\boldsymbol{\nu}$  is the normal to  $\partial\mathcal{O}$ . Let  $\varphi(x, y) = 0$  denote the equation of  $\partial\mathcal{O}$ , and consider the cylindrical domain  $Q$  in  $\mathbb{R}^2 \times [0, \infty[$  built on  $\mathcal{O}$ . Then  $\partial\mathcal{O}$  is non-characteristic, if the lateral surface  $\Gamma$  of the cylinder  $Q$  is non-characteristic in the sense of Definition 1.2. Indeed, this results from Lemma 1.1 since the normal to  $\Gamma$  is  $(0, \frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y})^T = (0, \boldsymbol{\nu})$ .  $\square$

A characteristic equation is obtained as follows: one makes an appropriate linear combination of the equations of the system (1.2),

$$\boldsymbol{\alpha} \cdot \left\{ \frac{\partial \mathbf{u}}{\partial t} + \sum_j \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) \right\} = 0, \quad \boldsymbol{\alpha} = (\alpha_i) \in \mathbb{R}^p,$$

---

<sup>1</sup> To simplify the notations, we skip some of the transpose signs when writing vectors such as  $(n_t, \boldsymbol{\nu})$  in  $\mathbb{R}^{d+1}$ .

or

$$\sum_{i=1}^p \alpha_i \left\{ \frac{\partial u_i}{\partial t} + \sum_{j,k=1}^d \frac{\partial f_{ij}}{\partial u_k} \frac{\partial u_k}{\partial x_j} \right\} = 0,$$

in order that the resulting equation contains only derivatives of each  $u_k$  in directions lying on a  $d$ -dimensional surface, instead of  $d+1$  as in (1.2) (for  $d \geq 2$ , we cannot in general impose that the directional derivatives have the same direction). We write equivalently that they are normal to the normal vector to a surface  $\Sigma$ ; if the equation of  $\Sigma$  is  $\varphi(\mathbf{x}, t) = 0$ , its normal vector is, as we have just seen,  $(n_t = \frac{\partial \varphi}{\partial t}, \boldsymbol{\nu} = (\frac{\partial \varphi}{\partial x_j})) \in \mathbb{R} \times \mathbb{R}^d$ . The directional derivative for each  $u_k$  is  $(\alpha_k, \Sigma_{ij} \alpha_i \frac{\partial f_{ij}}{\partial u_k} \mathbf{e}_j) \in \mathbb{R} \times \mathbb{R}^d$  ( $\mathbf{e}_j$  is the standard basis of  $\mathbb{R}^d$ ), and thus we impose

$$\alpha_k \frac{\partial \varphi}{\partial t} + \sum_{i,j} \alpha_i \frac{\partial f_{ij}}{\partial u_k} \frac{\partial \varphi}{\partial x_j} = 0, \quad k = 1, \dots, p.$$

This is a  $p \times p$  linear system in the  $\alpha_i$  that has a nontrivial solution. This supposes that the determinant vanishes, i.e.,

$$\det \left( \frac{\partial \varphi}{\partial t} \mathbf{I} + \sum_j \mathbf{f}'_j(\mathbf{u}) \frac{\partial \varphi}{\partial x_j} \right) = 0,$$

which is the definition of a characteristic surface. Therefore, for a characteristic equation, at any point  $(\mathbf{x}, t)$ , the directions in which all  $u_k$  ( $1 \leq k \leq p$ ) are differentiated lie in the tangent space to a surface  $\Sigma$  that is characteristic at this point (see Holt for examples [628]).

The main property is that singularities in the solution propagate along characteristic surfaces. Indeed, writing that a solution  $\mathbf{u}$ , which is assumed to be continuous across  $\Sigma$ , has discontinuous derivatives leads to exactly the same system as above (Whitham, Section 5.9 [1188], Courant and Friedrichs, Chapter II, no. 32 [371]).

If we can find  $p$  independent characteristic equations, the system is in characteristic form. In the multidimensional case, it is not obvious that a simplification may occur if the number of different directions involved for each characteristic equation is too important. However, in dimension  $d = 2$ , setting for  $\boldsymbol{\omega} = (\omega_1, \omega_2)^T \in \mathbb{R}^2$ ,

$$\boldsymbol{\omega}^\perp = (-\omega_2, \omega_1)^T,$$

which is the unit vector directly orthogonal to  $\boldsymbol{\omega}$ , we obtain the following result.

*Lemma 1.2*

*Consider a characteristic surface  $\Sigma$  and let  $\mathbf{n} = (n_t, \boldsymbol{\nu})$ , where  $n_t = -\lambda_k(\mathbf{u}, \boldsymbol{\nu})$  and  $\boldsymbol{\nu}$  is a unit vector in  $\mathbb{R}^2$ ; denote an outward normal vector to  $\Sigma$ . A smooth solution  $\mathbf{u}$  of (1.1) satisfies*

$$\mathbf{l}_k^T(\mathbf{u}, \boldsymbol{\nu}) \frac{d\mathbf{u}}{ds_k} + (\mathbf{l}_k^T(\mathbf{u}, \boldsymbol{\nu}) \mathbf{A}(\mathbf{u}, \boldsymbol{\nu}^\perp))(\boldsymbol{\nu}^\perp \cdot \mathbf{grad} \mathbf{u}) = 0, \quad (1.8)$$

where  $\mathbf{l}_k(\mathbf{u}, \boldsymbol{\omega})$  denote the eigenvector of  $\mathbf{A}(\mathbf{u}, \boldsymbol{\omega})^T$  and

$$\frac{d\mathbf{u}}{ds_k} = \frac{\partial \mathbf{u}}{\partial t} + \lambda_k(\mathbf{u}, \boldsymbol{\nu}) \left( \nu_1 \frac{\partial \mathbf{u}}{\partial x} + \nu_2 \frac{\partial \mathbf{u}}{\partial y} \right).$$

*Proof.* Note first that by (1.7), the vector  $(n_t, \boldsymbol{\nu})$ , which is normal to  $\Sigma$ , is orthogonal to  $(1, \lambda_k(\mathbf{u}, \boldsymbol{\nu})\boldsymbol{\nu})$  and  $(0, \boldsymbol{\nu}^\perp)$ ; thus, each equation in (1.8) contains only derivatives in directions that lie in the plane tangent to  $\Sigma$ ; they are particular characteristic equations.

Now, assuming that  $\mathbf{u}$  is a smooth solution, we have

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{u})}{\partial x} + \frac{\partial \mathbf{g}(\mathbf{u})}{\partial y} = \frac{\partial \mathbf{u}}{\partial t} + \mathbf{f}'(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} + \mathbf{g}'(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial y}.$$

For any unit vector  $\boldsymbol{\omega} \in \mathbb{R}^2$  and any  $\mathbf{w} \in \mathbb{R}^2$ , since  $\mathbf{w} = (\boldsymbol{\omega} \cdot \mathbf{w})\boldsymbol{\omega} + (\boldsymbol{\omega}^\perp \cdot \mathbf{w})\boldsymbol{\omega}^\perp$ , we have the obvious relation

$$\mathbf{w} \cdot \mathbf{u} = (\boldsymbol{\omega} \cdot \mathbf{w})(\boldsymbol{\omega} \cdot \mathbf{u}) + (\boldsymbol{\omega}^\perp \cdot \mathbf{w})(\boldsymbol{\omega}^\perp \cdot \mathbf{u}), \quad (1.9)$$

which yields, setting  $\mathbb{F}(\mathbf{u}) = (\mathbf{f}(\mathbf{u}), \mathbf{g}(\mathbf{u}))^T$ , and  $\mathbf{grad} \mathbf{u} = \left( \frac{\partial \mathbf{u}}{\partial x}, \frac{\partial \mathbf{u}}{\partial y} \right)^T$ ,

$$\mathbb{F}'(\mathbf{u}) \cdot \mathbf{grad} \mathbf{u} = (\mathbb{F}'(\mathbf{u}) \cdot \boldsymbol{\omega})(\mathbf{grad} \mathbf{u} \cdot \boldsymbol{\omega}) + (\mathbb{F}'(\mathbf{u}) \cdot \boldsymbol{\omega}^\perp)(\mathbf{grad} \mathbf{u} \cdot \boldsymbol{\omega}^\perp).$$

Thus, smooth solutions satisfy for any unit vector  $\boldsymbol{\omega} \in \mathbb{R}^2$

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbb{F}'(\mathbf{u}) \cdot \boldsymbol{\omega})(\mathbf{grad} \mathbf{u} \cdot \boldsymbol{\omega}) + (\mathbb{F}'(\mathbf{u}) \cdot \boldsymbol{\omega}^\perp)(\mathbf{grad} \mathbf{u} \cdot \boldsymbol{\omega}^\perp) = \mathbf{0},$$

and in particular for  $\boldsymbol{\omega} = \boldsymbol{\nu} = \left( \frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y} \right)^T$ ,

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u}, \boldsymbol{\nu})(\mathbf{grad} \mathbf{u} \cdot \boldsymbol{\nu}) + \mathbf{A}(\mathbf{u}, \boldsymbol{\nu}^\perp)(\mathbf{grad} \mathbf{u} \cdot \boldsymbol{\nu}^\perp) = \mathbf{0}. \quad (1.10)$$

Now, (1.10) implies

$$\begin{aligned} & \mathbf{l}_k(\mathbf{u}, \boldsymbol{\nu}) \cdot \left\{ \frac{\partial \mathbf{u}}{\partial t} + \operatorname{div} \mathbb{F}(\mathbf{u}) \right\} \\ &= \mathbf{l}_k^T(\mathbf{u}, \boldsymbol{\nu}) \left\{ \frac{\partial \mathbf{u}}{\partial t} + \lambda_k(\mathbf{u}, \boldsymbol{\nu}) \boldsymbol{\nu} \cdot \mathbf{grad} \mathbf{u} + \mathbf{A}(\mathbf{u}, \boldsymbol{\nu}^\perp)(\boldsymbol{\nu}^\perp \cdot \mathbf{grad} \mathbf{u}) \right\} = 0. \end{aligned}$$

Consider the differentiation along the integral curves  $C_k(\boldsymbol{\nu})$  of the differential system

$$\frac{d\mathbf{x}}{dt} = \lambda_k(\mathbf{u}, \boldsymbol{\nu})\boldsymbol{\nu}. \quad (1.11)$$

These curves are the integral curves of the vector field  $\lambda_k(\mathbf{u}, \boldsymbol{\nu})\boldsymbol{\nu}$ , which is tangent to  $\Sigma$ , and they lie on  $\Sigma$ . Indeed, we can assume that the curve is parametrized by  $t$ , and we have then

$$\frac{dx}{dt} = \lambda_k(\mathbf{u}, \boldsymbol{\nu})\nu_1, \quad \frac{dy}{dt} = \lambda_k(\mathbf{u}, \boldsymbol{\nu})\nu_2;$$

hence

$$\frac{d}{dt}\varphi(t, x(t), y(t)) = \frac{\partial \varphi}{\partial t} + \lambda_k(\mathbf{u}, \boldsymbol{\nu})\left\{\nu_1 \frac{\partial \varphi}{\partial x} + \nu_2 \frac{\partial \varphi}{\partial y}\right\} = 0.$$

The differentiation operator along  $C_k(\boldsymbol{\nu})$  is given by

$$\frac{d}{ds_k} = \frac{\partial}{\partial t} + \lambda_k(\mathbf{u}, \boldsymbol{\nu})(\boldsymbol{\nu} \cdot \nabla), \quad (1.12)$$

i.e.,

$$\frac{d\mathbf{u}}{ds_k} = \frac{\partial \mathbf{u}}{\partial t} + \lambda_k(\mathbf{u}, \boldsymbol{\nu})\left(\frac{\partial \mathbf{u}}{\partial x}\nu_1 + \frac{\partial \mathbf{u}}{\partial y}\nu_2\right),$$

which gives the desired result.  $\square$

*Remark 1.2.* In the scalar case ( $p = 1$ ), for a smooth solution  $u$  of (1.2), we can introduce the characteristic curves as in the one-dimensional case (Chap. I, Section 4); they are the integral curves  $t \rightarrow \mathbf{x}(t)$  of the ordinary differential system

$$\frac{d\mathbf{x}}{dt} = \mathbf{a}(u(\mathbf{x}, t)), \quad \mathbf{a} = \mathbb{F}' = (f'_1, \dots, f'_d)^T,$$

i.e.,

$$\frac{dx_j}{dt} = a_j(u(\mathbf{x}, t)), \quad j = 1, \dots, d, \quad a_j = f'_j.$$

We check easily that

$$\frac{d}{dt}(u(\mathbf{x}(t), t)) = 0,$$

so that again  $u$  is constant along characteristics which are thus straight lines. We can then follow the method of characteristics and prove the existence of a smooth solution for  $t$  small enough; if there exists  $\mathbf{y}$  such that  $\operatorname{div}_{\mathbf{x}} \mathbf{a}(u_0(\mathbf{y})) < 0$ , the critical time is

$$t^* = \left\{ - \min_{y \in \mathbb{R}^d} \operatorname{div} \mathbf{a}(u_0(\mathbf{y})) \right\}^{-1}.$$

Otherwise, there exists a smooth global solution (for details, we refer to Majda, Section 3.1 [840]).  $\square$

*Remark 1.3.* Notice that the definition (1.11) of  $C_k(\boldsymbol{\nu})$  and (1.12) can be extended in any dimension  $d$ . In the particular case where we look for a smooth solution of the form (1.5c)

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(\mathbf{x} \cdot \boldsymbol{\omega}, t),$$

where  $\boldsymbol{\omega}$  is a fixed unit vector in  $\mathbb{R}^d$  and  $\mathbf{v}(\zeta, t) : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}^p$ , we have seen that  $\mathbf{v}$  is a solution of a system (1.5a) in dimension  $d = 1$ ,

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial}{\partial \zeta} (\mathbf{F} \cdot \boldsymbol{\omega})(\mathbf{v}) = \frac{\partial \mathbf{v}}{\partial t} + \mathbf{A}(\mathbf{v} \cdot \boldsymbol{\omega}) \frac{\partial \mathbf{v}}{\partial \zeta} = \mathbf{0}.$$

One can write this system in characteristic form following the arguments of Chap. II, Sect. 5. Setting for fixed  $\boldsymbol{\omega}$

$$\frac{d\mathbf{v}}{ds_k} = \frac{\partial}{\partial t} + \lambda_k(\mathbf{v}, \boldsymbol{\omega}) \frac{\partial \mathbf{v}}{\partial \zeta},$$

which is the differentiation along the characteristic curve

$$\frac{d\zeta}{dt} = \lambda_k(\mathbf{v}(\zeta, t), \boldsymbol{\omega}), \quad (1.13)$$

we get

$$\mathbf{l}_k^T(\mathbf{v}, \boldsymbol{\omega}) \left\{ \frac{\partial \mathbf{v}}{\partial t} + \mathbf{A}(\mathbf{v} \cdot \boldsymbol{\omega}) \frac{\partial \mathbf{v}}{\partial \zeta} \right\} = \mathbf{l}_k^T(\mathbf{v}, \boldsymbol{\omega}) \frac{d\mathbf{v}}{ds_k} = 0.$$

Note that  $\frac{d\mathbf{v}}{ds_k}$  does coincide with  $\frac{d\mathbf{u}}{ds_k}$  defined by (1.12),

$$\begin{aligned} \frac{d\mathbf{u}}{ds_k} &= \frac{\partial \mathbf{u}}{\partial t} + \lambda_k(\mathbf{u}, \boldsymbol{\omega}) \boldsymbol{\omega} \cdot \mathbf{grad} \mathbf{u} \\ &= \frac{\partial \mathbf{v}}{\partial t} + \lambda_k(\mathbf{v}, \boldsymbol{\omega}) |\boldsymbol{\omega}| \frac{\partial \mathbf{v}}{\partial \zeta} = \frac{d\mathbf{v}}{ds_k}, \end{aligned}$$

which justifies the use of the same notation. Also, (1.13) yields

$$\frac{d}{dt}(\mathbf{x} \cdot \boldsymbol{\omega}) = \lambda_k(\mathbf{u}, \boldsymbol{\omega}),$$

which shows that the projection of  $C_k(\boldsymbol{\omega})$  in the direction  $\boldsymbol{\omega}$  (which lies in the  $(\zeta, t)$ -plane  $\mathbb{R}\boldsymbol{\omega} \times \mathbb{R}_+$  of the  $(\mathbf{x}, t)$ -space  $\mathbb{R}^d \times \mathbb{R}_+$ ) coincides with the usual characteristic of the one-dimensional projected Eq. (1.5).  $\square$

### 1.3 Simple Plane Waves

A plane-centered rarefaction wave is a “self-similar” continuous (piecewise  $C^1$ ) solution

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v} \left( \frac{\mathbf{x} \cdot \boldsymbol{\omega}}{t} \right), \quad (1.14)$$

where  $\mathbf{v} : \mathbb{R} \rightarrow \mathbb{R}^p$  is a curve in  $\mathbb{R}^p$  and  $\boldsymbol{\omega}$  is a given constant vector in  $\mathbb{R}^d$ . We shall set

$$\xi = \frac{\mathbf{x} \cdot \boldsymbol{\omega}}{t}.$$

For any fixed  $\bar{\xi}$ ,  $\mathbf{u}$  is constant on the (hyper)planes  $\mathbf{x} \cdot \boldsymbol{\omega} = \bar{\xi}t$ .

As in the one-dimensional case (see Chap. II, Sect. 3.1, Theorem 3.1), we find that  $\mathbf{v}$  must satisfy

$$(\mathbf{A}(\mathbf{v}(\xi), \boldsymbol{\omega}) - \xi \mathbf{I})\mathbf{v}'(\xi) = \mathbf{0},$$

so that either  $\mathbf{v}$  is constant or  $\mathbf{v}'$  is an eigenvector of  $\mathbf{A}(\mathbf{v}, \boldsymbol{\omega})$  associated with  $\xi$ ,

$$\begin{cases} \mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi), \boldsymbol{\omega}), \\ \xi = \lambda_k(\mathbf{v}(\xi), \boldsymbol{\omega}). \end{cases}$$

This can be solved if the  $k$ th field is genuinely nonlinear in the direction  $\boldsymbol{\omega}$ . If this is the case, we can connect two states  $\mathbf{u}_L$  and  $\mathbf{u}_R$  by a  $k$ -plane-centered rarefaction wave if  $\mathbf{u}_R$  lies on the integral curve  $\mathcal{R}_k(\mathbf{u}_L, \boldsymbol{\omega})$  of  $\mathbf{r}_k(\cdot, \boldsymbol{\omega})$ ,

$$\begin{cases} \mathbf{v}'(\xi) = \mathbf{r}_k(\mathbf{v}(\xi), \boldsymbol{\omega}), & \lambda_k(\mathbf{u}_L, \boldsymbol{\omega}) < \xi < \lambda_k(\mathbf{u}_L, \boldsymbol{\omega}) + \varepsilon, \\ \mathbf{v}(\lambda_k(\mathbf{u}_L, \boldsymbol{\omega})) = \mathbf{u}_L, \end{cases} \quad (1.15)$$

which exists for  $\varepsilon$  small enough and such that  $\lambda_k(\mathbf{v}(\xi), \boldsymbol{\omega})$  increases from  $\lambda_k(\mathbf{u}_L, \boldsymbol{\omega})$  to  $\lambda_k(\mathbf{u}_R, \boldsymbol{\omega})$ .

When  $\boldsymbol{\omega}$  varies, the  $k$ -rarefaction curves  $\mathcal{R}_k(\mathbf{u}_L, \boldsymbol{\omega})$  form a (rarefaction) cone  $\mathcal{R}_k(\mathbf{u}_L)$ .

*Example 1.1. (Revisited)* The rarefaction curve  $\mathcal{R}_{\pm}(\mathbf{u}_L, \boldsymbol{\omega})$  is defined by

$$\mathbf{V}'(\xi) = \mathbf{r}_{\pm}(\mathbf{V}(\xi), \boldsymbol{\omega}) = \left( p', \pm \omega_1 \frac{\sqrt{p'(\rho)}}{\rho}, \pm \omega_2 \frac{\sqrt{p'(\rho)}}{\rho} \right)^T.$$

We can parametrize the curve by  $\rho$  and, after elementary computations (similar to those of the  $p$ -system in Chap. II, Sect. 7), we find

$$\begin{aligned} u = u(\rho) &= u_L \pm \omega_1 \int_{\rho_L}^{\rho} \sqrt{p'(r)} \frac{dr}{r}, \\ v = v(\rho) &= v_L \pm \omega_2 \int_{\rho_L}^{\rho} \sqrt{p'(r)} \frac{dr}{r}, \end{aligned}$$

and the rarefaction cone  $\mathcal{R}_{\pm}(\mathbf{u}_L)$  is given in both cases by

$$(u - u_L)^2 + (v - v_L)^2 = \left( \int_{\rho_L}^{\rho} \sqrt{p'(r)} \frac{dr}{r} \right)^2.$$

Under the assumption made for genuine nonlinearity,  $p' > 0$  and  $\rho p'' + 2p' > 0$ , we can discriminate between the fields. Indeed, the hypothesis  $\rho p'' + 2p' > 0$  implies that  $\lambda_+$  (resp.  $\lambda_-$ ) increases (resp. decreases) along  $\mathcal{R}_+(\mathbf{u}_L)$  (resp.  $\mathcal{R}_-(\mathbf{u}_L)$ ) since

$$\frac{d\lambda_{\pm}}{d\rho} = D_{\mathbf{v}}\lambda_{\pm} \cdot \mathbf{r}_{\pm} = \pm \left( \frac{p''}{2\sqrt{p'}} + \frac{\sqrt{p'}}{\rho} \right) = \pm \frac{(\rho p'' + 2p')}{2\rho\sqrt{p'}}.$$

Given  $\mathbf{u}_R \in \mathcal{R}_{\pm}(\mathbf{u}_L)$ , it can be connected to  $\mathbf{u}_L$  by a plane rarefaction +wave (resp. -wave) if  $\rho_R > \rho_L$  (resp.  $\rho_R < \rho_L$ ).  $\square$

As in the one-dimensional case (Chap. II, Sect. 3.2), we can look more generally for continuous piecewise  $C^1$  solutions of the form

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(\varphi(\mathbf{x} \cdot \boldsymbol{\omega}, t)), \quad (1.16)$$

where  $\varphi(\zeta, t) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ ,  $\mathbf{v}(\varphi) : \mathbb{R} \rightarrow \mathbb{R}^p$ , and  $\boldsymbol{\omega}$  is a *given* constant vector in  $\mathbb{R}^d$ . Such a function is called a *simple plane wave* solution: it is constant at any time  $t$  along the hyperplane perpendicular to the direction  $\boldsymbol{\omega}$  and takes its values on a curve in  $\mathbb{R}^p$ . Substituting (1.16) into (1.2) gives

$$\mathbf{v}' \frac{\partial \varphi}{\partial t} + \mathbf{A}(\mathbf{v}, \boldsymbol{\omega}) \mathbf{v}' \frac{\partial \varphi}{\partial \zeta} = \mathbf{0}. \quad (1.17)$$

Following the arguments of Chap. II, Sect. 3.2, we solve (1.17) in two steps. First, we take for  $\mathbf{v}(\varphi) : \mathbb{R} \rightarrow \mathbb{R}^p$  an integral curve of the nonlinear ordinary differential equation

$$\begin{cases} \mathbf{v}'(\varphi) = \mathbf{r}_k(\mathbf{v}(\varphi), \boldsymbol{\omega}), \\ \mathbf{v}(0) = \mathbf{v}_0. \end{cases} \quad (1.18)$$

Such a solution exists at least locally. Then, given  $\mathbf{v} = \mathbf{v}(\cdot, \boldsymbol{\omega})$ ,  $\varphi$  satisfies the first-order quasi-linear equation

$$\begin{cases} \frac{\partial \varphi}{\partial t} + \lambda_k(\mathbf{v}, \boldsymbol{\omega}) \frac{\partial \varphi}{\partial \zeta}, & \zeta \in \mathbb{R}, \quad t > 0, \\ \varphi(\xi, 0) = \varphi_0(\xi). \end{cases}$$

As in Chap. II, formula (3.20), we obtain that  $\varphi$  is constant along the characteristics given by (1.13),

$$\frac{d\zeta}{dt} = \lambda_k(\mathbf{v}(\varphi(\zeta, t)), \boldsymbol{\omega}),$$

which are thus straight lines in the  $(\zeta, t)$ -plane, and

$$\varphi(\zeta, t) = \varphi_0(\zeta - \lambda_k(\mathbf{v}(\varphi(\zeta, t)), \boldsymbol{\omega})t)$$

so long as  $\varphi$  remains smooth. In the variables  $(\mathbf{x}, t)$ , the above characteristics give (hyper)planes

$$\mathbf{x} \cdot \boldsymbol{\omega} = \lambda_k(\mathbf{u}, \boldsymbol{\omega})t + C$$

along which  $\mathbf{u}$  is constant. Note that these (hyper)planes are obviously characteristic in the sense of Definition 1.2. Finally, we have the implicit formula

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(\varphi_0(\mathbf{x} \cdot \boldsymbol{\omega} - \lambda_k(\mathbf{u}, \boldsymbol{\omega})t)).$$

We can study some particular simple plane waves. On the one hand, assuming that the  $k$ th field is genuinely nonlinear in the direction  $\boldsymbol{\omega}$ , and normalizing the vector  $\mathbf{r}_k(\mathbf{v}, \boldsymbol{\omega})$  so that

$$D_{\mathbf{v}}\lambda_k(\mathbf{v}, \boldsymbol{\omega}) \cdot \mathbf{r}_k(\mathbf{v}, \boldsymbol{\omega}) = 1,$$

we see that the equation in  $\varphi$  resumes in Burgers' equation (see Example 3.5, Chap. II). We recover as particular simple plane waves the rarefaction waves studied above, corresponding to  $\varphi(\zeta, t) = \frac{\zeta}{t}$ . In the one-dimensional case  $d = 1$ , for a  $k$ -rarefaction wave, the characteristics of the  $k$ th field form a fan of straight lines in the  $(x, t)$ -plane (see Chap. II, Example 3.4). In the multidimensional case, for a plane rarefaction wave (1.14) connecting  $\mathbf{u}_L$  and  $\mathbf{u}_R$ , the characteristics (1.18) form a converging pencil of (hyper)planes in the  $(\mathbf{x}, t)$ -space passing through  $t = 0$ ,  $\mathbf{x} \cdot \boldsymbol{\omega} = 0$  (more generally  $t = t_0$ ,  $(\mathbf{x} - \mathbf{x}_0) \cdot \boldsymbol{\omega} = 0$ ).

On the other hand, assuming that the  $k$ th field is linearly degenerate in the direction  $\boldsymbol{\omega}$ , we see that  $\lambda_k(\mathbf{v}, \boldsymbol{\omega})$  is constant on a  $k$ -simple plane-wave since the variation of  $\lambda_k$  on the curve where  $\mathbf{v}$  takes its values is given by

$$D_{\mathbf{v}}\lambda_k \cdot \mathbf{v}' = D_{\mathbf{v}}\lambda_k(\mathbf{v}, \boldsymbol{\omega}) \cdot \mathbf{r}_k(\mathbf{v}, \boldsymbol{\omega}) = 0.$$

Thus, the functions

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(\varphi_0(\mathbf{x} \cdot \boldsymbol{\omega} - \bar{\lambda}_k t)), \quad \bar{\lambda}_k = \lambda_k(\mathbf{u}, \boldsymbol{\omega})$$

are  $k$ -simple plane waves, and the characteristics (1.18) of the  $k$ th field form a pencil of parallel (hyper)planes

$$\mathbf{x} \cdot \boldsymbol{\omega} - \bar{\lambda}_k t = C.$$

*Example 1.1.(Revisited).* The integral curve of  $\mathbf{r}_0(\mathbf{u}, \boldsymbol{\omega}) = (0, -\omega_2, \omega_1)^T = (0, \boldsymbol{\omega}^\perp)$  corresponding to the linearly degenerate field  $\lambda_0(\mathbf{u}, \boldsymbol{\omega}) = \mathbf{u} \cdot \boldsymbol{\omega}$  issued from  $\mathbf{u}_L$  is a straight line (intersection of the planes  $p = p_L$ , or equivalently  $\rho = \rho_L$ , and  $(u - u_L)\omega_1 + (v - v_L)\omega_2 = 0$ ) on which  $\lambda_0$  is obviously constant ( $\lambda_0(\mathbf{u}, \boldsymbol{\omega}) = \mathbf{u}_L \cdot \boldsymbol{\omega}$ ). The set spanned by these curves as  $\boldsymbol{\omega}$  varies is the plane  $\rho = \text{const.} = \rho_L$ .  $\square$

*Remark 1.4.* The definition and existence of general rarefaction waves that are not plane simple waves can be found in Alinhac [27].  $\square$

Now we can also look for some particular *discontinuous* solutions.

## 1.4 Shock Waves

Let us recall that a weak solution satisfies the Rankine–Hugoniot condition along the (hyper)surfaces of discontinuity  $\Sigma$  (see the Chap. I, (4.8))

$$n_t[\mathbf{u}] + \sum_{j=1}^d [\mathbf{f}_j(\mathbf{u})] n_{x_j} = \mathbf{0}, \quad (1.19)$$

where  $\mathbf{n} = (n_t, \mathbf{n}_\mathbf{x})$  is the normal vector to  $\Sigma$ . We can define in particular *plane* shock-wave solutions

$$\mathbf{u}(\mathbf{x}, t) = \begin{cases} \mathbf{u}_L, & \mathbf{x} \cdot \boldsymbol{\omega} - \sigma t < 0, \\ \mathbf{u}_R, & \mathbf{x} \cdot \boldsymbol{\omega} - \sigma t > 0, \end{cases} \quad (1.20a)$$

which satisfy the Rankine–Hugoniot condition across the (hyper)plane  $\Sigma = \{(\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R}_+ / \mathbf{x} \cdot \boldsymbol{\omega} - \sigma t = 0\}$ ,

$$-\sigma(\mathbf{u}_R - \mathbf{u}_L) + \sum_{j=1}^d (\mathbf{f}_j(\mathbf{u}_R) - \mathbf{f}_j(\mathbf{u}_L)) \omega_j = \mathbf{0}, \quad (1.20b)$$

or

$$-\sigma[\mathbf{u}] + [\mathbb{F}(\mathbf{u})] \cdot \boldsymbol{\omega} = \mathbf{0},$$

where  $\mathbf{u}_L, \mathbf{u}_R$  are *constant* states. Assuming, for instance, that the front propagates along the  $x_d$ -axis, i.e., choosing  $\boldsymbol{\omega} = (0, 0, \dots, 1)^T$ , we get

$$\mathbf{u}(\mathbf{x}, t) \begin{cases} \mathbf{u}_L, & x_d < \sigma t, \\ \mathbf{u}_R, & x_d > \sigma t, \end{cases}$$

$\Sigma = \{(\mathbf{x}, t) / x_d = \sigma t\}$ , and  $\mathbf{u}$  satisfies a one-dimensional jump condition

$$\sigma(\mathbf{u}_R - \mathbf{u}_L) + (\mathbf{f}_d(\mathbf{u}_R) - \mathbf{f}_d(\mathbf{u}_L)) = \mathbf{0}.$$

The theory of these shock waves is developed in Chap. II. Thus, for given  $\boldsymbol{\omega}$  and  $\mathbf{u}_L$ , assuming that the system is strictly hyperbolic, the Rankine–Hugoniot set of  $\mathbf{u}_L$  contains locally the  $k$ -shock curves  $\mathcal{S}_k(\mathbf{u}_L, \boldsymbol{\omega})$  associated to a (simple) eigenvalue  $\lambda_k(\mathbf{u}, \boldsymbol{\omega})$  (see Chap. II, Theorem 4.1). When  $\boldsymbol{\omega}$

varies, the curves  $\mathcal{S}_k(\mathbf{u}_L, \omega)$  form a “cone.” The curve  $\mathcal{S}_k(\mathbf{u}_L, \omega)$  is tangent to  $\mathbf{r}_k(\mathbf{u}_L, \omega)$  at  $\mathbf{u}_L$  and can be parametrized by

$$\mathbf{u}(\varepsilon, \omega) = \mathbf{u}_L + \varepsilon \mathbf{r}_k(\mathbf{u}_L, \omega) + \mathcal{O}(\varepsilon^2), \quad (1.21a)$$

$$\sigma = \lambda_k(\mathbf{u}_L, \omega) + \frac{\varepsilon}{2} D_{\mathbf{u}} \lambda_k(\mathbf{u}_L, \omega) \cdot \mathbf{r}_k(\mathbf{u}_L, \omega) + \mathcal{O}(\varepsilon^2). \quad (1.21b)$$

Now, assume first that the  $k$ th field is genuinely nonlinear, i.e.,

$$D\lambda_k(\mathbf{u}, \omega) \cdot \mathbf{r}_k(\mathbf{u}, \omega) \neq 0.$$

Since  $(n_t, \mathbf{n}_{\mathbf{x}}) = (-\sigma, \omega)$  is normal to  $\Sigma$ , it follows from (1.21b) that the eigenvalues  $\lambda_k(\mathbf{u}, \omega) - \sigma$  of the matrix (1.6),

$$M(\mathbf{u}, \mathbf{n}) = \frac{\partial \varphi}{\partial t} \mathbf{I} + \sum_{j=1}^d \mathbf{A}_j(\mathbf{u}) \frac{\partial \varphi}{\partial x_j} = -\sigma \mathbf{I} + \mathbf{A}(\mathbf{u}, \omega),$$

are  $\neq 0$  (i.e.,  $M(\mathbf{u}, \mathbf{n})$  is invertible) for  $\mathbf{u} = \mathbf{u}_L$  and  $\mathbf{u}_R$ , which means that  $\Sigma$  is non-characteristic.

If the  $k$ th field is linearly degenerate, i.e.,

$$D\lambda_k(\mathbf{u}, \omega) \cdot \mathbf{r}_k(\mathbf{u}, \omega) = 0,$$

then  $M(\mathbf{u}, \mathbf{n})$  is singular and,

$$\sigma = \lambda_k(\mathbf{u}_L, \omega) = \lambda_k(\mathbf{u}_R, \omega) = \bar{\lambda}_k,$$

and so the (hyper)plane  $\mathbf{x} \cdot \omega - \sigma t$  is characteristic. Thus, if  $\mathbf{u}_R \in \mathcal{S}_k(\mathbf{u}_L, \omega)$ , which is now an integral curve of  $\mathbf{r}_k(\mathbf{u}, \omega)$ , we have a  $k$ -contact discontinuity

$$\mathbf{u}(\mathbf{x}, t) = \begin{cases} \mathbf{u}_L, & \mathbf{x} \cdot \omega - \bar{\lambda}_k t < 0, \\ \mathbf{u}_R, & \mathbf{x} \cdot \omega - \bar{\lambda}_k t > 0. \end{cases}$$

For a planar shock connecting two states  $\mathbf{u}_-$  and  $\mathbf{u}_+$ , we can impose the Lax entropy conditions. These conditions were obtained in the one-dimensional case (see Chap. II, Sect. 5) by considering the number of characteristics impinging on  $\Sigma$  or emerging from  $\Sigma$  considered as a boundary. This number corresponds to the number of positive and negative eigenvalues of the matrix  $M(\mathbf{u}_{\pm}, \mathbf{n})$  defined by (1.6). For instance, if  $p = d = 2$ , a one-planar shock connecting  $\mathbf{u}_-$  and  $\mathbf{u}_+$  satisfies the Lax entropy conditions if

$$\lambda_1(\mathbf{u}_+, \omega) < \sigma < \lambda_1(\mathbf{u}_-, \omega), \quad \sigma < \lambda_2(\mathbf{u}_+, \omega),$$

and a two-shock if

$$\lambda_2(\mathbf{u}_+, \omega) < \sigma < \lambda_2(\mathbf{u}_-, \omega), \quad \lambda_1(\mathbf{u}_-, \omega) < \sigma.$$

In particular, if  $\boldsymbol{\omega} = (0, 1)^T$ , the  $\lambda_i(\mathbf{u}, \boldsymbol{\omega})$ ,  $i = 1, 2$ , are in this case the eigenvalues of  $\mathbf{A}(\mathbf{u}, \boldsymbol{\omega}) = \mathbf{f}'_2(\mathbf{u}) = \mathbf{g}'(\mathbf{u})$ .

*Remark 1.5.* One can prove more generally the existence of multidimensional shock fronts (at least for sufficiently short times  $t$ ) that are not plane shock waves. These are piecewise smooth solutions  $\mathbf{u}(\mathbf{x}, t) = \mathbf{u}_+(\mathbf{x}, t), \mathbf{u}_-(\mathbf{x}, t)$  on each side of a hypersurface  $\Sigma$  in the  $(\mathbf{x}, t)$ -space, satisfying (1.19) across  $\Sigma$ . For instance, in the two-dimensional case, if we assume that, at least for short times,  $\Sigma = \{(x, y, t), x = \sigma(y, t)\}$ , with  $\sigma(y, 0) = 0$ , then  $\mathbf{u}(x, y, t) = \mathbf{u}_-(x, y, t)$  for  $x < \sigma(y, t)$  and  $\mathbf{u}(x, y, t) = \mathbf{u}_+(x, y, t)$  for  $x > \sigma(y, t)$ , and  $\mathbf{u}_+$  and  $\mathbf{u}_-$  are linked by the Rankine–Hugoniot condition across  $\Sigma$ .

The construction of a shock front, i.e., the existence of  $\mathbf{u}_+$  satisfying (1.2) and (1.19), appears as a free boundary value problem since the surface  $\Sigma$  is not known in advance in the nonlinear case. If the equation of  $\Sigma$  is given by  $\varphi(\mathbf{x}, t) = 0$ , and the normal to  $\Sigma$  by  $(n_t, \mathbf{n}_\mathbf{x}) = \left( \frac{\partial \varphi}{\partial t}, \frac{\partial \varphi}{\partial x_1}, \dots, \frac{\partial \varphi}{\partial x_d} \right)^T$ , we shall require first that  $\Sigma$  is non-characteristic, which means, as we have already seen, that the matrix (1.6),

$$M(\mathbf{u}, \mathbf{n}) = \frac{\partial \varphi}{\partial t} \mathbf{I} + \sum_{j=1}^d \mathbf{A}_j(\mathbf{u}) \frac{\partial \varphi}{\partial x_j} = n_t \mathbf{I} + \mathbf{A}(\mathbf{u}, \mathbf{n}_\mathbf{x}),$$

is invertible for  $\mathbf{u} = \mathbf{u}_+$  and  $\mathbf{u}_-$ . Following the example of planar shock waves, we can also impose a number of boundary conditions for more general shock fronts. However, they are no longer sufficient (for  $d > 1$ ) to ensure that the problem is well-posed. Hence, stability conditions are needed (see Majda [840]).

The shock front solutions are constructed as progressing waves emanating from a “shock front” with initial data  $\mathbf{u}_0(\mathbf{x})$  given on each side of a surface  $\Sigma_0$ , respectively, by  $\mathbf{u}_{0+}(\mathbf{x}), \mathbf{u}_{0-}(\mathbf{x})$ . The stability of these shock fronts is studied by perturbation of plane shock waves (see Majda [840] for the case of strong shocks, Métivier [868] for weak shocks). For the thorough study of a solution presenting two shock waves in the case  $p = d = 2$ , and of their interactions, see Métivier [867].  $\square$

## 2 The Gas Dynamics Equations in Two Space Dimensions

The Euler system of gas dynamics in two dimensions is given by

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial y}(\rho v^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) + \frac{\partial}{\partial y}((\rho e + p)v) = 0 \end{cases} \quad (2.1)$$

(see Example 2.1 in the Chap. I). The pressure  $p$  is related to  $\varepsilon$  and  $\rho$  by an equation of state of the form

$$p = p(\rho, \varepsilon), \quad e = \varepsilon + \frac{1}{2}(u^2 + v^2).$$

For a polytropic ideal gas, we have  $p = (\gamma - 1)\rho\varepsilon$ , where  $\gamma > 1$  is a constant. When the flow is isentropic, the system (2.1) reduces to the system (1.4) of Example 1.1. Setting

$$\mathbf{U} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho e \end{pmatrix}, \quad \mathbf{f}(\mathbf{U}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (\rho e + p)u \end{pmatrix}, \quad \mathbf{g}(\mathbf{U}) = \begin{pmatrix} \rho v \\ \rho uv \\ \rho u^2 + p \\ (\rho e + p)v \end{pmatrix}, \quad (2.2a)$$

the system (2.1) can be written in the general form (1.1),

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{U}) + \frac{\partial}{\partial y}\mathbf{g}(\mathbf{U}) = \mathbf{0}. \quad (2.2b)$$

Note that the Jacobian matrices of  $\mathbf{f}$  and  $\mathbf{g}$  cannot be diagonalized simultaneously (i.e., in the same basis), but any linear combination of both can be diagonalized. See [1180] for 3d computations.

## 2.1 Entropy and Entropy Variables

Let us recall that the system is endowed with an entropy

$$\mathcal{U}(\mathbf{U}) = -\rho s, \quad (2.3a)$$

with associated entropy flux

$$\mathcal{F}(\mathbf{U}) = -\rho s \mathbf{u} \quad (\text{i.e., } \mathcal{F}_i = -\rho s u_i). \quad (2.3b)$$

(For ease of notation, we have set here  $\mathbf{u} = (u_1, u_2) = (u, v)$ ). In these expressions, we consider the (thermodynamic) specific entropy  $s$  as a function

$s = s(\tau, \varepsilon)$ , which is strictly convex, and we have seen that  $-\rho s\left(\frac{1}{\rho}, \frac{E}{\rho} - \frac{|\rho \mathbf{u}|^2}{2\rho^2}\right)$  is then a strictly convex function of the conservative variables  $\mathbf{U} = (\rho, \rho \mathbf{u}, E)^T$  (see the Chap. I, Example 5.3, and Chap. III, Sect. 1). The system (2.1) is thus symmetrizable, and the entropy variables that symmetrize the system (see Theorem 5.2, Chap. I) are given by the following lemma.

*Lemma 2.1*

*The entropy variables*

$$\mathbf{V}^T = \mathcal{U}'(\mathbf{U})$$

*for the Euler system (2.2) are given by*

$$\mathbf{V}^T = T^{-1}\left(h - Ts - \frac{|\mathbf{u}|^2}{2}, \mathbf{u}, -1\right), \quad (2.4)$$

*where  $h$  is the specific enthalpy,  $h = \varepsilon + \frac{p}{\rho}$ .*

*Proof.* Indeed, in order to compute

$$\mathbf{V}^T = \mathcal{U}'(\mathbf{U}) = -\left(\frac{\partial(\rho s)}{\partial \rho}, \frac{\partial(\rho s)}{\partial(\rho \mathbf{u})}, \frac{\partial(\rho s)}{\partial(\rho v)}, \frac{\partial(\rho s)}{\partial E}\right),$$

where  $\rho s$  is considered as a function of the conservative variables  $\mathbf{U}$ ,

$$\rho s(\mathbf{U}) = \rho s\left(\frac{1}{\rho}, \frac{E}{\rho} - \frac{|\rho \mathbf{u}|^2}{2\rho^2}\right),$$

we start from the relation (second law of thermodynamics)

$$Tds = d\varepsilon + p d\tau,$$

which gives  $\frac{\partial s(\tau, \varepsilon)}{\partial \varepsilon} = \frac{1}{T}$  and  $\frac{\partial s(\tau, \varepsilon)}{\partial \tau} = \frac{p}{T}$ . We then compute the partial derivatives using the chain rule; for instance,

$$\begin{aligned} \frac{\partial(\rho s(\mathbf{U}))}{\partial \rho} &= s + \rho \frac{\partial s(\tau, \varepsilon)}{\partial \tau} \left(-\frac{1}{\rho^2}\right) + \frac{\partial s(\tau, \varepsilon)}{\partial \varepsilon} \left(-\frac{E}{\rho^2} + \frac{|\mathbf{u}|^2}{\rho^3}\right) \\ &= s + T^{-1}\left(-\frac{p}{\rho} - \frac{E}{\rho} + |\mathbf{u}|^2\right) = T^{-1}\left(-h + Ts + \frac{|\mathbf{u}|^2}{2}\right), \end{aligned}$$

where the specific entropy  $h = \frac{p}{\rho} + \varepsilon$  was introduced in Chap. IV, Sect. 3, and so on. This gives (2.4).  $\square$

Notice that setting

$$\mu = h - Ts = \varepsilon + p\tau - Ts,$$

we can also write (2.4) as

$$\mathbf{V}^T = T^{-1} \left( \mu - \frac{|\mathbf{u}|^2}{2}, \mathbf{u}, -1 \right).$$

We have, moreover, the relation

$$Td(\rho s) = d(\rho \varepsilon) - \mu d\rho,$$

i.e.,  $\mu = -\frac{\partial(\rho s)(\rho, \rho \varepsilon)}{\partial \rho}$  ( $\mu$  is the Gibbs potential). For a polytropic ideal gas,  $T = \frac{\varepsilon}{C_v}$ ,  $\gamma = \frac{C_p}{C_v}$ ,  $s = C_v \log \varepsilon \tau^{(\gamma-1)}$  up to an additive constant,  $h = C_p T$ , and  $\mu = T(C_p - s)$ .

For the sake of completeness, let us give the expression of the conjugate (or polar) function,  $\mathcal{U}^*$  (resp.  $\mathcal{F}^*$ ) of the entropy  $\mathcal{U}$  (resp.  $\mathcal{F}$ ) defined by

$$\mathcal{U}^*(\mathbf{V}) = \mathbf{V}^T \mathbf{U}(\mathbf{V}) - \mathcal{U}(\mathbf{U}(\mathbf{V})) \quad (2.5a)$$

resp.

$$\mathcal{F}^*(\mathbf{V}) = \mathbf{V}^T \mathbf{F}(\mathbf{U}(\mathbf{V})) - \mathcal{F}(\mathbf{U}(\mathbf{V})), \quad (2.5b)$$

i.e.,

$$\mathcal{F} = (\mathcal{F}_i), \quad \mathcal{F}_i^* = \mathbf{V}^T \mathbf{f}_i - \mathcal{F}_i$$

( $\mathcal{U}^*$  and  $\mathcal{F}^*$  were defined in the proof of Theorem 5.2, Chap. I). The functions  $\mathcal{U}^*$  and  $\mathcal{F}^*$  satisfy

$$D_{\mathbf{v}} \mathcal{U}^* = \mathbf{V}^T \mathbf{U}'(\mathbf{V}) + \mathbf{U}(\mathbf{V}) - \mathcal{U}' \mathbf{U}'(\mathbf{V}) = \mathbf{U}(\mathbf{V})$$

and

$$D_{\mathbf{v}} \mathcal{F}^* = \mathbf{V}^T \mathbf{F}'(\mathbf{U}(\mathbf{V})) \mathbf{U}'(\mathbf{V}) + \mathbf{F}(\mathbf{U}(\mathbf{V})) - \mathcal{F}'(\mathbf{U}(\mathbf{V})) \mathbf{U}'(\mathbf{V}) = \mathbf{F}(\mathbf{U}(\mathbf{V}))$$

since  $\mathcal{F}' = \mathcal{U}' \mathbf{F}'$ .

In the case of the gas dynamics equations, the expressions for  $\mathcal{U}^*$  and  $\mathcal{F}^*$  are particularly simple.

*Lemma 2.2*

The functions  $\mathcal{U}^*$  and  $\mathcal{F}^*$  defined by (2.4) and (2.5) are given in the case of the gas dynamics equations by

$$\mathcal{U}^* = \frac{p}{T}, \quad \mathcal{F}^* = \frac{p\mathbf{u}}{T}.$$

*Proof.* The computations are very simple. First, we have by definition

$$\begin{aligned} \mathcal{U}^* &= T^{-1} \left( \mu - \frac{|\mathbf{u}|^2}{2}, \mathbf{u}, -1 \right) (\rho, \rho\mathbf{u}, E)^T + \rho s \\ &= T^{-1} \left\{ \rho \left( \varepsilon + \frac{p}{\rho} - Ts \right) - \rho \frac{|\mathbf{u}|^2}{2} + \rho |\mathbf{u}|^2 - E \right\} + \rho s = \frac{p}{T}. \end{aligned}$$

Similarly, we find

$$\mathcal{F}_1^* = T^{-1} \left( \mu - \frac{|\mathbf{u}|^2}{2}, u, v - 1 \right) (\rho u, \rho u^2 + p, \rho u v, E u + p u)^T + \rho s u = u \frac{p}{T}$$

and

$$\mathcal{F}_2^* = v \frac{p}{T}.$$

We refer to Bourdel et al. [183, 858], Mazet and Bourdel [859], and Croisille [374] for more general results and use of these formulas.  $\square$

## 2.2 Invariance of the Euler Equations

Let us now study the invariance of the Euler equations through a Galilean transformation. We consider a frame  $\mathcal{R}'$  moving with uniform speed  $\mathbf{v}$  w.r.t. the reference frame  $\mathcal{R}$  of  $\mathbb{R}^2 \times \mathbb{R}_+$ ; thus, the new independent variables in  $\mathcal{R}'$  are

$$\mathbf{x}' = \mathbf{x} - \mathbf{v}t, \quad t' = t.$$

Looking at the dependent variables in the frame  $\mathcal{R}'$ , the density  $\rho$ , pressure  $p$ , and internal energy  $\varepsilon$  are invariant,

$$\rho' = \rho, \quad \varepsilon' = \varepsilon, \quad p' = p,$$

while the velocity  $\mathbf{u}$  is transformed into

$$\mathbf{u}' = \mathbf{u} - \mathbf{v}.$$

The expression “Galilean invariance” means that setting

$$\mathbf{U}' = (\rho', \rho' \mathbf{u}', E'),$$

$\mathbf{U}'$  satisfies the Euler equations

$$\frac{\partial \mathbf{U}'}{\partial t'} + \frac{\partial \mathbf{f}(\mathbf{U}')}{\partial x'} + \frac{\partial \mathbf{g}(\mathbf{U}')}{\partial y'} = \mathbf{0}$$

with the same functions  $\mathbf{f}$  and  $\mathbf{g}$  as in (2.2), i.e.,

$$\mathbf{f}(\mathbf{U}') = (\rho' u', \rho' u'^2 + p', \rho' u' v', (\rho' e' + p') u')^T,$$

with the analog for  $\mathbf{g}(\mathbf{U}')$ .

*Remark 2.1.* Assume a general pressure law

$$p = p(\mathbf{U}) = p'(\mathbf{U}').$$

Then, Galilean invariance requires in particular that the functions  $p$  and  $p'$  coincide. In fact, this presupposes a pressure law of the form  $p = p(\rho, \rho\varepsilon)$ .

Indeed,  $p = p'$  implies, setting  $\mathbf{m} = \rho\mathbf{u}$ ,  $E = \rho\varepsilon + \frac{\rho|\mathbf{u}|^2}{2}$ ,

$$p(\rho, \mathbf{m}, E) = p\left(\rho, \mathbf{m} - \rho\mathbf{v}, E + \rho\frac{1}{2}|\mathbf{v}|^2 - \rho\mathbf{v}^T\mathbf{u}\right);$$

differentiating this identity w.r.t.  $\mathbf{v}$  gives

$$0 = -\rho\frac{\partial p}{\partial m_i} + \rho(v_i - u_i)\frac{\partial p}{\partial E}, \quad 1 \leq i \leq 2$$

(for ease of notation, we have set  $\mathbf{u} = (u_i)$ ), and at  $\mathbf{v} = \mathbf{0}$  it yields

$$\rho\frac{\partial p}{\partial m_i} + m_i\frac{\partial p}{\partial E} = 0.$$

Integrating this differential system, we obtain that  $p$  is a function of the form

$$p(\rho, \mathbf{m}, E) = p\left(\rho, E - \frac{|\mathbf{m}|^2}{2\rho}\right) = p(\rho, \rho\varepsilon),$$

which gives the result.  $\square$

*Lemma 2.3*

*Assuming an equation of state of the form*

$$p = p(\rho, \rho\varepsilon), \tag{2.6}$$

*the Euler equations (2.1) are invariant under Galilean transformations.*

*Proof.* The invariance of the equations is easily established using the rules

$$\begin{aligned} \text{grad}_{\mathbf{x}'} &= \text{grad}_{\mathbf{x}} \\ \frac{\partial}{\partial t'} &= \frac{\partial}{\partial t} + \mathbf{v} \cdot \text{grad}_{\mathbf{x}} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y}. \end{aligned}$$

Let us check, for instance, the first equation:

$$\begin{aligned} \frac{\partial \rho}{\partial t'} + \text{div}_{\mathbf{x}'} \rho \mathbf{u}' &= \frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \text{grad}_{\mathbf{x}} \rho + \text{div}_{\mathbf{x}} \rho (\mathbf{u} - \mathbf{v}) \\ &= \frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \text{grad}_{\mathbf{x}} \rho + \text{div}_{\mathbf{x}} \rho \mathbf{u} - \mathbf{v} \cdot \text{grad}_{\mathbf{x}} \rho \\ &= \frac{\partial \rho}{\partial t} + \text{div}_{\mathbf{x}} \rho \mathbf{u} = 0. \end{aligned}$$

The other computations are similar.  $\square$

*Remark 2.2.* We have chosen to write the equation of state in the form (2.6) rather than in the usual form  $p = p(\rho, \varepsilon)$  because of Remark 2.1; more-

over, as we have already seen, it leads to simpler algebraic computations (see Chap. IV, Sect. 4.2).  $\square$

We come now to the rotational invariance of the Euler equations which is used extensively in the numerical schemes. Let us first specify some notations. A rotation of angle  $\theta$  in the  $(x, y)$ -plane corresponds to the matrix

$$R = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}. \quad (2.7)$$

Setting

$$\mathbf{X} = (x, y)^T = R\tilde{\mathbf{X}}, \quad \tilde{\mathbf{X}} = (\zeta, \tau)^T = (\mathbf{X} \cdot \mathbf{n}, \mathbf{X} \cdot \mathbf{n}^\perp)^T,$$

and

$$\mathbf{n} = R\mathbf{e}_1 = (\cos \theta, \sin \theta)^T, \quad \mathbf{n}^\perp = R\mathbf{e}_2 = (-\sin \theta, \cos \theta)^T, \quad (2.8)$$

we see geometrically that  $(\mathbf{n}, \mathbf{n}^\perp)$  are the new basis vectors obtained from  $(\mathbf{e}_1, \mathbf{e}_2)$  by the rotation  $R$ . If  $M$  is a point with coordinates  $(x, y)$  (in the basis  $(\mathbf{e}_1, \mathbf{e}_2)$ ), then  $(\zeta, \tau)$  are the coordinates of  $M$  in the basis  $(\mathbf{n}, \mathbf{n}^\perp)$ .

Similarly, we can write

$$\mathbf{u} = u\mathbf{e}_1 + v\mathbf{e}_2 = (\mathbf{u} \cdot \mathbf{n})\mathbf{n} + (\mathbf{u} \cdot \mathbf{n}^\perp)\mathbf{n}^\perp,$$

and the vector

$$\tilde{\mathbf{u}} = (u_n, u_\tau)^T = (\mathbf{u} \cdot \mathbf{n}, \mathbf{u} \cdot \mathbf{n}^\perp)^T \quad (2.9)$$

gives the coefficients of the velocity vector in the basis  $(\mathbf{n}, \mathbf{n}^\perp)$ , which we can write precisely as

$$\tilde{\mathbf{u}} = R^{-1}\mathbf{u} \circ R, \quad \text{i.e., } \tilde{\mathbf{u}}(\tilde{\mathbf{X}}) = \tilde{\mathbf{u}}(R^{-1}\mathbf{X}) = R^{-1}\mathbf{u}(\mathbf{X}) = R^{-1}\mathbf{u}(R\tilde{\mathbf{X}}),$$

or algebraically (equality of matrices),

$$(u_n, u_\tau)^T = \tilde{\mathbf{u}} = R^{-1}\mathbf{u} = R^{-1}(u, v)^T.$$

Now, since  $\rho, \varepsilon$ , and  $e$  are not changed, we define

$$\tilde{\mathbf{U}}(\zeta, \tau, t) = (\rho, \rho\tilde{\mathbf{u}}, \rho e)^T,$$

and the invariance means that  $\tilde{\mathbf{U}}$  is a solution of the Euler equations, i.e., we have

$$\frac{\partial \tilde{\mathbf{U}}}{\partial t} + \frac{\partial \mathbf{f}(\tilde{\mathbf{U}})}{\partial \zeta} + \frac{\partial \mathbf{g}(\tilde{\mathbf{U}})}{\partial \tau} = \mathbf{0}$$

with the same functions  $\mathbf{f}$  and  $\mathbf{g}$  as in (2.2),

$$\begin{aligned}\mathbf{f}(\tilde{\mathbf{U}}) &= (\rho u_n, \rho(u_n)^2 + p, \rho u_n u_\tau, (\rho e + p)u_n))^T, \\ \mathbf{g}(\tilde{\mathbf{U}}) &= (\rho u_\tau, \rho u_n u_\tau, \rho(u_\tau)^2 + p, (\rho e + p)u_\tau))^T.\end{aligned}$$

*Lemma 2.4*

The Euler equations (2.1) are invariant under rotation.

*Proof.* We consider a rotation with angle  $\theta$  in the  $(x, y)$ -plane, corresponding to the matrix (2.7), and we use the notations (2.8), (2.9). The computations can be put in a rather compact form using matrix notations. For the reader's convenience, we present a detailed proof. Since

$$\tilde{\mathbf{X}} = (\zeta, \tau)^T = R^{-1}\mathbf{X},$$

setting

$$\nabla = \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)^T, \quad \tilde{\nabla} = \left( \frac{\partial}{\partial \zeta}, \frac{\partial}{\partial \tau} \right)^T,$$

we have

$$\tilde{\nabla} = R^{-1}\nabla, \tag{2.10a}$$

which implies

$$\tilde{\nabla} \cdot \tilde{\mathbf{u}} = \nabla \cdot \mathbf{u}, \tag{2.10b}$$

i.e.,

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = \operatorname{div}_{\mathbf{X}} \mathbf{u} = \nabla \cdot \mathbf{u} = R\tilde{\nabla} \cdot R\tilde{\mathbf{u}} = \tilde{\nabla} \cdot \tilde{\mathbf{u}} = \operatorname{div}_{\tilde{\mathbf{X}}} \tilde{\mathbf{u}} = \frac{\partial u_n}{\partial \zeta} + \frac{\partial u_\tau}{\partial \tau}$$

(the dot “.” means the scalar product of vectors in  $\mathbb{R}^2$ , which is invariant under rotation, i.e.,  $R\mathbf{a} \cdot R\mathbf{b} = \mathbf{a} \cdot \mathbf{b}$ ). This gives the invariance of the first equation of (2.1) and also of the last since  $|\mathbf{u}| = |\tilde{\mathbf{u}}|$  and  $\nabla \cdot p\mathbf{u} = \tilde{\nabla} \cdot p\tilde{\mathbf{u}}$ .

Consider now the second and third equations of the system (2.1), which we can write

$$\frac{\partial}{\partial t}(\rho\mathbf{u})(\mathbf{X}, t) + \nabla \cdot (\rho u\mathbf{u} + p\mathbf{e}_1, \rho v\mathbf{u} + p\mathbf{e}_2)^T,$$

where we use the notation

$$\nabla \cdot (\mathbf{v}, \mathbf{w})^T = \frac{\partial \mathbf{v}}{\partial x} + \frac{\partial \mathbf{w}}{\partial y}. \tag{2.11}$$

Multiplying on the left by  $R^{-1}$ , we get

$$\frac{\partial}{\partial t}(\rho\tilde{\mathbf{u}})(\tilde{\mathbf{X}}, t) + R^{-1}\nabla \cdot (\rho u\mathbf{u} + p\mathbf{e}_1, \rho v\mathbf{u} + p\mathbf{e}_2)^T.$$

Now, using the obvious identities

$$\begin{aligned} R^{-1} \nabla \cdot (\mathbf{v}, \mathbf{w})^T &= R^{-1} \left( \frac{\partial \mathbf{v}}{\partial x} + \frac{\partial \mathbf{w}}{\partial y} \right) \\ &= \left( \frac{\partial R^{-1} v}{\partial x} + \frac{\partial R^{-1} w}{\partial y} \right) = \nabla \cdot (R^{-1} \mathbf{v}, R^{-1} \mathbf{w})^T \end{aligned}$$

and

$$\begin{aligned} \nabla \cdot (\mathbf{v} + \mathbf{v}', \mathbf{w} + \mathbf{w}')^T &= \frac{\partial(\mathbf{v} + \mathbf{v}')}{\partial x} + \frac{\partial(\mathbf{w} + \mathbf{w}')}{\partial x} \\ &= \nabla \cdot (\mathbf{v}, \mathbf{w})^T + \nabla \cdot (\mathbf{v}', \mathbf{w}')^T, \end{aligned}$$

we get

$$\begin{aligned} R^{-1} \nabla \cdot (\rho u \mathbf{u} + p \mathbf{e}_1, \rho v \mathbf{u} + p \mathbf{e}_2)^T &= \nabla \cdot (R^{-1}(\rho u \mathbf{u} + p \mathbf{e}_1), R^{-1}(\rho v \mathbf{u} + p \mathbf{e}_2))^T \\ &= \nabla \cdot (\rho u R^{-1} \mathbf{u}, \rho v R^{-1} \mathbf{u})^T + R^{-1} \nabla \cdot (p \mathbf{e}_1, p \mathbf{e}_2)^T. \end{aligned}$$

Using again the invariance of the scalar product in  $\mathbb{R}^2$  under rotation, we have

$$\nabla \cdot (\rho u \tilde{\mathbf{u}}, \rho v \tilde{\mathbf{u}})^T = \tilde{\nabla} \cdot (\rho u_n \tilde{\mathbf{u}}, \rho u_\tau \tilde{\mathbf{u}})^T,$$

which yields

$$\begin{aligned} R^{-1} \nabla \cdot (\rho u \mathbf{u} + p \mathbf{e}_1, \rho v \mathbf{u} + p \mathbf{e}_2)^T &= \tilde{\nabla} \cdot (\rho u_n \tilde{\mathbf{u}}, \rho u_\tau \tilde{\mathbf{u}}) + \tilde{\nabla} \cdot (p \mathbf{e}_1, p \mathbf{e}_2)^T \\ &= \tilde{\nabla} \cdot (\rho u_n \tilde{\mathbf{u}} + p \mathbf{e}_1, \rho u_\tau \tilde{\mathbf{u}} + p \mathbf{e}_2) \end{aligned}$$

and proves the invariance.

Let us extend the definition of  $R^{-1}$  on vectors of  $\mathbb{R}^4 = \mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}$  as the transformation

$$R^{-1}(a, \mathbf{b}, c) = (a, R^{-1}\mathbf{b}, c). \quad (2.12)$$

Thus,

$$R^{-1} \mathbf{U} = (\rho, \rho \tilde{\mathbf{u}}, \rho e)^T = \tilde{\mathbf{U}}.$$

To prove the rotational invariance of the Euler equations, we have checked that

$$R^{-1} \nabla \cdot \mathbf{F}(\mathbf{U}) = \tilde{\nabla} \cdot \mathbf{F}(\tilde{\mathbf{U}}), \quad \text{where } \mathbf{F} = (\mathbf{f}, \mathbf{g})^T$$

with the notations (2.10), (2.11).  $\square$

*Remark 2.3.* We can choose to expand the above identities. Let us check, for instance, that

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u_n)}{\partial \zeta} + \frac{\partial(\rho u_\tau)}{\partial \tau} = 0.$$

Since  $\mathbf{X}' = R^{-1} \mathbf{X}$ , we have

$$\frac{\partial}{\partial \zeta} = \cos \theta \frac{\partial}{\partial x} + \sin \theta \frac{\partial}{\partial y}, \quad \frac{\partial}{\partial \tau} = -\sin \theta \frac{\partial}{\partial x} + \cos \theta \frac{\partial}{\partial y},$$

and

$$\begin{aligned}\frac{\partial(\rho u_n)}{\partial \zeta} + \frac{\partial(\rho u_\tau)}{\partial \tau} &= \left( \cos \theta \frac{\partial}{\partial x} + \sin \theta \frac{\partial}{\partial y} \right) (\rho(u \cos \theta + v \sin \theta)) \\ &\quad + \left( -\sin \theta \frac{\partial}{\partial x} + \cos \theta \frac{\partial}{\partial y} \right) (\rho(-u \sin \theta + v \cos \theta)) \\ &= \frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y}.\end{aligned}$$

Consider now the second equation:

$$\begin{aligned}\frac{\partial}{\partial t}(\rho u_n) + \frac{\partial}{\partial \zeta}(\rho(u_n)^2 + p) + \frac{\partial}{\partial \tau}(\rho u_n u_\tau) &= \frac{\partial}{\partial t}(\rho(u \cos \theta + v \sin \theta)) \\ &\quad + \left( \cos \theta \frac{\partial}{\partial x} + \sin \theta \frac{\partial}{\partial y} \right) (\rho(u_n)^2 + p) - \left( \sin \theta \frac{\partial}{\partial x} - \cos \theta \frac{\partial}{\partial y} \right) (\rho u_n u_\tau) \\ &= \frac{\partial}{\partial t}(\rho(u \cos \theta + v \sin \theta)) + \frac{\partial}{\partial x} \rho u_n (u_n \cos \theta - u_\tau \sin \theta) \\ &\quad + \frac{\partial}{\partial y} \rho u_n (u_n \sin \theta + u_\tau \cos \theta) + \left( \cos \theta \frac{\partial p}{\partial x} + \sin \theta \frac{\partial p}{\partial y} \right) \\ &= \frac{\partial}{\partial t}(\rho(u \cos \theta + v \sin \theta)) + \frac{\partial}{\partial x} \rho(u \cos \theta + v \sin \theta) u \\ &\quad + \frac{\partial}{\partial y} \rho(u \cos \theta + v \sin \theta) v + \left( \cos \theta \frac{\partial p}{\partial x} + \sin \theta \frac{\partial p}{\partial y} \right).\end{aligned}$$

We obtain a combination of the second and third equations in (2.2),

$$\begin{aligned}\cos \theta \left\{ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho u v) \right\} \\ + \sin \theta \left\{ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho u v) + \frac{\partial}{\partial y}(\rho v^2 + p) \right\} = 0.\end{aligned}$$

The other computations are similar.

We can also use the notation of the projected equations introduced in (1.5) and (1.9), i.e., in dimension  $d = 2$ ,

$$\begin{aligned}\frac{\partial}{\partial \zeta} &= \cos \theta \frac{\partial}{\partial x} + \sin \theta \frac{\partial}{\partial y} = \mathbf{n} \cdot \nabla, \\ \frac{\partial}{\partial \tau} &= -\sin \theta \frac{\partial}{\partial x} + \cos \theta \frac{\partial}{\partial y} = \mathbf{n}^\perp \cdot \nabla.\end{aligned}$$

As we have already seen in (1.9), we can write

$$\nabla \cdot \mathbf{u} = (\mathbf{n} \cdot \nabla)(\mathbf{n} \cdot \mathbf{u}) + (\mathbf{n}^\perp \cdot \nabla)(\mathbf{n}^\perp \cdot \mathbf{u}) = \tilde{\nabla} \cdot \tilde{\mathbf{u}},$$

which gives (2.10b). Also, for  $\mathbb{F} = (\mathbf{f}, \mathbf{g})^T$ ,

$$\begin{aligned}\operatorname{div}_{\mathbf{X}} \cdot \mathbb{F}(\mathbf{U}(\mathbf{X}, t)) &= \nabla \cdot \mathbb{F} = (\mathbf{n} \cdot \nabla)(\mathbf{n} \cdot \mathbb{F}) + (\mathbf{n}^\perp \cdot \nabla)(\mathbf{n}^\perp \cdot \mathbb{F}) \\ &= \tilde{\nabla}(\mathbf{n} \cdot \mathbb{F}^T, \mathbf{n}^\perp \cdot \mathbb{F}^T)^T,\end{aligned}$$

where

$$\mathbf{n} \cdot \mathbb{F} = \cos \theta \mathbf{f} + \sin \theta \mathbf{g}, \quad \mathbf{n}^\perp \cdot \mathbb{F} = -\sin \theta \mathbf{f} + \cos \theta \mathbf{g}.$$

Thus, for any system  $\frac{\partial}{\partial t} \mathbf{U} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{U}) + \frac{\partial}{\partial y} \mathbf{g}(\mathbf{U}) = \mathbf{0}$ , we have

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial(\mathbf{n} \cdot \mathbb{F})}{\partial \zeta} + \frac{\partial(\mathbf{n}^\perp \cdot \mathbb{F})}{\partial \tau} = \mathbf{0}.$$

We now focus on the corresponding system for the Euler equations. Consider the second and third equations, which we can write

$$\frac{\partial}{\partial t}(\rho \mathbf{u}) + \frac{\partial}{\partial \zeta}(\mathbf{n} \cdot \mathbf{F}_{23}) + \frac{\partial}{\partial \tau}(\mathbf{n}^\perp \cdot \mathbf{F}_{23}) = \mathbf{0}$$

with  $\mathbf{F}_{23} = (\mathbf{f}_{23}, \mathbf{g}_{23})$ ,  $\mathbf{f}_{23} = (f_2, f_3)$ , i.e.,

$$\begin{aligned}\mathbf{n} \cdot \mathbf{F}_{23} &= \mathbf{n} \cdot (\rho u \mathbf{u} + p \mathbf{e}_1, \rho v \mathbf{u} + p \mathbf{e}_2)^T \\ &= \cos \theta (\rho u \mathbf{u} + p \mathbf{e}_1) + \sin \theta (\rho v \mathbf{u} + p \mathbf{e}_2), \\ \mathbf{n}^\perp \cdot \mathbf{F}_{23} &= \mathbf{n}^\perp (\rho u \mathbf{u} + p \mathbf{e}_1, \rho v \mathbf{u} + p \mathbf{e}_2)^T \\ &= -\sin \theta (\rho u \mathbf{u} + p \mathbf{e}_1) + \cos \theta (\rho v \mathbf{u} + p \mathbf{e}_2).\end{aligned}$$

In order to obtain the system in the dependent variables  $\tilde{\mathbf{U}}$ , we take the scalar product of this system of two equations with  $\mathbf{n}$  and  $\mathbf{n}^\perp$  (which corresponds to applying  $R^{-1}$ ),

$$\mathbf{n} \cdot \left\{ \frac{\partial}{\partial t}(\rho \mathbf{u})(\mathbf{X}, t) + \tilde{\nabla} \cdot (\mathbf{n} \cdot \mathbf{F}_{23}, \mathbf{F}_{23} \cdot \mathbf{n}^\perp)^T \right\} = 0, \quad (2.13a)$$

$$\mathbf{n}^\perp \cdot \left\{ \frac{\partial}{\partial t}(\rho \mathbf{u})(\mathbf{X}, t) + \tilde{\nabla} \cdot (\mathbf{n} \cdot \mathbf{F}_{23}, \mathbf{F}_{23} \cdot \mathbf{n}^\perp)^T \right\} = 0. \quad (2.13b)$$

We obtain for Eq. (2.13a)

$$\begin{aligned}\mathbf{n} \cdot \left\{ \frac{\partial}{\partial t}(\rho \mathbf{u}) + \tilde{\nabla} \cdot (\mathbf{n} \cdot \mathbf{F}_{23}, \mathbf{n}^\perp \cdot \mathbf{F}_{23})^T \right\} &= \frac{\partial}{\partial t}(\rho \mathbf{u} \cdot \mathbf{n}) \\ &+ \frac{\partial}{\partial \zeta} \mathbf{n} \cdot (\mathbf{n} \cdot \mathbf{F}_{23}) + \frac{\partial}{\partial \tau} \mathbf{n} \cdot (\mathbf{n}^\perp \cdot \mathbf{F}_{23})\end{aligned}$$

and

$$\begin{aligned}\mathbf{n} \cdot (\mathbf{F}_{23} \cdot \mathbf{n}) &= \mathbf{n} \cdot \{ \cos \theta (\rho u \mathbf{u} + p \mathbf{e}_1) + \sin \theta (\rho v \mathbf{u} + p \mathbf{e}_2) \} \\ &= \cos \theta (\rho u \mathbf{u} \cdot \mathbf{n} + p \mathbf{e}_1 \cdot \mathbf{n}) + \sin \theta (\rho v \mathbf{u} \cdot \mathbf{n} + p \mathbf{e}_2 \cdot \mathbf{n}),\end{aligned}$$

which is the scalar product with  $\mathbf{n}$  of the vector  $\rho \mathbf{u}(\mathbf{u} \cdot \mathbf{n}) + p \mathbf{n}$  with component

$$(\rho u \mathbf{u} \cdot \mathbf{n} + p \mathbf{e}_1 \cdot \mathbf{n}, \rho v \mathbf{u} \cdot \mathbf{n} + p \mathbf{e}_2 \cdot \mathbf{n}),$$

and this scalar product gives

$$(\rho \mathbf{u}(\mathbf{u} \cdot \mathbf{n}) + p \mathbf{n}) \cdot \mathbf{n} = \rho u_n^2 + p = f_2(\tilde{\mathbf{U}}).$$

Similarly,

$$\begin{aligned} \mathbf{n} \cdot (\mathbf{F}_{23} \cdot \mathbf{n}^\perp) &= \mathbf{n} \cdot \{-\sin \theta(\rho u \mathbf{u} + p \mathbf{e}_1) + \cos \theta(\rho v \mathbf{u} + p \mathbf{e}_2)\} \\ &= (-\sin \theta(\rho u \mathbf{u} \cdot \mathbf{n} + p \mathbf{e}_1 \cdot \mathbf{n} + \cos \theta(\rho v \mathbf{u} \cdot \mathbf{n} + p \mathbf{e}_2 \cdot \mathbf{n})) \end{aligned}$$

is the scalar product with  $\mathbf{n}^\perp$  of the vector  $\rho \mathbf{u}(\mathbf{u} \cdot \mathbf{n}) + p \mathbf{n}$  and

$$(\rho \mathbf{u}(\mathbf{u} \cdot \mathbf{n}) + p \mathbf{n}) \cdot \mathbf{n}^\perp = \rho u_n u_\tau = g_2(\tilde{\mathbf{U}}).$$

Hence, we have

$$\frac{\partial}{\partial \zeta} \mathbf{n} \cdot (\mathbf{n} \cdot \mathbf{F}_{23}) + \frac{\partial}{\partial \tau} \mathbf{n} \cdot (\mathbf{n}^\perp \cdot \mathbf{F}_{23}) = \frac{\partial}{\partial \zeta} (\rho u_n^2) + \frac{\partial}{\partial \tau} (\rho u_\tau u_n).$$

The computations for Eq. (2.13b) are similar:

$$\begin{aligned} \mathbf{n}^\perp \cdot \left\{ \frac{\partial}{\partial t} (\rho \mathbf{u})(\mathbf{X}, t) + \tilde{\nabla} \cdot (\mathbf{n} \cdot \mathbf{F}_{23}, \mathbf{F}_{23} \cdot \mathbf{n}^\perp)^T \right\} \\ = \frac{\partial}{\partial t} (\rho u_\tau) + \frac{\partial}{\partial \zeta} (\rho u_n u_\tau) + \frac{\partial}{\partial \tau} (\rho u_\tau^2 + p). \end{aligned}$$

We can check easily that, using (2.12), we have also proven

$$R^{-1} \mathbf{n} \cdot \mathbb{F}(\mathbf{U}) = \mathbf{f}(\tilde{\mathbf{U}}), \quad R^{-1} \mathbf{n}^\perp \cdot \mathbb{F}(\mathbf{U}) = \mathbf{g}(\tilde{\mathbf{U}}), \quad (2.14a)$$

which implies in particular by differentiation

$$\mathbf{A}(\mathbf{U}, \mathbf{n}) = R \mathbf{A}(\tilde{\mathbf{U}}) R^{-1}, \quad \text{where } \mathbf{A}(\cdot) = \mathbf{f}'(\cdot), \quad (2.14b)$$

a formula that is used in numerical schemes.  $\square$

## 2.3 Eigenvalues

Let us now study the hyperbolicity of the system (2.2). As in Chap. II, Example 2.4, we use the nonconservative variables  $(\rho, u, v, s)$  to compute the eigenvalues of the matrix  $\mathbf{A}(\mathbf{U}, \boldsymbol{\omega})$ . From

$$T ds = d\varepsilon - \frac{p}{\rho^2} d\rho,$$

we get

$$\begin{aligned} T \left( \frac{\partial s}{\partial t} + u \frac{\partial s}{\partial x} + v \frac{\partial s}{\partial y} \right) &= \frac{\partial \varepsilon}{\partial t} + u \frac{\partial \varepsilon}{\partial x} + v \frac{\partial \varepsilon}{\partial y} \\ - \left( \frac{p}{\rho^2} \right) \left( \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + v \frac{\partial \rho}{\partial y} \right) &= 0, \end{aligned}$$

so that the system can be written in nonconservation form

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + \rho \frac{\partial u}{\partial x} + v \frac{\partial \rho}{\partial y} + \rho \frac{\partial v}{\partial y} = 0, \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + \left( \frac{1}{\rho} \right) \frac{\partial p}{\partial x} = 0, \\ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + \left( \frac{1}{\rho} \right) \frac{\partial p}{\partial y} = 0, \\ \frac{\partial s}{\partial t} + u \frac{\partial s}{\partial x} + v \frac{\partial s}{\partial y} = 0. \end{array} \right. \quad (2.15)$$

We take  $p = p(\rho, s)$  and define the local speed of sound  $c$  by

$$c^2 = \frac{\partial p}{\partial \rho}(\rho, s) = p_\rho.$$

For instance, for a polytropic ideal gas,  $p = A(s) \rho^\gamma$ ,  $c^2 = \gamma \frac{p}{\rho}$ . Denoting the velocity field by

$$\mathbf{u} = (u, v)^T,$$

the Mach number is defined by

$$M = \frac{|\mathbf{u}|}{c}.$$

Replacing in (2.15)  $\frac{\partial p}{\partial x}$  (resp.  $\frac{\partial p}{\partial y}$ ) by  $p_\rho \frac{\partial \rho}{\partial x} + p_s \frac{\partial s}{\partial x}$  (resp.  $p_\rho \frac{\partial \rho}{\partial y} + p_s \frac{\partial s}{\partial y}$ ), we compute the matrix  $\mathbf{A}(\mathbf{U}, \omega)$ ,

$$\mathbf{A}(\mathbf{U}, \omega) = \begin{pmatrix} \mathbf{u} \cdot \boldsymbol{\omega} & \rho \omega_1 & \rho \omega_2 & 0 \\ p_\rho \omega_1 / \rho & \mathbf{u} \cdot \boldsymbol{\omega} & 0 & p_s \omega_1 / \rho \\ p_\rho \omega_2 / \rho & 0 & \mathbf{u} \cdot \boldsymbol{\omega} & p_s \omega_2 / \rho \\ 0 & 0 & 0 & \mathbf{u} \cdot \boldsymbol{\omega} \end{pmatrix},$$

and the eigenvalues are easily found to be  $\mathbf{u} \cdot \boldsymbol{\omega}$  and  $\mathbf{u} \cdot \boldsymbol{\omega} \pm c|\boldsymbol{\omega}|$ . Thus, we have two simple eigenvalues, which for  $|\boldsymbol{\omega}| = 1$  are  $\lambda_1 = \mathbf{u} \cdot \boldsymbol{\omega} - c$  and  $\lambda_4 = \mathbf{u} \cdot \boldsymbol{\omega} + c$ , and give genuinely nonlinear fields (“acoustic” or “pressure” waves), and  $\lambda_2 = \lambda_3 = \mathbf{u} \cdot \boldsymbol{\omega}$  (of multiplicity 2) associated to the entropy waves and “vorticity” (or shear) waves, which are linearly degenerate. We can choose

$$\mathbf{r}_1 = \begin{pmatrix} -\rho/c \\ \omega_1 \\ \omega_2 \\ 0 \end{pmatrix}, \quad \mathbf{r}_4 = \begin{pmatrix} \rho/c \\ \omega_1 \\ \omega_2 \\ 0 \end{pmatrix}, \quad (2.16a)$$

while the eigenvectors associated to the eigenvalue  $\mathbf{u} \cdot \boldsymbol{\omega}$  may be taken as

$$\mathbf{r}_2 = \begin{pmatrix} p_s \\ 0 \\ 0 \\ -p_\rho \end{pmatrix}, \quad \mathbf{r}_3 = \begin{pmatrix} 0 \\ -\omega_2 \\ \omega_1 \\ 0 \end{pmatrix}. \quad (2.16b)$$

The vector  $\mathbf{r}_2(\mathbf{U}, \boldsymbol{\omega})$  does not depend on  $\boldsymbol{\omega}$ , and  $\mathbf{r}_3(\mathbf{U}, \boldsymbol{\omega}) = (0, \boldsymbol{\omega}^\perp, 0)$  does not depend on  $\mathbf{U}$ ; its form explains the term “shear wave” which we have used above (see Sect. 2.5 below; see also Roe [980] and Roe [982] and Rumsey et al. [996] for the use of shear waves in numerical schemes).

*Remark 2.4.* One can prove more generally that the multiple eigenvalues of a hyperbolic system give linearly degenerate fields, and therefore genuinely nonlinear fields are simple (see Chap. II, Remark 6.1).  $\square$

The 2,3-Riemann invariants in the direction  $\boldsymbol{\omega}$  are  $\mathbf{u} \cdot \boldsymbol{\omega}$  and  $p$  (since we have a system of four equations, and the multiplicity of the eigenvalue is 2, we can find only  $4 - 2 = 2$  Riemann invariants whose gradients  $Dw(\mathbf{U})$  satisfying  $Dw(\mathbf{U}) \cdot \mathbf{r}_i(\mathbf{U}, \boldsymbol{\omega}) = 0$ ,  $i = 2, 3$ , are independent; see Chap. II, Sect. 3.2). The three 1- (resp. 4-) Riemann invariants are  $u + \ell\omega_1, v + \ell\omega_2, s$  (resp.  $u - \ell\omega_1, v - \ell\omega_2, s$ ), where  $\ell(\rho, s)$  is the function defined (in Chap. II, formula (3.18)) by

$$\frac{\partial \ell}{\partial \rho} = \frac{c}{\rho}.$$

We could as well take  $\mathbf{u} \cdot \boldsymbol{\omega} \pm \ell, \mathbf{u} \cdot \boldsymbol{\omega}^\perp, s$ .

*Remark 2.5.* Let us see why, as is well known, the simple eigenvalues  $\mathbf{u} \cdot \boldsymbol{\omega} \pm c$  are associated to the sound (or acoustic) waves (see Courant and Friedrichs, Section 11 [371], Chorin and Marsden, Section 3.1 [305], J.D. Anderson, Section 7.5 [40], Whitham, Section 6.6 [1188]). Indeed, consider a small, smooth (therefore isentropic as we shall see in (2.20) below) perturbation

$$\rho = \rho_0 + \delta\rho_1,$$

where  $\rho_0$  is constant, and take for simplicity  $\mathbf{u}_0 = 0$ , so that  $\mathbf{u} = \delta\mathbf{u}_1$  is small. Since  $ds = 0$ , we can write

$$dp = \frac{\partial p}{\partial \rho}(\rho, s)d\rho = c^2 d\rho,$$

which yields

$$\frac{\partial p}{\partial x} = c^2 \frac{\partial \rho}{\partial x}, \quad \frac{\partial p}{\partial y} = c^2 \frac{\partial \rho}{\partial y}.$$

Substituting these expressions in the equations (2.15), and neglecting terms of order higher than 1 in  $\delta$ , we obtain

$$\begin{cases} \frac{\partial \rho_1}{\partial t} + \rho_0 \left( \frac{\partial u_1}{\partial x} + \frac{\partial v_1}{\partial y} \right) = 0, \\ \frac{\partial u_1}{\partial t} + \left( \frac{1}{\rho_0} \right) \frac{\partial p}{\partial \rho}(\rho, s) \frac{\partial \rho_1}{\partial x} = 0, \\ \frac{\partial v_1}{\partial t} + \left( \frac{1}{\rho_0} \right) \frac{\partial p}{\partial \rho}(\rho, s) \frac{\partial \rho_1}{\partial y} = 0. \end{cases}$$

Now, also expanding  $\frac{\partial p}{\partial \rho}(\rho, s) = c^2$  about  $\rho_0$ ,

$$\frac{\partial p}{\partial \rho}(\rho, s) = c_0^2 + O(\delta),$$

we differentiate the first equation (resp. the second and third) w.r.t.  $t$  (resp. w.r.t.  $x$  and  $y$ ) and obtain that the disturbance  $\rho_1$  satisfies the wave equation associated to the velocity  $c_0$ ,

$$\frac{\partial^2(\rho_1)}{\partial t^2} = c_0^2 \left( \frac{\partial^2 \rho_1}{\partial x^2} + \frac{\partial^2 \rho_1}{\partial y^2} \right) = c_0^2 \Delta \rho_1.$$

In the one-dimensional case, the general solution is a function of the form  $f(x + c_0 t) + g(x - c_0 t)$ , where  $f$  and  $g$  are arbitrary, i.e., a superposition of two waves traveling with constant speed  $\pm c_0$ ; small disturbances propagate with speed  $c_0$ .  $\square$

*Remark 2.6.* In order to compute the characteristic fields, we might also have used the primitive variables  $(\rho, u, v, p)$ . From

$$dp = c^2 d\rho + \frac{\partial p}{\partial s} ds,$$

we get from the first and last equations (2.1),

$$\begin{aligned} \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + v \frac{\partial p}{\partial y} &= c^2 \frac{\partial \rho}{\partial t} + \frac{\partial p}{\partial s} \frac{\partial s}{\partial t} \\ &\quad + u \left\{ c^2 \frac{\partial \rho}{\partial x} + \frac{\partial p}{\partial s} \frac{\partial s}{\partial x} \right\} + v \left\{ c^2 \frac{\partial \rho}{\partial y} + \frac{\partial p}{\partial s} \frac{\partial s}{\partial y} \right\} \\ &= c^2 \left\{ \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + v \frac{\partial \rho}{\partial y} \right\}. \end{aligned}$$

Setting  $\text{grad } p = \nabla p = (\frac{\partial p}{\partial x}, \frac{\partial p}{\partial y})^T$ , this yields

$$\frac{\partial p}{\partial t} + \mathbf{u} \cdot \text{grad } p - c^2 \left\{ \frac{\partial \rho}{\partial t} + \mathbf{u} \cdot \text{grad } \rho \right\} = 0. \quad (2.17)$$

Using the primitive variables  $(u, v, p)$ , the system can be written in the form

$$\left\{ \frac{\partial \rho}{\partial t} + \mathbf{u} \cdot \nabla \rho \right\} - c^{-2} \left\{ \frac{\partial p}{\partial t} + \mathbf{u} \cdot \nabla p \right\} = 0, \quad (2.18a)$$

$$\begin{cases} \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + \left( \frac{1}{\rho} \right) \frac{\partial p}{\partial x} = 0, \\ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + \left( \frac{1}{\rho} \right) \frac{\partial p}{\partial y} = 0, \\ \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + v \frac{\partial p}{\partial y} + \rho c^2 \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) = 0. \end{cases} \quad (2.18b)$$

We consider the nonconservative system (2.18b) of the last three equations in  $\mathbf{V} = (u, v, p)$ . The density is supposed to be determined from  $p$  through Eq. (2.18a), which holds along the particles' paths, as we shall see below. The corresponding matrix  $\mathbf{A}(\mathbf{V}, \boldsymbol{\omega}) = \mathbf{A}_1(\mathbf{V})\boldsymbol{\omega}_1 + \mathbf{A}_2(\mathbf{V})\boldsymbol{\omega}_2$  is then

$$\begin{pmatrix} \mathbf{u} \cdot \boldsymbol{\omega} & 0 & \omega_1/\rho \\ 0 & \mathbf{u} \cdot \boldsymbol{\omega} & \omega_2/\rho \\ \rho c^2 \omega_1 & \rho c^2 \omega_2 & \mathbf{u} \cdot \boldsymbol{\omega} \end{pmatrix},$$

whose (simple) eigenvalues are naturally  $\mathbf{u} \cdot \boldsymbol{\omega}$  and  $\mathbf{u} \cdot \boldsymbol{\omega} \pm c$ . The complete matrix associated to system (2.18b) completed by the equation

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + v \frac{\partial \rho}{\partial y} + \rho \frac{\partial u}{\partial x} + \rho \frac{\partial v}{\partial y} = 0$$

is

$$\begin{pmatrix} \mathbf{u} \cdot \boldsymbol{\omega} & \rho \omega_1 & \rho \omega_2 & 0 \\ 0 & \mathbf{u} \cdot \boldsymbol{\omega} & 0 & \omega_1/\rho \\ 0 & 0 & \mathbf{u} \cdot \boldsymbol{\omega} & \omega_2/\rho \\ 0 & \rho c^2 \omega_1 & \rho c^2 \omega_2 & \mathbf{u} \cdot \boldsymbol{\omega} \end{pmatrix},$$

and the eigenvectors are  $\mathbf{r}_1 = (1, -c \frac{\omega_1}{\rho}, -c \frac{\omega_2}{\rho}, c^2)^T$ ,  $\mathbf{r}_2 = (1, 0, 0, 0)^T$ ,  $\mathbf{r}_3 = (1, -\omega_2, \omega_1, 0)^T$ ,  $\mathbf{r}_4 = (1, c \frac{\omega_1}{\rho}, c \frac{\omega_2}{\rho}, c^2)^T$ .  $\square$

We shall need the expression for the eigenvectors in conservative variables in Sect. 4.3.3 (for Roe's scheme); setting  $H = \frac{(E+p)}{\rho}$ , which is the total enthalpy, we have

$$\mathbf{r}_1(\mathbf{U}, \boldsymbol{\omega}) = (1, u - c\omega_1, v - c\omega_2, H - \mathbf{u} \cdot \boldsymbol{\omega}c)^T,$$

$$\mathbf{r}_4(\mathbf{U}, \boldsymbol{\omega}) = (1, u + c\omega_1, v + c\omega_2, H + \mathbf{u} \cdot \boldsymbol{\omega}c)^T,$$

$$\mathbf{r}_2(\mathbf{U}, \boldsymbol{\omega}) = \left( 1, u, v \frac{|\mathbf{u}|^2}{2} \right)^T, \mathbf{r}_3(\mathbf{U}, \boldsymbol{\omega}) = (0, -\omega_2, \omega_1, \mathbf{u} \cdot \boldsymbol{\omega}^\perp)^T.$$

By rotational invariance, these values can also be obtained from the formula (2.14b),  $\mathbf{A}(\mathbf{U}, \boldsymbol{\omega}) = R\mathbf{A}(\tilde{\mathbf{U}})R^{-1}$ , where  $R$  is a rotation such that  $\boldsymbol{\omega} = R\mathbf{e}_1$ ,

so that we need only to compute the eigenvectors of  $\mathbf{f}' = \mathbf{A}$ . Note that the vector  $\mathbf{r}_2(\mathbf{U}, \boldsymbol{\omega})$  is an eigenvector of both  $\mathbf{f}'$  and  $\mathbf{g}'$ .

## 2.4 Characteristics

Let us introduce now the particle path or trajectory. It is an integral curve of the velocity field  $\mathbf{u} = (u, v)^T$ , i.e., a curve  $t \rightarrow \mathbf{x}(t) = (x(t), y(t))$ , parametrized by  $t$ , such that

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{u}(\mathbf{x}(t), t), \quad (2.19a)$$

i.e.,

$$\frac{dx}{dt} = u, \quad \frac{dy}{dt} = v.$$

One sometimes uses instead of “particle path” the term *streamline*; a streamline is a function  $s \rightarrow \mathbf{x}(s) = (x(s), y(s))$ , which is an integral curve at fixed time  $t$ ,

$$\frac{d}{ds}\mathbf{x}(s) = \mathbf{u}(\mathbf{x}(s), t).$$

A streamline coincides with a particle path for a stationary flow.

The particle derivative is the differential along the particle path,

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x} + v \frac{\partial}{\partial y}. \quad (2.19b)$$

In fact this corresponds to (1.12) with  $\boldsymbol{\nu} = \frac{\mathbf{u}}{|\mathbf{u}|}$  and  $\lambda_k(\mathbf{u}, \boldsymbol{\nu}) = \mathbf{u} \cdot \boldsymbol{\nu} = |\mathbf{u}|$ , i.e.,  $\lambda_k(\mathbf{u}, \boldsymbol{\nu})\boldsymbol{\nu} = \mathbf{u}$ , and the trajectories are the corresponding curves  $C_k$  (1.11) ( $k = 2$  or  $3$ ).

Note that with the notation  $\frac{D}{Dt}$ , the last equation (2.15) can be written

$$\frac{Ds}{Dt} = 0, \quad (2.20)$$

which means that the entropy is constant along the particle paths (in the smooth parts of the flow). Also, Eq. (2.17) can be written equivalently

$$\frac{Dp}{Dt} - c^2 \frac{D\rho}{Dt} = 0. \quad (2.21)$$

Recall that a characteristic surface  $\varphi(x, y, t) = 0$  is such that the matrix  $\frac{\partial \varphi}{\partial t} \mathbf{I} + \mathbf{f}'(\mathbf{U}) \frac{\partial \varphi}{\partial x} + \mathbf{g}'(\mathbf{U}) \frac{\partial \varphi}{\partial y}$  is singular, which, by Lemma 1.1, implies that  $-\frac{\partial \varphi}{\partial t}$  is an eigenvalue of  $\mathbf{A}(\mathbf{u}, \boldsymbol{\nu})$ ; for instance, the characteristic surfaces associated to the eigenvalue  $\lambda_4$  satisfy

$$-\frac{\partial \varphi}{\partial t} = \mathbf{u} \cdot \boldsymbol{\nu} + c,$$

where  $\boldsymbol{\nu} = (\frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y})$ , which gives the condition

$$\frac{\partial \varphi}{\partial t} + u \frac{\partial \varphi}{\partial x} + v \frac{\partial \varphi}{\partial y} + c = 0.$$

*Lemma 2.5*

The envelope of all the characteristic surfaces through a point  $(x_0, y_0, t_0)$  consists of the streamline through the point and a conoid.

*Proof.* Consider first the envelope of the tangent planes to the characteristic surfaces through the point  $M_0 = (t_0, x_0, y_0)$  (it corresponds to a linearization about the state  $\mathbf{U}(M_0)$ ). The equation of the tangent plane to a surface  $\varphi(t, x, y) = 0$  at the point  $M_0$  is

$$\frac{\partial \varphi}{\partial t}(M_0)(t - t_0) + \frac{\partial \varphi}{\partial x}(M_0)(x - x_0) + \frac{\partial \varphi}{\partial y}(M_0)(y - y_0) = 0.$$

If the characteristic surface is associated, say, to the eigenvalue  $\lambda_4$ , we add the condition

$$\frac{\partial \varphi}{\partial t}(M_0) + u_0 \frac{\partial \varphi}{\partial x}(M_0) + v_0 \frac{\partial \varphi}{\partial y}(M_0) + c(M_0) = 0,$$

together with the normalization

$$\left( \frac{\partial \varphi}{\partial x}(M_0) \right)^2 + \left( \frac{\partial \varphi}{\partial y}(M_0) \right)^2 = 1.$$

We can set

$$\frac{\partial \varphi}{\partial x}(M_0) = \cos \beta, \quad \frac{\partial \varphi}{\partial y}(M_0) = \sin \beta,$$

and we get

$$\frac{\partial \varphi}{\partial t}(M_0) + u_0 \cos \beta + v_0 \sin \beta + c_0 = 0.$$

Hence, the tangent plane satisfies

$$-(u_0 \cos \beta + v_0 \sin \beta + c_0)(t - t_0) + \cos \beta(x - x_0) + \sin \beta(y - y_0) = 0$$

or

$$\cos \beta(x - x_0 - u_0(t - t_0)) + \sin \beta(y - y_0 - v_0(t - t_0)) = c_0(t - t_0). \quad (2.22)$$

The envelope of this family of planes is derived by differentiating (2.22) w.r.t.  $\beta$ ,

$$-\sin \beta(x - x_0 - u_0(t - t_0)) + \cos \beta(y - y_0 - v_0(t - t_0)) = 0, \quad (2.23)$$

and we obtain from (2.22), (2.23)

$$\begin{cases} x - x_0 - u_0(t - t_0) = \cos \beta c_0(t - t_0), \\ y - y_0 - v_0(t - t_0) = \sin \beta c_0(t - t_0); \end{cases} \quad (2.24)$$

and finally, eliminating  $\beta$  between the two equations (2.24), we get the (sonic) cone through the point  $M_0 = (t_0, x_0, y_0)$ ,

$$(x - x_0 - u_0(t - t_0))^2 + (y - y_0 - v_0(t - t_0))^2 = (c_0(t - t_0))^2. \quad (2.25)$$

The intersection of the cone (2.25) with the tangent plane (2.22) is precisely the line (2.24), which is called a bicharacteristic (see Holt [628], for instance).

If we consider the envelope of the characteristic surfaces, we write similarly the identity

$$\frac{\partial \varphi}{\partial t} dt + \frac{\partial \varphi}{\partial x} dx + \frac{\partial \varphi}{\partial y} dy = 0,$$

and then the condition that the surfaces be characteristic can be written

$$\frac{\partial \varphi}{\partial t} + u \frac{\partial \varphi}{\partial x} + v \frac{\partial \varphi}{\partial y} + c = 0,$$

and the normalization identity

$$\nu_1 = \frac{\partial \varphi}{\partial x} = \cos \beta, \quad \nu_2 = \frac{\partial \varphi}{\partial y} = \sin \beta.$$

We have

$$\cos \beta \left( \frac{dx}{dt} - u \right) + \sin \beta \left( \frac{dy}{dt} - v \right) = c.$$

We differentiate w.r.t.  $\beta$ ,

$$-\sin \beta \left( \frac{dx}{dt} - u \right) + \cos \beta \left( \frac{dy}{dt} - v \right) = 0$$

and obtain

$$\begin{cases} \frac{dx}{dt} - u = c \cos \beta, \\ \frac{dy}{dt} - v = c \sin \beta. \end{cases} \quad (2.26)$$

Finally,

$$\left( \frac{dx}{dt} - u \right)^2 + \left( \frac{dy}{dt} - v \right)^2 = c^2. \quad (2.27)$$

The characteristic surface touches the envelope at the line of tangency (2.26), which is the integral curve of the vector field  $\mathbf{u} + c \boldsymbol{\nu}$  and is usually called a bicharacteristic; we shall note it  $B_k$ . If  $\Sigma$  is a characteristic surface, both

curves  $B_4(\frac{dx}{dt} = \mathbf{u} + c\boldsymbol{\nu})$  and  $C_4$  (defined by (1.11), i.e.,  $\frac{dx}{dt} = (\mathbf{u} \cdot \boldsymbol{\nu} + c)\boldsymbol{\nu}$ ), lie on  $\Sigma$ .

We have similar results with the other eigenvalues. Notice that the cone associated to  $\lambda_1$ , obtained by changing  $c$  to  $-c$ , coincides with (2.27). Now, if we take  $c = 0$ , corresponding to the double eigenvalue, the corresponding conoid (2.27) degenerates into the streamline.  $\square$

We have already obtained the characteristic equations (2.20), (2.21), which hold along the streamlines. The other characteristic equations are obtained following exactly the computations of Example 5.1, Chap. II.

Equation (2.21) together with the first equation in (2.15) gives

$$\frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + v \frac{\partial p}{\partial y} + \rho c^2 \left\{ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right\} = 0$$

or, with  $\operatorname{div} \mathbf{u} = \nabla \cdot \mathbf{u} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$ ,

$$\frac{\partial p}{\partial t} + \mathbf{u} \cdot \operatorname{grad} p + \rho c^2 \operatorname{div} \mathbf{u} = 0. \quad (2.28)$$

The 2nd and 3rd equations in (2.15) can be written with  $\operatorname{grad} = \nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})^T$ ,

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \operatorname{grad}) \mathbf{u} + \left( \frac{1}{\rho} \right) \operatorname{grad} p = \mathbf{0} \quad (2.29)$$

or, with the particle derivative  $\frac{D}{Dt}$  defined in (2.19b),

$$\frac{D\mathbf{u}}{Dt} + \left( \frac{1}{\rho} \right) \operatorname{grad} p = \mathbf{0}.$$

We can now take the scalar product of (2.29) by  $\boldsymbol{\mu}$ , where  $\boldsymbol{\mu}$  is any unit vector, to obtain the characteristic equation

$$\boldsymbol{\mu} \cdot \left\{ \frac{D\mathbf{u}}{Dt} + \left( \frac{1}{\rho} \right) \operatorname{grad} p \right\} = 0 \quad (2.30)$$

(the differentiation takes place in a plane tangent to the particle path and parallel to  $\boldsymbol{\mu}$ ).

If we multiply (2.30), written for some unit vector  $\boldsymbol{\omega}$ , by  $\pm \rho c$  and add it to (2.28), we get

$$\frac{\partial p}{\partial t} + \mathbf{u} \cdot \operatorname{grad} p + \rho c^2 \operatorname{div} \mathbf{u} \pm \rho c \boldsymbol{\omega} \frac{D\mathbf{u}}{Dt} \pm c \boldsymbol{\omega} \operatorname{grad} p = 0,$$

which yields

$$\left( \frac{D}{Dt} \pm c \boldsymbol{\omega} \cdot \operatorname{grad} \right) p \pm \rho c \left( \boldsymbol{\omega} \cdot \frac{D}{Dt} \pm c \operatorname{div} \right) \mathbf{u} = 0,$$

or

$$\left( \frac{D}{Dt} \pm c\boldsymbol{\omega} \cdot \text{grad} \right) p \pm \rho c \boldsymbol{\omega} \cdot \left( \frac{D}{Dt} \pm c\boldsymbol{\omega} \cdot \text{div} \right) \mathbf{u} = 0. \quad (2.31)$$

The characteristic equations are thus

$$\begin{cases} \frac{Dp}{Dt} - c^2 \frac{D\rho}{Dt} = 0 \quad (\text{can be equivalently replaced by } \frac{Ds}{Dt} = 0), \\ \boldsymbol{\mu} \cdot \left( \frac{D\mathbf{u}}{Dt} + \left( \frac{1}{\rho} \right) \text{grad } p \right) = 0, \\ \left( \frac{D}{Dt} + c\boldsymbol{\omega} \cdot \text{grad} \right) p + \rho c \boldsymbol{\omega} \cdot \left( \frac{D}{Dt} + c\boldsymbol{\omega} \cdot \text{div} \right) \mathbf{u} = 0, \\ \left( \frac{D}{Dt} - c\boldsymbol{\omega} \cdot \text{grad} \right) p + \rho c \boldsymbol{\omega} \cdot \left( \frac{D}{Dt} - c\boldsymbol{\omega} \cdot \text{div} \right) \mathbf{u} = 0. \end{cases} \quad (2.32)$$

It is easy to check that in (2.31) the differential operators inside each bracket all act in one plane. Indeed, we can write equation (2.31) using the decomposition  $\mathbf{u} = (\mathbf{u} \cdot \boldsymbol{\omega})\boldsymbol{\omega} + (\mathbf{u} \cdot \boldsymbol{\omega}^\perp)\boldsymbol{\omega}^\perp$  and  $\text{div } \mathbf{u} = (\boldsymbol{\omega} \cdot \text{grad})(\mathbf{u} \cdot \boldsymbol{\omega}) + (\boldsymbol{\omega}^\perp \cdot \text{grad})(\mathbf{u} \cdot \boldsymbol{\omega}^\perp)$  in the form

$$\begin{aligned} & \left\{ \frac{\partial}{\partial t} + (\mathbf{u} + c\boldsymbol{\omega}) \cdot \text{grad} \right\} p + \rho c \boldsymbol{\omega} \cdot \left\{ \frac{\partial}{\partial t} + \mathbf{u} \cdot \text{grad} + c\boldsymbol{\omega} \cdot \text{div} \right\} \mathbf{u} \\ &= \left\{ \frac{\partial}{\partial t} + (\mathbf{u} + c\boldsymbol{\omega}) \cdot \text{grad} \right\} p + \rho c \left\{ \frac{\partial}{\partial t} + (\mathbf{u} + c\boldsymbol{\omega}) \cdot \text{grad} \right\} (\mathbf{u} \cdot \boldsymbol{\omega}) \\ & \quad + \rho c^2 (\boldsymbol{\omega}^\perp \cdot \text{grad})(\mathbf{u} \cdot \boldsymbol{\omega}^\perp) = 0. \end{aligned}$$

The operator  $\left\{ \frac{\partial}{\partial t} + (\mathbf{u} + c\boldsymbol{\omega}) \cdot \text{grad} \right\}$  acts along an integral curve of the vector  $\mathbf{u} + c\boldsymbol{\omega}$ , i.e.,

$$\frac{dx}{dt} = u + c\omega_1, \quad \frac{dy}{dt} = u + c\omega_2,$$

which is a bicharacteristic  $B_4$  (see (2.26)). Thus in (2.31), the directions of differentiation lie in a plane passing through this bicharacteristic (tangent to the sonic cone) and through  $(0, \boldsymbol{\omega}^\perp)$  (i.e., the intersection with the plane  $t = 0$  is orthogonal to  $\boldsymbol{\omega}$ ).

*Remark 2.7.* We have derived some characteristic equations using the particular expression of the eigenvalues in terms of the velocity field. We might also consider the form mentioned in Lemma 1.2 with the notation (1.12),

$$\frac{d}{ds_4} = \frac{\partial}{\partial t} + (\mathbf{u} \cdot \boldsymbol{\omega} + c)\boldsymbol{\omega} \cdot \text{grad}.$$

We have, for instance,

$$\begin{aligned} \frac{\partial}{\partial t} + (\mathbf{u} + c\boldsymbol{\omega}) \cdot \text{grad} &= \frac{\partial}{\partial t} + (\mathbf{u} \cdot \boldsymbol{\omega} + c)\boldsymbol{\omega} \cdot \text{grad} + (\mathbf{u} \cdot \boldsymbol{\omega}^\perp)\boldsymbol{\omega}^\perp \cdot \text{grad} \\ &= \frac{d}{ds_4} + (\mathbf{u} \cdot \boldsymbol{\omega}^\perp)\boldsymbol{\omega}^\perp \cdot \text{grad}, \end{aligned}$$

and we get

$$\begin{aligned} & \frac{\partial p}{\partial t} + (\mathbf{u} + c\boldsymbol{\omega}) \cdot \operatorname{grad} p + \rho c \left\{ \frac{\partial}{\partial t} + (\mathbf{u} + c\boldsymbol{\omega}) \cdot \operatorname{grad} \right\} (\mathbf{u} \cdot \boldsymbol{\omega}) \\ & \quad + \rho c^2 (\boldsymbol{\omega}^\perp \cdot \operatorname{grad})(\mathbf{u} \cdot \boldsymbol{\omega}^\perp) \\ & = \frac{dp}{ds_4} + \rho c \frac{d}{ds_4} (\mathbf{u} \cdot \boldsymbol{\omega}) + (\mathbf{u} \cdot \boldsymbol{\omega}^\perp) \boldsymbol{\omega}^\perp \cdot \operatorname{grad} p \\ & \quad + \rho c (\mathbf{u} \cdot \boldsymbol{\omega}^\perp) \boldsymbol{\omega}^\perp \cdot \operatorname{grad}(\mathbf{u} \cdot \boldsymbol{\omega}) + \rho c^2 (\boldsymbol{\omega}^\perp \cdot \operatorname{grad})(\mathbf{u} \cdot \boldsymbol{\omega}^\perp) = 0. \end{aligned}$$

Thus, as we have already observed, the equation contains only derivatives in the directions  $(1, \lambda_k(\mathbf{u}, \boldsymbol{\omega})\boldsymbol{\omega})$  and  $(0, \boldsymbol{\omega}^\perp)$ .  $\square$

## 2.5 Plane Wave Solutions: Self-Similar Solutions

### 2.5.1 Simple Plane Waves and Contact Discontinuities

For what concerns simple plane wave solutions (see Chap. II, Sect. 3.2)

$$\mathbf{U}(\mathbf{x}, t) = \mathbf{V}(\varphi(\mathbf{x} \cdot \boldsymbol{\omega}, t)),$$

we are led to consider the integral curves (1.18) of the vector fields  $\mathbf{r}_k$  given by (2.16a)–(2.16b) in the system of variables  $\mathbf{U} = (\rho, u, v, s)^T$ .

For  $k = 1$  and  $k = 4$ , using the Riemann invariants given in Sect. 2.3, we obtain that a state  $\mathbf{U}$  belongs to the set  $\mathcal{R}_k(\mathbf{U}_L, \boldsymbol{\omega})$  if

$$\begin{cases} u + \ell(\rho, s)\omega_1 = u_L + \ell(\rho_L, s_L)\omega_1, \\ v + \ell(\rho, s)\omega_2 = v_L + \ell(\rho_L, s_L)\omega_2, \\ s = s_L, \end{cases} \quad (2.33)$$

and in particular that  $\mathbf{u} \cdot \boldsymbol{\omega}^\perp$  is continuous,

$$\mathbf{u} \cdot \boldsymbol{\omega}^\perp = \mathbf{u}_L \cdot \boldsymbol{\omega}^\perp.$$

For instance, for a perfect gas,  $\ell = \frac{2c}{(\gamma-1)}$ , (see Example 3.2, Chap. III). When  $\boldsymbol{\omega}$  varies in (2.33), we obtain the set

$$s = s_L, \quad (u - u_L)^2 + (v - v_L)^2 = (\ell(\rho_L, s_L) - \ell(\rho, s_L))^2.$$

For  $k = 2$  or  $3$ , since  $\lambda_2 = \lambda_3 = \mathbf{u} \cdot \boldsymbol{\omega}$ , the system is not strictly hyperbolic and we must extend slightly the results of Sects. 1.3, 1.4. If  $\mathbf{r}$  belongs to the eigenspace spanned by  $\mathbf{r}_2$  and  $\mathbf{r}_3$ , for a curve

$$\frac{d\mathbf{U}}{d\varphi} = \mathbf{r}(\mathbf{U})$$

we have the relations

$$\frac{dp}{d\varphi} = p_\rho \frac{d\rho}{d\varphi} + p_s \frac{ds}{d\varphi} = 0$$

and

$$\omega_1 \frac{du}{d\varphi} + \omega_2 \frac{dv}{d\varphi} = 0$$

(these relations hold obviously for  $\mathbf{r}_2$  and  $\mathbf{r}_3$ ). Hence, the set of states  $\mathbf{U}$  that can be connected to a given state  $\mathbf{U}_L$  by a plane contact discontinuity corresponding to the eigenvalue  $\mathbf{u} \cdot \boldsymbol{\omega}$  is a two-dimensional manifold in the  $(\rho, u, v, s)$ -space

$$\begin{cases} p = p_L, \\ \omega_1(u - u_L) + \omega_2(v - v_L) = 0, \end{cases} \quad (2.34)$$

which results from the fact that  $p$  and  $\mathbf{u} \cdot \boldsymbol{\omega}$  are Riemann invariants. When  $\boldsymbol{\omega}$  varies, these manifolds span the set  $p = p_L$ . In particular, (2.34) says that normal velocity components are continuous and that there is a discontinuity of the tangential velocity component only, which is characteristic of a shear wave; for a pure “contact” or “entropy” wave

$$\frac{d\mathbf{U}}{d\varphi} = \mathbf{r}_2(\mathbf{U}, \boldsymbol{\omega}) = (p_s, 0, 0, -p_\rho)^T,$$

it is obvious that  $\frac{d\mathbf{u}}{d\varphi} = 0$ , and hence there is no discontinuity in the velocity. The plane of discontinuity in the  $(x, y, t)$ -space is:

$$\mathbf{x} \cdot \boldsymbol{\omega} = (\mathbf{u} \cdot \boldsymbol{\omega})t = (\mathbf{u}_L \cdot \boldsymbol{\omega})t \quad (\text{i.e., } \sigma = \lambda_2 = \mathbf{u} \cdot \boldsymbol{\omega} = \mathbf{u}_L \cdot \boldsymbol{\omega}).$$

### 2.5.2 Plane Shock Waves

The computations for a 1-wave or 4-wave-plane shock follow exactly those for the one-dimensional case. One writes the Rankine–Hugoniot conditions across the plane  $\mathbf{x} \cdot \boldsymbol{\omega} = \sigma t$  as follows:

$$\begin{cases} \sigma[\rho] = \boldsymbol{\omega} \cdot [\rho \mathbf{u}], \\ \sigma[\rho u] = \omega_1[\rho u^2 + p] + \omega_2[\rho uv], \\ \sigma[\rho v] = \omega_1[\rho uv] + \omega_2[\rho v^2 + p], \\ \sigma[\rho e] = \boldsymbol{\omega} \cdot [(\rho e + p)\mathbf{u}] = \omega_1[(\rho e + p)u] + \omega_2[(\rho e + p)v]. \end{cases} \quad (2.35)$$

The analog of the velocity relative to the discontinuity (see Chap. III, formula (2.6)) is now

$$v = \mathbf{u} \cdot \boldsymbol{\omega} - \sigma,$$

so that the first equation (2.35) is equivalent to

$$\rho v = \rho_L v_L,$$

and we shall set

$$\mathcal{M} = \rho v = \rho_L v_L.$$

Note that, as in the one-dimensional case,  $\mathcal{M} = 0$  corresponds to a contact discontinuity, which we have just studied (see (2.34)). In the same way, multiplying the second (resp. the third) equation by  $\omega_1$  (resp.  $\omega_2$ ) and adding, which corresponds to taking the scalar product by  $\boldsymbol{\omega}$  of the system of two equations

$$\sigma[\rho \mathbf{u}] = (\boldsymbol{\omega} \cdot [\rho u \mathbf{u} + p \mathbf{e}_1], \boldsymbol{\omega} \cdot [\rho v \mathbf{u} + p \mathbf{e}_2])^T,$$

yield, together with the first equation,

$$\rho v^2 + p = \rho_L v_L^2 + p_L.$$

And if we take the scalar product by  $\boldsymbol{\omega}^\perp = (-\omega_2, \omega_1)^\perp$ , we get

$$\rho v \mathbf{u} \cdot \boldsymbol{\omega}^\perp = \rho_L v_L \mathbf{u}_L \cdot \boldsymbol{\omega}^\perp.$$

Thus, if  $\mathcal{M} \neq 0$ ,

$$\mathbf{u} \cdot \boldsymbol{\omega}^\perp = \mathbf{u}_L \cdot \boldsymbol{\omega}^\perp,$$

i.e., there is no change in the tangential velocity component (normal to  $\boldsymbol{\omega}$ ), which implies that

$$\mathbf{u} - \mathbf{u}_L = ((\mathbf{u} - \mathbf{u}_L) \cdot \boldsymbol{\omega}) \boldsymbol{\omega} \quad (2.36)$$

is collinear to  $\boldsymbol{\omega}$ . Also, we check that the last equation gives

$$\left\{ \rho \left( \varepsilon + \frac{v^2}{2} \right) + p \right\} v = \left\{ \rho_L \left( \varepsilon_L + \frac{v_L^2}{2} \right) + p_L \right\} v_L,$$

and we have the exact analog of (2.7) in Chap. III.

Then, we have

$$\mathcal{M} = \frac{(\mathbf{u} - \mathbf{u}_L) \cdot \boldsymbol{\omega}}{(\tau - \tau_L)} \quad (2.37)$$

and

$$\mathcal{M}v + p = \rho v^2 + p = \rho_L v_L^2 + p_L = \mathcal{M}v_L + p_L,$$

which gives

$$\mathcal{M} = -\frac{(p - p_L)}{(v - v_L)} = -\frac{(p - p_L)}{(\mathbf{u} - \mathbf{u}_L) \cdot \boldsymbol{\omega}}$$

and

$$\mathcal{M}^2 = -\frac{(p - p_L)}{(\tau - \tau_L)}. \quad (2.38)$$

Finally, we also obtain the equation of the Hugoniot curve (see Chap. III, (2.18)),

$$\varepsilon - \varepsilon_L + \frac{1}{2}(p + p_L)(\tau - \tau_L) = 0. \quad (2.39)$$

Assuming that the Hugoniot curve can be parametrized by  $p$ , i.e., can be represented by an equation of the form

$$\tau = h_L(p),$$

and using (2.36)–(2.38), we proceed as in the one-dimensional case to obtain the shock curves in the  $(\mathbf{u}, p)$ -space

$$\begin{aligned} \mathbf{u} - \mathbf{u}_L &= ((\mathbf{u} - \mathbf{u}_L) \cdot \boldsymbol{\omega})\boldsymbol{\omega} = \pm(-(p - p_L)(\tau - \tau_L))^{1/2}\boldsymbol{\omega} \\ &= \pm\left(\frac{(p - p_L)(\rho - \rho_L)}{\rho\rho_L}\right)^{1/2}\boldsymbol{\omega}. \end{aligned} \quad (2.40)$$

The speed  $\sigma$  is given by (2.35),

$$\begin{aligned} \sigma &= \boldsymbol{\omega} \cdot \frac{(\rho\mathbf{u} - \rho_L\mathbf{u}_L)}{(\rho - \rho_L)} = \boldsymbol{\omega} \cdot \mathbf{u}_L + \rho\boldsymbol{\omega} \cdot \frac{(\mathbf{u} - \mathbf{u}_L)}{(\rho - \rho_L)}, \\ \sigma &= \boldsymbol{\omega} \cdot \mathbf{u}_L + \tau_L\boldsymbol{\omega} \cdot \frac{(\mathbf{u} - \mathbf{u}_L)}{(\tau_L - \tau)}, \end{aligned}$$

and by (2.37), (2.38) we have

$$\begin{aligned} \sigma &= \boldsymbol{\omega} \cdot \mathbf{u}_L \pm \tau_L \left(-\frac{(p - p_L)}{(\tau - \tau_L)}\right)^{1/2} \\ &= \boldsymbol{\omega} \cdot \mathbf{u}_L \pm \left(\frac{\rho(p - p_L)}{\rho_L(\rho - \rho_L)}\right)^{1/2}. \end{aligned}$$

The sign “–” (resp. “+”) corresponds to an admissible shock for the first field  $\lambda_1 = \mathbf{u} \cdot \boldsymbol{\omega} - c = \mathbf{u} \cdot \boldsymbol{\omega} - (\frac{\partial p}{\partial \rho}(\rho, s))^{1/2}$  (resp. the fourth field  $\lambda_4 = \mathbf{u} \cdot \boldsymbol{\omega} + c = \mathbf{u} \cdot \boldsymbol{\omega} + (\frac{\partial p}{\partial \rho}(\rho, s))^{1/2}$ ), and the “Lax entropy conditions”

$$\lambda_1(\mathbf{u}_R, \boldsymbol{\omega}) < \sigma < \lambda_1(\mathbf{u}_L, \boldsymbol{\omega}), \quad \sigma < \lambda_2(\mathbf{u}_R, \boldsymbol{\omega})$$

(resp.

$$\lambda_4(\mathbf{u}_R, \boldsymbol{\omega}) < \sigma < \lambda_4(\mathbf{u}_L, \boldsymbol{\omega}), \quad \sigma > \lambda_3(\mathbf{u}_L, \boldsymbol{\omega})$$

hold.

When  $\boldsymbol{\omega}$  varies, we see from (2.40) that the set spanned by the shock curves is

$$|\mathbf{u} - \mathbf{u}_L|^2 = (u - u_L)^2 + (v - v_L)^2 = -(p - p_L)(\tau - \tau_L) = \frac{(p - p_L)(\rho - \rho_L)}{\rho \rho_L}.$$

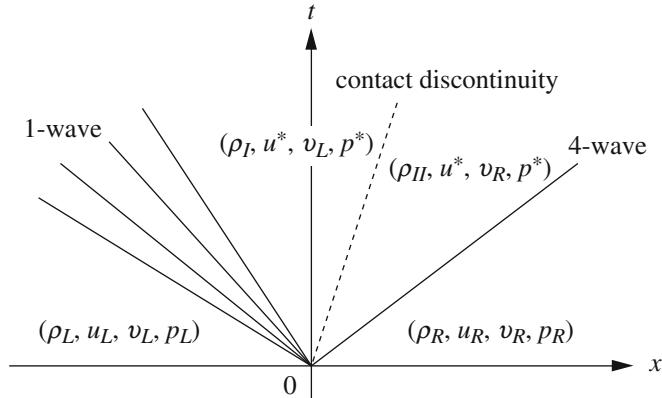
### 2.5.3 Plane Riemann Problem

In the case of the gas dynamics equations that are invariant under rotation, it is important to consider solutions  $\mathbf{U}(\mathbf{x}, t) = \mathbf{V}(\mathbf{x} \cdot \boldsymbol{\omega}, t)$  in the direction of, say,  $\boldsymbol{\omega} = (1, 0)^T$ . Since then,  $\mathbf{U} = \mathbf{U}(x, t)$  and  $\mathbf{A}(\mathbf{U}, \boldsymbol{\omega}) = \mathbf{f}'(\mathbf{U})$ , we are led to the system

$$\begin{cases} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \\ \frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) = 0, \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = 0. \end{cases} \quad (2.41)$$

System (2.41) is a one-dimensional system that is hyperbolic but not strictly hyperbolic. However, it is easy to prove that the result of Chap. II, Theorem 6.1, concerning the solution of the Riemann problem can be extended in this particular case (see Remark 6.1, Chap. II). The solution has the same structure as that of the system which is described in Chap. III, Sect. 3, Remark 3.5. It consists of at most four constant states separated by a 1-wave (shock or rarefaction), a 2-wave (or 3-wave) contact discontinuity, and 4-wave (shock or rarefaction), as illustrated in Fig. 2.1. Across 1- and 4-wave,  $v$  is continuous as results from (2.33) and (2.36) (with  $\boldsymbol{\omega} = (1, 0)^T$ ), while across the 2-wave  $p$  and  $u$  are continuous, as results from (2.34); in primitive variables,  $\mathbf{U}_{L,R} = (\rho, u, v, p)_{L,R}^T$ ,  $\mathbf{U}_I = (\rho_I, u^*, v_L, p^*)^T$ ,  $\mathbf{U}_{II} = (\rho_{II}, u^*, v_R, p^*)^T$ .

*Remark 2.8.* The Riemann problem for (1.1) is the initial value problem with piecewise constant initial data in each of the four quadrants or more generally in sectors meeting at the origin. The problem is invariant under the transformation  $(x, y, t) \mapsto (cx, cy, ct)$ , and the solution is self-similar as in (2.42),  $\mathbf{u}(x, y, t) = \mathbf{v}\left(\frac{x}{t}, \frac{y}{t}\right)$ . The construction of explicit solutions has been investigated in the scalar case by Wagner [1174], Lindquist [806, 807], Klingenberg and Osher [703] essentially when  $g = f$ , and by Chang and Hsiao [278], and



**Fig. 2.1** Solution of the Riemann problem

[1098], Zhang and Zheng [1220, 1222] in general, [4, 5], for its approximation. The case of linear hyperbolic systems of two equations has been investigated by Gilquin et al. [513, 514] and Abgrall [3]. In the case of isentropic or polytropic ideal gas dynamics, the Riemann problem is studied in Zhang and Zheng [1221] for gas dynamics, also Schulz and Rinne [1026, 1027] see also Glimm et al. [533], and the more recent [791].  $\square$

### 2.5.4 Characteristic Equations

Introducing the differentiation along the particle paths,

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x},$$

system (2.41) written in the variables  $\rho, u, s$  reads

$$\begin{aligned}\frac{D\rho}{Dt} + \rho \frac{\partial u}{\partial x} &= 0, \\ \rho \frac{Du}{Dt} + \frac{\partial p}{\partial x} &= 0, \\ \frac{Dv}{Dt} &= 0, \\ \frac{Ds}{Dt} &= 0.\end{aligned}$$

The last two equations  $\frac{Ds}{Dt} = 0, \frac{Dv}{Dt} = 0$ , are characteristic, corresponding to the double eigenvalue  $\mathbf{u} \cdot \boldsymbol{\omega} = u$  (since  $\boldsymbol{\omega} = (1, 0)^T$ ); the characteristic equations that correspond to the simple eigenvalues  $\mathbf{u} \cdot \boldsymbol{\omega} \pm c = u \pm c$  are

$$\frac{\partial p}{\partial t} + (u \pm c) \frac{\partial p}{\partial x} \pm \rho c \left\{ \frac{\partial u}{\partial t} + (u \pm c) \frac{\partial u}{\partial x} \right\} = 0;$$

the differential operator  $\frac{\partial}{\partial t} + (u \pm c) \frac{\partial}{\partial x}$  acts along the characteristic  $\frac{dx}{dt} = u \pm c$ .

### 2.5.5 Self-Similar Solutions

If we are interested in a self-similar (or pseudostationary) solution

$$\mathbf{U}(x, y, t) = \mathbf{V} \left( \frac{x}{t}, \frac{y}{t} \right) \quad (2.42)$$

of system (2.1) (see Remark 2.8 above), we set  $\Xi = (\xi, \zeta)$ , where  $\xi = \frac{x}{t}$ ,  $\zeta = \frac{y}{t}$ , and

$$\begin{aligned} \frac{\partial}{\partial t} &= -\frac{x}{t^2} \frac{\partial}{\partial \xi} - \frac{y}{t^2} \frac{\partial}{\partial \zeta} = -\frac{1}{t} \left( \xi \frac{\partial}{\partial \xi} + \zeta \frac{\partial}{\partial \zeta} \right), \\ \frac{\partial}{\partial x} &= \frac{1}{t} \frac{\partial}{\partial \xi}, \quad \frac{\partial}{\partial y} = \frac{1}{t} \frac{\partial}{\partial \zeta}. \end{aligned}$$

The equation of conservation of mass in (2.1) becomes

$$-\xi \frac{\partial \rho}{\partial \xi} - \zeta \frac{\partial \rho}{\partial \zeta} + \frac{\partial}{\partial \xi}(\rho u) + \frac{\partial}{\partial \zeta}(\rho v) = 0,$$

keeping the same notations for the function  $\mathbf{U}$  and  $\mathbf{V}$ , or

$$\frac{\partial(\rho(u - \xi))}{\partial \xi} + \frac{\partial(\rho(v - \zeta))}{\partial \zeta} = -2\rho.$$

Hence, we are led to set

$$\tilde{\mathbf{u}} = \mathbf{u} - \Xi = (u - \xi, v - \zeta).$$

The other equations are computed in the same way, and system (2.1) becomes

$$\begin{aligned} \frac{\partial(\rho \tilde{u})}{\partial \xi} + \frac{\partial(\rho \tilde{v})}{\partial \zeta} &= -2\rho, \\ \frac{\partial}{\partial \xi}(\rho \tilde{u}^2 + p) + \frac{\partial}{\partial \zeta}(\rho \tilde{u} \tilde{v}) &= -3\rho \tilde{u}, \\ \frac{\partial}{\partial \xi}(\rho \tilde{u} \tilde{v}) + \frac{\partial}{\partial \zeta}(\rho \tilde{v}^2 + p) &= -3\rho \tilde{v}, \\ \frac{\partial}{\partial \xi}((\rho \tilde{e} + p)\tilde{u}) + \frac{\partial}{\partial \zeta}((\rho \tilde{e} + p)\tilde{v}) &= -\rho |\tilde{\mathbf{u}}|^2 - 2\rho \left( \tilde{e} + \frac{p}{\rho} \right), \end{aligned}$$

where  $\bar{e} = e + \frac{(\tilde{u}^2 + \tilde{v}^2)}{2} = e + \frac{|\tilde{\mathbf{u}}|^2}{2}$ . Thus, the system satisfied by a self-similar solution is the steady Euler system in the variable  $\tilde{\mathbf{U}} = (\rho, \rho\tilde{u}, \rho\tilde{v}, \rho\tilde{e})$  with a source term

$$\frac{\partial \mathbf{f}(\tilde{\mathbf{U}})}{\partial \xi} + \frac{\partial \mathbf{g}(\tilde{\mathbf{U}})}{\partial \zeta} = \tilde{\mathbf{S}},$$

where  $\mathbf{f}, \mathbf{g}$  are defined by (2.2a). The overtaking of two shocks in steady flow (above system without source term) is studied in Chang and Hsiao [278]; see also Marshall and Plohr [851] and Glaz [528].

*Remark 2.9.* Let us have a look at the equation of two-dimensional steady flow that we have just encountered, i.e.,

$$\frac{\partial \mathbf{f}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{g}(\mathbf{U})}{\partial y} = \mathbf{0}, \quad (2.43)$$

where  $\mathbf{f}, \mathbf{g}$  are defined by (2.2a). If we seek a smooth, simple wave solution  $\mathbf{U}(\mathbf{x}) = \mathbf{V}(\mathbf{x} \cdot \boldsymbol{\omega})$  constant along the characteristic lines  $\mathbf{x} \cdot \boldsymbol{\omega} = \text{const.}$ , we find that

$$(\omega_1 \mathbf{f}'(\mathbf{V}) + \omega_2 \mathbf{g}'(\mathbf{V})) \mathbf{V}' = \mathbf{A}(\mathbf{V}, \boldsymbol{\omega}) \mathbf{V}' = \mathbf{0},$$

and  $\mathbf{A}(\mathbf{V}, \boldsymbol{\omega})$  is not invertible. We have already computed

$$\det \mathbf{A}(\mathbf{V}, \boldsymbol{\omega}) - \lambda \mathbf{I} = (\lambda - \mathbf{u} \cdot \boldsymbol{\omega})^2 \{(\mathbf{u} \cdot \boldsymbol{\omega} - \lambda)^2 - c^2 |\boldsymbol{\omega}|^2\}.$$

Thus, the roots  $\boldsymbol{\omega}$  of

$$\det \mathbf{A}(\mathbf{V}, \boldsymbol{\omega}) = (\mathbf{u} \cdot \boldsymbol{\omega})^2 (\mathbf{u} \cdot \boldsymbol{\omega}^2 - c^2 |\boldsymbol{\omega}|^2) = 0$$

are firstly those of

$$\mathbf{u} \cdot \boldsymbol{\omega} = 0,$$

which always exist and correspond to the streamlines (since the direction of the characteristics  $\mathbf{x} \cdot \boldsymbol{\omega} = \text{const.}$  is then  $\mathbf{u}$ ), and secondly those of the quadratic polynomial

$$(\mathbf{u} \cdot \boldsymbol{\omega})^2 - c^2 |\boldsymbol{\omega}|^2 = 0.$$

Considered as a polynomial in  $\frac{\omega_1}{\omega_2}$  (slope of the characteristic), it has real roots if

$$|\mathbf{u}|^2 \geq c^2,$$

i.e., in supersonic regions where the system (2.41) is thus hyperbolic. There, we get

$$\frac{\omega_1}{\omega_2} = \frac{(uv \pm c(u^2 + v^2 - c^2)^{1/2})}{(u^2 - c^2)}.$$

Since

$$\mathbf{u} \cdot \frac{\boldsymbol{\omega}}{|\boldsymbol{\omega}|} = |\mathbf{u}| \cos(\mathbf{u}, \boldsymbol{\omega}) = \pm c,$$

the characteristics  $\mathbf{x} \cdot \boldsymbol{\omega} = c$  (which are orthogonal to  $\boldsymbol{\omega}$ ) make with the streamlines an angle  $\pm\mu$  such that  $\sin\mu = \frac{c}{|\mathbf{u}|} = \frac{1}{M}$ , where  $M$  is the Mach number, and are often called Mach lines.

Using the characteristic equations, we can transform the system into characteristic form (following the arguments of Sect. 1.2, which can easily be extended). For details concerning this “hodograph” transformation, and more generally steady supersonic plane flows, we refer to Courant and Friedrichs [371], J.D. Anderson [40], Whitham [1188], and Menikoff [861]. A physical example of 2d stationary simple waves is given by the flow around a bend (Prandtl–Meyer expansion waves), which is studied in the above references (respectively in Section 111, Section 4.13, and Section 6.17).

The flow may also be subsonic ( $M = \frac{|\mathbf{u}|}{c} < 1$ ) so that the sound wave characteristics become complex and the system is elliptic. The flow is called transonic if it involves mixed subsonic–supersonic regions. The steady equations (2.43) are of mixed type, while the time-dependent problem is always hyperbolic.  $\square$

### 3 Multidimensional Finite Difference Schemes

#### 3.1 Direct Approach

##### 3.1.1 Difference Schemes in Conservation Form

Many schemes can be extended to two dimensions in the finite difference setting. We consider a two-dimensional uniform Cartesian spatial grid  $\Delta = \Delta x \times \Delta y$  with space increments  $\Delta x$  and  $\Delta y$ , and we set

$$\lambda_x = \frac{\Delta t}{\Delta x}, \quad \lambda_y = \frac{\Delta t}{\Delta y}. \quad (3.1)$$

Let  $\mathbf{v}_{j,k}^n$  denote an approximation of the solution at the grid point ( $x_j = j\Delta x$ ,  $y_k = k\Delta y$ ,  $t_n = n\Delta t$ ) (or rather an approximation of its average value on the rectangular cell with center  $(x_j, y_k)$  at time  $t_n$ ) and define the sequences  $\mathbf{v}^n$  by  $\mathbf{v}^n = (\mathbf{v}_{j,k}^n)$ ,  $j, k \in \mathbb{Z}, n \in \mathbb{N}$ . For approximating the system

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) + \frac{\partial}{\partial y} \mathbf{g}(\mathbf{u}) = 0, \quad (3.2)$$

we can use a finite difference scheme of the form

$$\mathbf{v}^{n+1} = \mathbf{H}_\Delta(\mathbf{v}^n), \quad (3.3a)$$

i.e.,

$$\mathbf{v}_{j,k}^{n+1} = \mathbf{H}(\mathbf{v}_{j-J,k-K}^n, \dots, \mathbf{v}_{j+J,k+K}^n), \quad (3.3b)$$

where the discrete solution operator  $\mathbf{H} : \mathbb{R}^{(2J+1) \times (2K+1) \times p} \rightarrow \mathbb{R}^p$  and  $\mathbf{v}^0 = (\mathbf{v}_{j,k}^0)$  are given, for instance,

$$\begin{aligned} \mathbf{v}_{j,k}^0 &= \frac{1}{\Delta x \Delta y} \int_{\Omega_{j,k}} \mathbf{u}_0(\boldsymbol{\xi}) d\boldsymbol{\xi}, \\ \Omega_{j,k} &= (x_{j-1/2}, x_{j+1/2}) \times (y_{k-1/2}, y_{k+1/2}). \end{aligned}$$

The scheme is in *conservation form* if there exist continuous functions  $\mathbf{F} : \mathbb{R}^{2J \times (2K+1) \times p} \rightarrow \mathbb{R}^p$  and  $\mathbf{G} : \mathbb{R}^{(2J+1) \times 2K \times p} \rightarrow \mathbb{R}^p$  such that

$$\mathbf{u}_{j,k}^{n+1} = \mathbf{u}_{j,k}^n - \lambda_x(\mathbf{F}_{j+1/2,k}^n - \mathbf{F}_{j-1/2,k}^n) - \lambda_y(\mathbf{G}_{j,k+1/2}^n - \mathbf{G}_{j,k-1/2}^n), \quad (3.4)$$

where

$$\begin{aligned} \mathbf{F}_{j+1/2,k}^n &= \mathbf{F}(\mathbf{v}_{j-J+1,k-K}^n, \dots, \mathbf{v}_{j+J,k+K}^n), \\ \mathbf{G}_{j,k+1/2}^n &= \mathbf{G}(\mathbf{v}_{j-J,k-K+1}^n, \dots, \mathbf{v}_{j+J,k+K}^n). \end{aligned}$$

For instance, if  $J = K = 1$  (nine-point schemes), each component of  $\mathbf{F}$  and  $\mathbf{G}$  is a function of six variables,

$$\begin{aligned} \mathbf{F}_{j+1/2,k} &= \mathbf{F}(\mathbf{v}_{j,k-1}, \mathbf{v}_{j,k}, \mathbf{v}_{j,k+1}, \mathbf{v}_{j+1,k-1}, \mathbf{v}_{j+1,k}, \mathbf{v}_{j+1,k+1}), \\ \mathbf{G}_{j,k+1/2} &= \mathbf{G}(\mathbf{v}_{j-1,k}, \mathbf{v}_{j,k}, \mathbf{v}_{j+1,k}, \mathbf{v}_{j-1,k+1}, \mathbf{v}_{j,k+1}, \mathbf{v}_{j+1,k+1}). \end{aligned}$$

Let us observe that the conservation form implies the following: if  $\mathbf{v}_{j-J,k} = \dots = \mathbf{v}_{j+J,k} = \mathbf{v}_k, \forall k, -K \leq k \leq K$ , then  $\mathbf{F}_{j-1/2,k} = \mathbf{F}_{j+1/2,k}$ ; similarly, if  $\mathbf{v}_{j,k-K} = \dots = \mathbf{v}_{j,k+K} = \mathbf{v}_j, \forall j, -J \leq j \leq J$ , then  $\mathbf{G}_{j,k+1/2} = \mathbf{G}_{j,k-1/2}$ .

We shall assume, moreover, that the numerical fluxes  $\mathbf{F}$  and  $\mathbf{G}$  are consistent with  $\mathbf{f}$  and  $\mathbf{g}$ , respectively, i.e.,

$$\mathbf{F}(\mathbf{u}, \mathbf{u}, \dots, \mathbf{u}) = \mathbf{f}(\mathbf{u}), \quad \mathbf{G}(\mathbf{u}, \mathbf{u}, \dots, \mathbf{u}) = \mathbf{g}(\mathbf{u}).$$

One can study the order of accuracy of a difference scheme exactly as in the one-dimensional case (G.R., Chapter 3, Section 1.2) by expanding a smooth solution in Taylor series and deriving the equivalent system (see Lerat [761], Jeng and Chen [652], Billet [149]). For instance, a nine-point conservative scheme is second-order accurate if the fluxes  $\mathbf{F}(\mathbf{u}_{-1}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{v}_{-1}, \mathbf{v}_0, \mathbf{v}_1)$  and  $\mathbf{G}(\mathbf{u}_{-1}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{v}_{-1}, \mathbf{v}_0, \mathbf{v}_1)$  satisfy

$$\begin{cases} \sum_j \left( \frac{\partial \mathbf{F}}{\partial u_j} + \frac{\partial \mathbf{F}}{\partial v_j} \right) (\mathbf{u}, \mathbf{u}, \dots, \mathbf{u}) = \mathbf{A}(\mathbf{u}), \\ \sum_j \left( \frac{\partial \mathbf{G}}{\partial u_j} + \frac{\partial \mathbf{G}}{\partial v_j} \right) (\mathbf{u}, \mathbf{u}, \dots, \mathbf{u}) = \mathbf{B}(\mathbf{u}), \end{cases} \quad (3.5a)$$

where  $\mathbf{A}(\mathbf{u})$  (resp.  $\mathbf{B}(\mathbf{u})$ ) is the Jacobian matrix of  $\mathbf{f}$  (resp.  $\mathbf{g}$ ); this relation comes from the consistency (order one), and

$$\begin{cases} \sum_j \left( \frac{\partial \mathbf{F}}{\partial u_j} - \frac{\partial \mathbf{F}}{\partial v_j} \right) (\mathbf{u}, \dots, \mathbf{u}) = \lambda_x \mathbf{A}^2(\mathbf{u}), \\ \sum_j \left( \frac{\partial \mathbf{G}}{\partial u_j} - \frac{\partial \mathbf{G}}{\partial v_j} \right) (\mathbf{u}, \dots, \mathbf{u}) = \lambda_y \mathbf{B}^2(\mathbf{u}), \end{cases} \quad (3.5b)$$

$$\begin{cases} \sum_j j \left( \frac{\partial \mathbf{F}}{\partial u_j} + \frac{\partial \mathbf{F}}{\partial v_j} \right) (\mathbf{u}, \dots, \mathbf{u}) = -\frac{\lambda_y}{2} \mathbf{A}(\mathbf{u}) \mathbf{B}(\mathbf{u}), \\ \sum_j j \left( \frac{\partial \mathbf{G}}{\partial u_j} + \frac{\partial \mathbf{G}}{\partial v_j} \right) (\mathbf{u}, \dots, \mathbf{u}) = -\frac{\lambda_x}{2} \mathbf{B}(\mathbf{u}) \mathbf{A}(\mathbf{u}). \end{cases} \quad (3.5c)$$

For details concerning the study of nine-point linear schemes ( $J = K = 3$ ), we refer to Lerat [761].

*Example 3.1. The Lax–Wendroff scheme.* It is derived, as in the one-dimensional case, from a Taylor expansion of a (smooth) solution and corresponds to the numerical fluxes

$$\begin{aligned} \mathbf{F}(\mathbf{v}_{j,k-1}, \dots, \mathbf{v}_{j+1,k+1}) &= \frac{1}{2}(\mathbf{f}(\mathbf{v}_{j,k}) + \mathbf{f}(\mathbf{v}_{j,k})) \\ &\quad - \frac{\lambda_x}{2} \mathbf{A}_{j+1/2,k}(\mathbf{f}(\mathbf{v}_{j+1,k}) - \mathbf{f}(\mathbf{v}_{j,k})) \\ &\quad - \frac{\lambda_y}{8} \left\{ \mathbf{A}(\mathbf{v}_{j,k})(\mathbf{g}(\mathbf{v}_{j,k+1}) - \mathbf{g}(\mathbf{v}_{j,k-1})) \right. \\ &\quad \left. + \mathbf{A}(\mathbf{v}_{j+1,k})(\mathbf{g}(\mathbf{v}_{j+1,k+1}) - \mathbf{g}(\mathbf{v}_{j+1,k-1})) \right\}, \end{aligned}$$

Where, for instance,  $\mathbf{A}_{j+1/2,k} = \mathbf{A}\left(\frac{\mathbf{v}_{j,k} + \mathbf{v}_{j+1,k}}{2}\right)$ , with an analogous formula for  $\mathbf{G}$  obtained by exchanging  $\lambda_x$  and  $\lambda_y$ ,  $\mathbf{A}$  and  $\mathbf{B}$ ,  $\mathbf{f}$  and  $\mathbf{g}$ ,  $j$  and  $k$ . Thus, one notices the presence of “crossed derivative” terms. The scheme is second-order accurate, but in the linear case it is not the only one since the six linear relations (3.5), which are required for second-order accuracy, do not determine the nine coefficients uniquely as was the case in dimension one G.R., Chapter 3, Section 1.3 [539]; see Lax and Wendroff [747].  $\square$

If we restrict ourselves to the *scalar case* ( $p = 1$ ),

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) + \frac{\partial}{\partial y} g(u) = 0, \quad (3.6a)$$

then recall that Theorem 5.4 in the Chap. I gives for  $u_0 \in \mathbf{L}^\infty(\mathbb{R}^2)$  the existence and uniqueness of the entropy solution  $u$  of the scalar conservation law (3.6a) satisfying

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}). \quad (3.6b)$$

The consistency of the numerical fluxes ensures that, in the scalar case, when the sequence of approximate solutions converges in some sensible way, the limit is indeed a weak solution, i.e., satisfies the Rankine–Hugoniot jump condition (Lax–Wendroff Theorem, see Chap. IV, Theorem 1.1. and for details G.R., Chapter 3, Theorem 1.1).

Still, in the scalar case, the scheme (3.3) is *monotone* if  $H$  is a nondecreasing function of each of its arguments.

In many cases, we use much simpler formulas where  $\mathbf{F}_{j+1/2,k} = \mathbf{F}(\mathbf{v}_{j-J+1,k}, \dots, \mathbf{v}_{j+J,k})$  (resp.  $\mathbf{G}_{j,k+1/2} = \mathbf{G}(\mathbf{v}_{j,k-K+1}, \dots, \mathbf{v}_{j,k+K})$ ) depends only on the  $2J$  values  $\mathbf{v}_{p,k}$ ,  $j - J + 1 \leq p \leq j + J$ , (resp. on the  $2K$  values  $\mathbf{v}_{j,q}$ ,  $k - K + 1 \leq q \leq k + K$ ). This occurs naturally if we start from one-dimensional numerical fluxes  $\mathbf{F} : \mathbb{R}^{2J \times p} \rightarrow \mathbb{R}^p$  and  $\mathbf{G} : \mathbb{R}^{2K \times p} \rightarrow \mathbb{R}^p$  consistent, respectively, with  $\mathbf{f}$  and  $\mathbf{g}$ . We can also take a combination of one-dimensional numerical fluxes consistent, respectively, with  $2\mathbf{f}$  and  $2\mathbf{g}$ , i.e., construct

$$\mathbf{v}_{j,k}^{n+1} = \mathbf{v}_{j,k}^n + \frac{\lambda_x}{2} (\mathbf{F}_{j+1/2,k}^n - \mathbf{F}_{j-1/2,k}^n) + \frac{\lambda_y}{2} (\mathbf{G}_{j,k+1/2}^n - \mathbf{G}_{j,k-1/2}^n).$$

This scheme corresponds to the discretization of the system (3.2) written in the form

$$\frac{1}{2} \left\{ \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} 2\mathbf{f}(\mathbf{u}) \right\} + \frac{1}{2} \left\{ \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial y} 2\mathbf{g}(\mathbf{u}) \right\} = \mathbf{0}.$$

If  $F$  and  $G$  are the numerical fluxes of monotone schemes, the resulting scheme is monotone in the scalar case.

*Example 3.2. The Lax–Friedrichs Scheme.* We take for  $F$  and  $G$  the one-dimensional Lax–Friedrichs numerical flux (see Chap. IV, Example 1.1, and G.R., Chapter 3, Example 2.1, [539]), consistent with  $2f$  (i.e., we replace  $f$  in the formula by  $2f$ ). The above construction gives

$$\begin{aligned} v_{j,k}^{n+1} &= v_{j,k}^n + \frac{v_{j+1,k}^n - 2v_{j,k}^n + v_{j-1,k}^n}{4} - \frac{\lambda_x}{2} \{ f(v_{j+1,k}^n) - f(v_{j-1,k}^n) \} \\ &\quad + \frac{v_{j,k+1}^n - 2v_{j,k}^n + v_{j,k-1}^n}{4} - \frac{\lambda_y}{2} \{ g(v_{j,k+1}^n) - g(v_{j,k-1}^n) \} \\ &= \frac{v_{j+1,k}^n + v_{j-1,k}^n + v_{j,k+1}^n + v_{j,k-1}^n}{4} - \frac{\lambda_x}{2} \{ f(v_{j+1,k}^n) - f(v_{j-1,k}^n) \} \\ &\quad - \frac{\lambda_y}{2} \{ g(v_{j,k+1}^n) - g(v_{j,k-1}^n) \}. \end{aligned}$$

This scheme is monotone if  $\lambda_x \max |f'| \leq \frac{1}{2}$  and  $\lambda_y \max |g'| \leq \frac{1}{2}$ . This has to be compared with the one-dimensional CFL condition  $\lambda \max |f'| \leq 1$ .

If we had just added the usual Lax–Friedrichs fluxes in each direction, we would have obtained the following formula:

$$\begin{aligned}
v_{j,k}^{n+1} &= v_{j,k}^n + \frac{v_{j+1,k}^n - 2v_{j,k}^n + v_{j-1,k}^n}{2} - \frac{\lambda_x}{2} \{f(v_{j+1,k}^n) - f(v_{j-1,k}^n)\} \\
&\quad + \frac{v_{j,k+1}^n - 2v_{j,k}^n + v_{j,k-1}^n}{2} - \frac{\lambda_y}{2} \{g(v_{j,k+1}^n) - g(v_{j,k-1}^n)\} \\
&= \frac{1}{2}(v_{j+1,k}^n + v_{j-1,k}^n + v_{j,k+1}^n + v_{j,k-1}^n - 2v_{j,k}^n) \\
&\quad - \frac{\lambda_x}{2} \{f(v_{j,k+1}^n) - f(v_{j,k-1}^n)\} - \frac{\lambda_y}{2} \{g(v_{j,k+1}^n) - g(v_{j,k-1}^n)\},
\end{aligned}$$

in which the dependence on  $v_{j,k}^n$  is effective.  $\square$

*Example 3.3. “Five-Point” Schemes.* For  $J = K = 1$  (a nine-point scheme), the simpler formulas  $F_{j+1/2,k}^n = F(v_{j-J+1,k}^n, \dots, v_{j+J,k}^n)$  (resp.  $G_{j,k+1/2}^n = G(v_{j,k-K+1}^n, \dots, v_{j,k+K}^n)$ ) give that  $F$  and  $G$  are functions of only two variables,

$$F_{j+1/2,k} = F(v_{j,k}, v_{j+1,k}), \quad G_{j,k+1/2} = G(v_{j,k}, v_{j,k+1}),$$

and  $u_{j,k}^{n+1}$  depends on only five values. However, we cannot get second-order accuracy for which at least six points are necessary. Indeed, relations (3.5a), (3.5b) yield, as expected,

$$\begin{aligned}
\frac{\partial F}{\partial u_0} &= \frac{a}{2}(1 + \lambda_x a), & \frac{\partial F}{\partial v_0} &= \frac{a}{2}(1 - \lambda_x a), \\
\frac{\partial G}{\partial u_0} &= \frac{b}{2}(1 + \lambda_y b), & \frac{\partial G}{\partial v_0} &= \frac{b}{2}(1 - \lambda_y b),
\end{aligned}$$

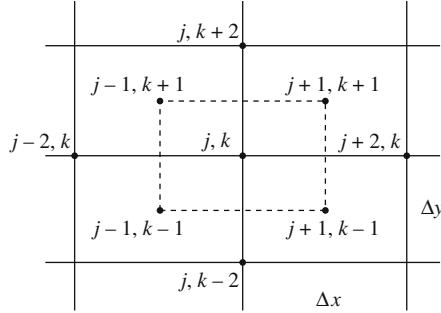
but  $\frac{\partial F}{\partial u_{\pm 1}} = \frac{\partial F}{\partial v_{\pm 1}} = \frac{\partial G}{\partial u_{\pm 1}} = \frac{\partial G}{\partial v_{\pm 1}} = 0$  is clearly incompatible with (3.5c), which means that “crossed” terms must be involved.  $\square$

*Example 3.4. The Two-Step Lax–Wendroff Scheme.* This last drawback can be avoided by using the two-step version of the Lax–Wendroff scheme proposed by Richtmyer (Chap. IV, Example 1.2, G.R., Chapter 3, (2.19), [539]; see Richtmyer and Morton [974], Section 13.4). This two-step scheme is given by:

$$\begin{aligned}
v_{j,k}^{n+1/2} &= \frac{1}{4}(v_{j+1,k}^n + v_{j-1,k}^n + v_{j,k+1}^n + v_{j,k-1}^n) \\
&\quad - \frac{\lambda_x}{2}(f(v_{j+1,k}^n) - f(v_{j-1,k}^n)) - \frac{\lambda_y}{2}(g(v_{j,k+1}^n) - g(v_{j,k-1}^n)) \\
v_{j,k}^{n+1} &= v_{j,k}^n - \lambda_x(f(v_{j+1,k}^{n+1/2}) - f(v_{j-1,k}^{n+1/2})) - \lambda_y(g(v_{j,k+1}^{n+1/2}) - g(v_{j,k-1}^{n+1/2})).
\end{aligned}$$

The two steps involve staggered meshes, and the resulting scheme is in fact nine-point (see Fig. 3.1).  $\square$

In this frame of finite difference schemes constructed from one-dimensional schemes, the quasi-monotone schemes (Cockburn [315]), schemes with flux



**Fig. 3.1** The two-step Lax–Wendroff scheme

limiter can be extended to the two-dimensional case (Spekreijse [1071], Venkatakrishnan [1167]). Also, a fully multidimensional (one-step) extension of the F.C.T. scheme has been derived (Zalesak [1216]). Similarly, Colella has derived upwind methods (Colella [328], Saltzman [1004], Pember et al. [936]), and ENO schemes can be extended to two- and three-dimensional flows (Harten [592], Shu et al. [1060], Casper and Atkins [239]).

### 3.1.2 $L^2$ Stability

One can first study the  $L^2$  stability in the linear case

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} + \mathbf{B} \frac{\partial \mathbf{u}}{\partial y} = \mathbf{0}, \quad (3.7)$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are constant matrices. Following the ideas of G.R., Chapter 3, Section 1.3, and the present Chap. IV, Sect. 1.2, we consider a linear scheme that can be written

$$\mathbf{v}_{j,k}^{n+1} = \sum_{\substack{\ell=-J, \dots, J \\ m=-K, \dots, K}} \mathbf{C}_{\ell,m} \mathbf{v}_{j+\ell, k+m}^n,$$

where the matrices  $\mathbf{C}$  are polynomials in  $\lambda_x \mathbf{A}$  and  $\lambda_y \mathbf{B}$ . By extending the scheme to the whole space  $\mathbb{R}^2$ ,

$$\mathbf{v}^{n+1}(x, y) = \sum_{\ell,m} \mathbf{C}_{\ell,m} \mathbf{v}^n(x + \ell\Delta x, y + m\Delta y),$$

where  $\mathbf{v}^n$  denotes the piecewise constant function with value  $\mathbf{v}_{j,k}^n$  on  $\Omega_{j,k}$ , and by using the Fourier transform

$$\hat{\varphi}(\xi, \eta) = (2\pi)^{-1} \int_{\mathbb{R}^2} e^{i(x\xi + y\eta)} \varphi(x, y) dx dy, \quad (\xi, \eta) \in \mathbb{R}^2,$$

we get

$$\hat{\mathbf{v}}^{n+1}(\xi, \eta) = \mathbf{G}^a(\xi, \eta)\hat{\mathbf{v}}^n(\xi, \eta),$$

where the amplification matrix is defined by

$$\mathbf{G}^a(\xi, \eta) = \sum_{\ell, m} \mathbf{C}_{\ell, m} e^{i(\ell \xi \Delta x + m \eta \Delta y)}.$$

The  $\ell^2$  norm of the sequence  $(\mathbf{v}_{j,k}^n)_{j,k}$  is equal to the  $\mathbf{L}^2$  norm of the function  $\mathbf{v}^n$  up to the factor  $(\Delta x \Delta y)^{1/2}$ , and the Fourier transform is an isometry of  $\mathbf{L}^2(\mathbb{R}^2)$ .

Then a simple, necessary condition for  $L^2$  stability, known as the von Neumann condition, is that the spectral radius of  $\mathbf{G}^a$  be less than 1. A necessary and sufficient condition is that the powers  $(\mathbf{G}^a)^n(\xi, \eta)$  of the amplification matrix be bounded uniformly in  $(\xi, \eta)$  and  $n$ . This condition was proven by Kreiss to be equivalent to the so-called resolvent condition or to the Hermitean norm condition

$$\sup_n \sup_{\xi, \eta \in \mathbb{R}^2} \|(\mathbf{G}^a)^n(\xi, \eta)\| < \infty$$

where the norm denotes the spectral norm (see [573], Chap. 5).

Note that  $\mathbf{G}^a$  depends on  $(\xi, \eta)$  but also on  $(\lambda_x \mathbf{A}, \lambda_y \mathbf{B})$ , which means that besides the dissipation linked to the modulus of the eigenvalues of  $\mathbf{G}^a$  and the “phase error” induced by the imaginary part, there appears “numerical anisotropy” due to the dependence of  $\mathbf{G}^a$  on the advection direction. A thorough study of  $\mathbf{G}^a$  in the general case is difficult, and one often assumes particular values (such as square mesh) in order to carry out the computations (see Desideri et al. [413] for an example).

Following the same approach as in Chap. IV, Sect. 1.3, we see that by Fourier transform the linear system (3.7) gives

$$\frac{\partial \hat{\mathbf{u}}}{\partial t} + i(\xi \mathbf{A} + \eta \mathbf{B})\hat{\mathbf{u}} = \mathbf{0},$$

and the exact amplification matrix is

$$\mathbf{G}^{ex}(\xi, \eta) = \exp(-i(\xi \mathbf{A} + \eta \mathbf{B})\Delta t).$$

If the system is hyperbolic, the matrix  $\xi \mathbf{A} + \eta \mathbf{B}$  has real eigenvalues and is diagonalizable, and

$$\rho(\mathbf{G}^{ex}(\xi, \eta)) = 1.$$

Equivalently, if we look for an elementary solution (Fourier mode) of the form

$$\mathbf{u}(\mathbf{x}, t) = \hat{\mathbf{u}} e^{i(\mathbf{k} \cdot \mathbf{x} - \omega t)},$$

where  $\mathbf{k} = (\xi, \eta)^T$  is the vector wave number and  $\omega$  the frequency,  $\omega$  must be an eigenvalue of  $\xi\mathbf{A} + \eta\mathbf{B}$  and is therefore real, so that the amplitude remains constant.

For results on  $L^2$  stability of some classical schemes (upwind, Lax-Wendroff) for a symmetric hyperbolic system in 2d, we refer to [368] and the references therein.

In the scalar case

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = 0,$$

substituting an elementary wave

$$u(\mathbf{x}, t) = \hat{u} e^{i(\mathbf{k} \cdot \mathbf{x} - \omega t)},$$

we get the dispersion relation (see Chap. IV, Sect. 1.3)  $\omega(\mathbf{k}) = \mathbf{c} \cdot \mathbf{k}$ , where  $\mathbf{c} = (a, b)^T$  is the advection vector. The phase surfaces  $\mathbf{k} \cdot \mathbf{x} - \omega t = \text{constant}$  are parallel planes that propagate in the direction  $\mathbf{k}$  with normal velocity  $\frac{\omega}{|\mathbf{k}|}$ , and the phase velocity is  $\frac{\omega \mathbf{k}}{|\mathbf{k}|^2}$ . Note that the group velocity is  $\text{grad } \omega(\mathbf{k}) = \mathbf{c}$ , which is constant and represents the advection direction, whereas the phase velocity may point in any direction (Whitham, Chapter 11 [1188], Higdon [615]).

Looking for discrete Fourier mode solutions of the scalar numerical scheme

$$v_{j,k}^n = \hat{u} e^{i(\xi j \Delta x + \eta k \Delta y - \omega n \Delta t)}$$

leads to the discrete dispersion relation

$$e^{i\omega \Delta t} = g^a(\xi, \eta),$$

where  $g^a$  is the amplification factor. Writing  $\tan(\omega \Delta t) = -\arg(g^a(\xi, \eta))$  yields by implicit derivation the discrete group velocity  $(\frac{\partial \omega}{\partial \xi}, \frac{\partial \omega}{\partial \eta})^T$ . Again, the error in group velocity yields an error not only in the speed (dispersion) but also in the direction (anisotropy) (see Trefethen [1132, 1133]).

*Example 3.5. The upwid scheme.* The natural upwind approximation of the scalar linear equation

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = 0, \quad a > 0, \quad b > 0,$$

whose exact solutions satisfy  $u(x, y, t) = u(x - at, y - bt)$ , would be

$$v_{j,k}^{n+1} = v_{j,k}^n - \lambda_x a(v_{j,k}^n - v_{j-1,k}^n) - \lambda_y b(v_{j,k}^n - v_{j,k-1}^n).$$

It involves only three points:  $j, k; j - 1, k;$  and  $j, k - 1$ . Setting  $\nu_x = \lambda_x a$ ,  $\nu_y = \lambda_y b$ , we write

$$v_{j,k}^{n+1} = v_{j,k}^n - \nu_x (v_{j,k}^n - v_{j-1,k}^n) - \nu_y (v_{j,k}^n - v_{j,k-1}^n);$$

the coefficient of amplification is

$$\begin{aligned} g^a(\xi, \eta) &= 1 - \nu_x(1 - \cos \xi \Delta x) - \nu_y(1 - \cos \eta \Delta y) \\ &\quad - i(\nu_x \sin \xi \Delta x + \nu_y \sin \eta \Delta y). \end{aligned}$$

For  $\xi \Delta x = \eta \Delta y = \pi$ , we find the necessary condition

$$|1 - 2(\nu_x + \nu_y)| \leq 1$$

or, equivalently,

$$0 \leq \nu_x + \nu_y = \Delta t \left( \frac{a}{\Delta x} \right) + \left( \frac{b}{\Delta y} \right) \leq 1.$$

This CFL condition seems too restrictive (for a given mesh, it is most restrictive if the convection  $\mathbf{c} = (a, b)^T$  is parallel to  $(\Delta x, \Delta y)^T$ ).

Instead, consider the split scheme (see Sect. 3.2)

$$\begin{aligned} u_{j,k}^{n,1} &= u_{j,k}^n - \lambda_x a(u_{j,k}^n - u_{j-1,k}^n) = (1 - \nu_x)u_{j,k}^n + \nu_x u_{j-1,k}^n, \\ u_{j,k}^{n+1} &= u_{j,k}^{n,1} - \lambda_y b(u_{j,k}^{n,1} - u_{j,k-1}^{n,1}) = (1 - \nu_y)u_{j,k}^{n,1} + \nu_y u_{j,k-1}^{n,1}, \end{aligned}$$

which differs from the above scheme by a second-order cross term

$$\begin{aligned} u_{j,k}^{n+1} &= u_{j,k}^n - \nu_x(u_{j,k}^n - u_{j-1,k}^n) - \nu_y(u_{j,k}^n - u_{j,k-1}^n) \\ &\quad + \nu_x \nu_y \{(u_{j,k}^n - u_{j-1,k}^n) - (u_{j,k-1}^n - u_{j-1,k-1}^n)\}. \end{aligned}$$

The amplification factor is the product

$$\begin{aligned} g^a(\xi, \eta) &= (1 - \nu_x(1 - \cos \xi \Delta x) - i\nu_x \sin \xi \Delta x) \\ &\quad (1 - \nu_y(1 - \cos \eta \Delta y) - i\nu_y \sin \eta \Delta y), \end{aligned}$$

and the split scheme is stable under the CFL condition

$$0 \leq \nu_x \leq 1, \quad 0 \leq \nu_y \leq 1,$$

which appears to be less restrictive.  $\square$

Together with the preceding examples, this shows that there are several ways of extending a one-dimensional scheme to two dimensions. For a study of optimum linear schemes for advection, see Roe and Sidilkover [991].

### 3.1.3 Total Variation

Let us now define the following “ $\mathbf{L}^1(\Delta)$  norm” of a scalar sequence  $u = (u_{j,k})$ ,

$$\|u\|_{\mathbf{L}^1(\Delta)} = \Delta x \Delta y \sum_{j,k} |u_{j,k}|;$$

it is the  $\ell^1$  norm of the sequence  $(u_{j,k})$  (up to a factor  $\Delta x \Delta y$ ) or the  $\mathbf{L}^1$  norm of the associated piecewise constant  $u_\Delta$  defined in (3.9). Then define the total variation of  $u$  by

$$TV(u) = \sum_{j,k} \{\Delta y |u_{j+1,k} - u_{j,k}| + \Delta x |u_{j,k+1} - u_{j,k}|\}.$$

If we introduce the notation

$$TV_{x,1}(u) = \Delta y \sum_{j,k} |u_{j+1,k} - u_{j,k}|, \quad (3.8a)$$

we have

$$TV_{x,1}(u) = \Delta y \sum_k \left\{ \sum_j |u_{j+1,k} - u_{j,k}| \right\} = \Delta y \sum_k TV_x u_{\cdot,k}.$$

$TV_{x,1}(u)$  is the one-dimensional (in  $y$ )  $\mathbf{L}^1$ -norm of the sequence  $k \mapsto TV_x u_{\cdot,k}$ , where  $TV_x u_{\cdot,k}$  is the total variation of the sequence  $j \mapsto u_{j,k}$ . Similarly, setting

$$TV_{y,1}(u) = \Delta x \sum_{j,k} |u_{j,k+1} - u_{j,k}|, \quad (3.8b)$$

then

$$TV(u) = TV_{x,1}(u) + TV_{y,1}(u). \quad (3.8c)$$

It is the discrete norm associated with the continuous BV norm (see G.R., Chapter 2, Section 1 [539]). Assuming that the numerical fluxes  $F$  and  $G$  are Lipschitz continuous, we can prove that a scheme (3.4) that is monotone converges to the unique entropy solution. More precisely, we define

$$\Omega_{jk} = (x_{j-1/2}, x_{j+1/2}) \times (y_{k-1/2}, y_{k+1/2})$$

and associate as usual a piecewise constant function  $v_\Delta$  with value  $v_{j,k}^n$  on  $\Omega_{jk}$ ,

$$v_\Delta(\mathbf{x}, t) = v_{j,k}^n, \quad \mathbf{x} \in \Omega_{jk}, \quad t_n < t \leq t_{n+1}, \quad (3.9)$$

and set

$$v_{j,k}^0 = \frac{1}{\Delta x \Delta y} \int_{\Omega_{jk}} u_0(\mathbf{x}) d\mathbf{x},$$

and assume, moreover, that  $u_0 \in BV(\mathbb{R}^2)$ . One can then prove the following result.

*Theorem 3.1*

Assume that the scheme (3.4) is monotone. The sequence  $v_\Delta$  associated by (3.9) converges in  $\mathbf{L}^\infty(0, T; \mathbf{L}_{loc}^1(\mathbb{R}^2))$  to the unique entropy solution of the scalar Cauchy problem (3.6a)–(3.6b) as  $\Delta x \rightarrow 0, \Delta y \rightarrow 0, \Delta t = \lambda_x \Delta x = \lambda_y \Delta y$ , with  $\lambda_x, \lambda_y$  kept constant.

*Proof.* The convergence of monotone conservative schemes is proven exactly as in the one-dimensional case (see G.R., Chapter 3, Theorem 3.4, and Crandall and Majda [372], LeRoux [767], Sanders [1005]). In particular, it can be proven, using Crandall–Tartar’s Lemma (see G.R., Chapter 2, Lemma 5.2, and Chapter 3, Theorem 3.2), that monotone schemes are contractions in  $\mathbf{L}^1$ ,

$$\|H_\Delta(u) - H_\Delta(v)\|_{\mathbf{L}^1(\Delta)} \leq \|u - v\|_{\mathbf{L}^1(\Delta)}.$$

They are also  $L^\infty$ -stable and TVD.  $\square$

*Remark 3.1.* A generalization of this result, the proof of which does not involve BV-estimates but instead uses measure-valued solutions [426], and DiPerna’s uniqueness result, can be found in Szepessy [1084], Coquel, and LeFloch [355, 356]. We shall discuss this result in Remark 3.3 and in Sect. 4.2.3.

For similar results concerning the approximation of conservation laws with source terms, we refer to Chalabi [259].  $\square$

We state now the limiting result concerning the maximum order of a TVD scheme.

*Proposition 3.1*

A scheme (3.3) that is TVD w.r.t. the norm (3.8) is at most first-order accurate.

*Proof.* Goodman and LeVeque [549] have proven that, given a two-dimensional TVD scheme, there exists a set of “one space dimensional” data such that the restriction of the scheme to these data gives a monotone one-dimensional scheme; we refer to Goodman and LeVeque [549] for details.  $\square$

*Remark 3.2.* In fact, in several space dimensions, an estimate of the total variation of either the exact solution of a hyperbolic system (see Rauch [967]) or its approximate solution often fails. For all these reasons, the notion of TVD or TVB (total variation bounded) scheme is not as well adapted as in the one-dimensional scalar case (see also Lemma 3.4).  $\square$

We can still define an incremental form. Setting

$$\begin{cases} \Delta v_{j+1/2,k} = \Delta_x v_{j,k} = v_{j+1,k} - v_{j,k}, \\ \Delta v_{j,k+1/2} = \Delta_y v_{j,k} = v_{j,k+1} - v_{j,k}, \end{cases} \quad (3.10)$$

we say that scheme (3.3) can be put in incremental form if there exist coefficients  $C_x, D_x, C_y, D_y$  (incremental coefficients) such that

$$\begin{cases} v_{j,k}^{n+1} = v_{j,k}^n + C_{x,j+1/2,k}^n \Delta v_{j+1/2,k}^n - D_{x,j-1/2,k}^n \Delta v_{j-1/2,k}^n \\ \quad + C_{y,j,k+1/2}^n \Delta v_{j,k+1/2}^n - D_{y,j,k-1/2}^n \Delta v_{j,k-1/2}^n. \end{cases} \quad (3.11)$$

This form can be useful for proving monotonicity (see Spekreijse [1071]) and  $L^\infty$  estimates.

*Lemma 3.1*

Assume that the scheme (3.3) can be put in incremental form (3.11) and that the incremental coefficients satisfy for any  $j, k, n$

$$\begin{aligned} C_{x,j+1/2,k}^n &\geq 0, \quad D_{x,j-1/2,k}^n \geq 0, \quad C_{y,j,k+1/2}^n \geq 0, \quad D_{y,j,k-1/2}^n \geq 0, \\ C_{x,j+1/2,k}^n + D_{x,j-1/2,k}^n + C_{y,j,k+1/2}^n + D_{y,j,k-1/2}^n &\leq 1. \end{aligned}$$

Then, the scheme is  $L^\infty$ -stable.

*Proof.* By assumption, we can write  $v_{j,k}^{n+1}$  as a convex combination of the values  $v_{j,k}^n, v_{j\pm 1,k}^n, v_{j,k\pm 1}^n$ .  $\square$

*Remark 3.3.* The result holds if the incremental form differs from (3.11) by a term  $E_{j,k}^n$  such that  $n|E_{j,k}^n|$  can be bounded. It has been applied for higher-order schemes built from monotone or  $E$ -schemes following a corrected anti-diffusive flux approach (see G.R., Chapter 4, Section 1): the one-dimensional numerical  $E$ -flux  $F_{j+1/2,k}^n$  (resp.  $G_{j,k+1/2}^n$ ) is corrected by an antidiffusive flux  $h_{j+1/2,k}^n$  (resp.  $k_{j,k+1/2}^n$ )

$$\tilde{F}_{j+1/2,k}^n = \frac{F_{j+1/2,k}^n + h_{j+1/2,k}^n}{\lambda_x}, \quad \tilde{G}_{j,k+1/2}^n = \frac{G_{j,k+1/2}^n + k_{j,k+1/2}^n}{\lambda_y}.$$

This  $L^\infty$ -estimate is used by Coquel and LeFloch [356] to prove the convergence of this type of schemes. As we shall detail in Sect. 4.2.3, the  $L^\infty$ -estimate ensures that a Young measure [74],  $\nu$ , can be constructed from the family  $(u_h)$ . However, since  $L^\infty$  stability is not sufficient to ensure the convergence of the nonlinear terms, some other estimates are needed (which are linked to the local entropy production of the scheme). We just sketch the ideas of the proof. The  $E$ -schemes satisfy, for any convex entropy  $U$  and the associated numerical entropy fluxes  $\psi_x$  and  $\psi_y$ , a discrete entropy inequality (see G.R., Chapter 3, Theorem 4.3)

$$\begin{aligned} U(v_{j,k}^{n+1}) &\leq U(v_{j,k}^n) - \lambda_x(\psi_{x,j+1/2,k}^n - \psi_{x,j-1/2,k}^n) \\ &\quad - \lambda_y(\psi_{y,j,k+1/2}^n - \psi_{y,j,k-1/2}^n). \end{aligned}$$

Following some ideas of the proof of Theorem 1.1 in G.R., Chapter 4, the authors show for the modified scheme an estimate of the form

$$\begin{aligned} U(v_{j,k}^{n+1}) &\leq U(v_{j,k}^n) - \lambda_x(\psi_{x,j+1/2,k}^n - \psi_{x,j-1/2,k}^n) \\ &\quad - \lambda_y(\psi_{y,j,k+1/2}^n - \psi_{y,j,k-1/2}^n) + R_{j,k}^n. \end{aligned}$$

The term  $R_{j,k}^n$  is shown to tend to zero. More precisely, it is evaluated by obtaining a sharp evaluation of the entropy production generated by the whole scheme for the particular entropy  $U(u) = \frac{u^2}{2}$ . Note that the proof is based on a decomposition of the scheme in a form analogous to (3.6a), i.e., as a convex combination of the two one-dimensional schemes

$$\begin{aligned} v_{j,k}^{n+1} &= \frac{1}{2}\{v_{j,k}^n - 2\lambda_x(F_{j+1/2,k}^n - F_{j-1/2,k}^n)\} \\ &\quad + \frac{1}{2}\{v_{j,k}^n - 2\lambda_y(G_{j,k+1/2}^n - G_{j,k-1/2}^n)\}. \end{aligned}$$

The sharp estimate of the entropy production leads to a weak uniform BV-estimate

$$\begin{aligned} \Delta t \sum_{j \leq T/\Delta t} TV(v^n) &= \Delta t \sum_n \sum_{j,k} \{\Delta y |v_{j+1,k} - v_{j,k}| + \Delta x |v_{j,k+1} - v_{j,k}|\} \\ &\leq C \Delta t^{-2/3}; \end{aligned}$$

it does not imply a compactness property but is enough to pass to the limit, which proves that the Young measure  $\nu$  associated with  $u_h$  is indeed a measure-valued solution (see Definition 4.1 below). Together with DiPerna's uniqueness result (see Remark 3.1), the discrete entropy inequalities imply that  $\nu$  is the unique entropy weak  $L^\infty$ -solution.  $\square$

### 3.2 Dimensional Splitting

However, still assuming a Cartesian grid, another popular way of constructing multidimensional schemes starting from one-dimensional schemes is to use a splitting method. The most practical calculations use a dimensional splitting or alternating direction technique (see Beam and Warming [95, 1179], Woodward and Colella [1190], Yee, Warming and Harten [1209], Daru and Lerat [389], van Leer and Mulder [1160], Le Gruyer and LeRoux [748], Glaister [525], Osher and Solomon [922], Clarke et al. [313]), even if it may create spurious waves [984]. Note that a few attempts have been made to use operator splitting in some particular cases (Dukowicz and Dvinsky [445], Baraille et al. [82], Buffard and Hérard [216], Buffard [212], Liou and Steffen [817], Chalabi and Vila [262]).

Let us first present the general time-stepping technique, without specifying the space discretization or the chosen decomposition. Suppose that we have discretized in some way the terms  $\frac{\partial}{\partial x} f(u)$  and  $\frac{\partial}{\partial y} g(u)$ , and consider the method of lines. We shall even restrict ourselves to the simple case of a linear differential system

$$\frac{du}{dt} + (A + B)u = 0, \tag{3.12}$$

where  $A$  and  $B$  are constant matrices. The exact solution of (3.12) for given  $u(t_0) = u_0$  is simply  $u(t) = e^{-(A+B)t}u_0$ . The simplest procedure consists of two steps. Assume that we have an approximation  $u^n$  of  $u$  at time  $t_n$ . In order to compute  $u^{n+1}$ , in the first step one solves the equations

$$\begin{cases} \frac{du}{dt} + Au = 0, & t \in (t_n, t_{n+1}), \\ u(t_n) = u^n, \end{cases} \quad (3.13a)$$

which give  $u^{n,1} = u(t_{n+1}) = e^{-\Delta t A}u^n$ .

In the second step, one solves

$$\begin{cases} \frac{du}{dt} + Bu = 0, \\ u(t_n) = u^{n,1}, \end{cases} \quad (3.13b)$$

and then one takes

$$u^{n+1} = u(t_{n+1}) = e^{-\Delta t B}u^{n,1} = e^{-\Delta t B}e^{-\Delta t A}u^n. \quad (3.14)$$

Thus, we have

$$u^{n+1} = (e^{-\Delta t B}e^{-\Delta t A})^n u^0.$$

This leads to a first-order-in-time method, as we shall see later. A more subtle timestepping introduces three steps: one solves (3.13a) during a half timestep, then (3.13b) on a timestep, and last (3.13a) on a half timestep. Hence, we first solve

$$\begin{cases} \frac{du}{dt} + Au = 0, & t \in (t_n, t_{n+1/2}), \\ u(t_n) = u^n, \end{cases} \quad (3.15a)$$

which gives

$$u^{n,1} = u(t_{n+1/2}) = e^{-\Delta t A/2}u^n.$$

In the second step, one solves

$$\begin{cases} \frac{dv}{dt} + Bv = 0, & t \in (t_n, t_{n+1}), \\ v(t_n) = u^{n,1}, \end{cases} \quad (3.15b)$$

which gives

$$u^{n,2} = v(t_{n+1}) = e^{-\Delta t B}u^{n,1}.$$

Finally,

$$\begin{cases} \frac{dw}{dt} + Aw = 0, & t \in (t_{n+1/2}, t_{n+1}), \\ w(t_{n+1/2}) = u^{n,2}, \end{cases} \quad (3.15c)$$

which yields

$$u^{n+1} = w(t_{n+1}) = e^{-\Delta t A / 2} u^{n,2} = e^{-\Delta t A / 2} e^{-\Delta t B} e^{-\Delta t A / 2} u^n. \quad (3.16)$$

### Lemma 3.2

The scheme (3.13a)–(3.13b), (3.14) yields a first-order accurate method if the matrices  $A$  and  $B$  do not commute, whereas (3.15), (3.16) is second-order accurate.

*Proof.* We use the Taylor expansion of the exponential function

$$\begin{aligned} & e^{-\Delta t A / 2} e^{-\Delta t B} e^{-\Delta t A / 2} \\ &= \left( I - \frac{\Delta t}{2} A + \frac{\Delta t^2}{8} A^2 + \dots \right) \left( I - \Delta t B + \frac{\Delta t^2}{2} B^2 + \dots \right) \\ & \quad \left( I - \frac{\Delta t}{2} A + \frac{\Delta t^2}{8} A^2 + \dots \right) \\ &= I - \Delta t(A + B) + \frac{\Delta t^2}{2}(A^2 + AB + BA + B^2) + \dots \\ &= e^{-\Delta t(A+B)} + O(\Delta t^3), \end{aligned}$$

which proves that the resulting three-timestep scheme is second-order accurate (this is Strang's result). If, however, matrices  $A$  and  $B$  do not commute,

$$\begin{aligned} e^{-\Delta t B} e^{-\Delta t A} &= \left( I - \frac{\Delta t}{2} B + \frac{\Delta t^2}{2} B^2 + \dots \right) \left( I - \Delta t A + \frac{\Delta t^2}{2} A^2 + \dots \right) \\ &= I - \Delta t(A + B) + \frac{\Delta t^2}{2}(A^2 + 2BA + B^2) + \dots \\ &= e^{-\Delta t(A+B)} + O(\Delta t^2), \end{aligned}$$

and the two-step scheme is only first-order accurate.  $\square$

*Remark 3.4.* The convergence of scheme (3.13a)–(3.13b) is a very simple example of the application of the Trotter formula, which holds for generators of unbounded operators  $A$ ,  $B$  in a Banach space,

$$\exp\{t(A + B)\} = \lim_{n \rightarrow \infty} \left\{ \exp\left(\frac{tA}{n}\right) \exp\left(\frac{tB}{n}\right) \right\}^n.$$

(For a detailed result and proof, see Chorin et al. [306]).  $\square$

Let us now study the application for a two-dimensional conservation law (3.6a)–(3.6b)

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) + \frac{\partial}{\partial y} g(u) = 0,$$

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

Using the notations of G.R., Chapter 2, Section 5, we can write the unique entropy solution as  $u(\cdot, t) = S(t)u_0$ , where  $S(t)$  is the (exact) solution operator. We consider the one-dimensional conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0$$

and

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial y} g(u) = 0.$$

The associated solution operators  $S_x(t)$  and  $S_y(t)$  take the place of the above exponential function. Thus, the fractional step (dimensional splitting) method approximates  $u(x, y, n\Delta t) = S(n\Delta t)u_0(x, y)$  by  $(S_y(\Delta t)S_x(\Delta t))^n u_0(x, y)$ . (The accuracy of this semi-discrete splitting is studied in Teng [1114].)

In order to define an alternate-direction, fully discrete scheme, we introduce finite difference one-dimensional schemes,

$$v_j^{n+1} = H_x(v_{j-J}^n, \dots, v_{j+J}^n), \quad (3.17a)$$

$$v_j^{n+1} = H_y(v_{j-K}^n, \dots, v_{j+K}^n), \quad (3.17b)$$

and we set

$$\begin{cases} v_{j,k}^{n,1} = H_x(v_{j-J,k}^n, \dots, v_{j+J,k}^n), \\ v_{j,k}^{n+1} = H_y(v_{j,k-K}^{n,1}, \dots, v_{j,k+K}^{n,1}). \end{cases} \quad (3.18)$$

*Lemma 3.3*

Assume that the difference schemes (3.17) are in conservation form and consistent with  $f$  and  $g$ , respectively. Then, the resulting scheme (3.18) can be put in conservation form and is consistent with the 2d conservation law. Moreover, in the scalar case, if the schemes are monotone, the scheme (3.18) is also monotone.

*Proof.* By assumption on scheme (3.17a), we can write  $v_{j,k}^{n,1}$  in the form

$$v_{j,k}^{n,1} = v_{j,k}^n - \lambda_x \{F_{j+1/2,k}^n - F_{j-1/2,k}^n\},$$

and similarly,

$$v_{j,k}^{n+1} = v_{j,k}^{n,1} - \lambda_y \{G_{j,k+1/2}^{n,1} - G_{j,k-1/2}^{n,1}\},$$

where

$$G_{j,k+1/2}^{n,1} = G(v_{j,k-K+1}^{n,1}, \dots, v_{j,k+K}^{n,1}).$$

Thus, (3.18) yields

$$v_{j,k}^{n+1} = v_{j,k}^n - \lambda_x \{F_{j+1/2,k}^n - F_{j-1/2,k}^n\} - \lambda_y \{G_{j,k+1/2}^{n,1} - G_{j,k-1/2}^{n,1}\}.$$

Substituting the expressions for  $v_{j,k}^{n,1}$  in  $G$ , we check that we can put the scheme in conservation form. Indeed, we have

$$G_{j,k+1/2}^{n,1} = \tilde{G}(v_{j-J,k-K+1}^n, \dots, v_{j+J,k+K}^n) = \tilde{G}_{j,k+1/2}^n$$

for some function  $\tilde{G}$ . It is also consistent: taking

$$v_{j-J,k-K+1}^n = \dots = v_{j+J,k+K}^n = u,$$

we have by the consistency of  $F$  with  $f$ ,

$$v_{j,k-K+1}^{n,1} = \dots = v_{j,k+K}^{n,1} = u,$$

which implies by the consistency of  $G$  with  $g$

$$G(u, \dots, u) = g(u).$$

The fact that the resulting scheme (3.18) is monotone if both schemes (3.17) are monotone is easily proven. Since  $H_x$  and  $H_y$  are nondecreasing functions of each of their arguments,

$$H_\Delta(v_{j-J,k-K}^n, \dots, v_{j+J,k+K}^n) = H_y(H_x(\dots), \dots, H_x(\dots))$$

is also a nondecreasing function.  $\square$

Is it possible to prove an analogous property for a scheme split from TVD schemes? In order to study this question, we assume that the schemes (3.17) are TVD and can be written in incremental form,

$$\begin{aligned} v_j^{n+1} &= H_x(v_{j-J}^n, \dots, v_{j+J}^n) \\ &= v_j^n + C_{x,j+1/2}^n \Delta v_{j+1/2}^n - D_{x,j-1/2}^n \Delta v_{j-1/2}^n \end{aligned} \tag{3.19a}$$

$$\begin{aligned} v_j^{n+1} &= H_y(v_{j-K}^n, \dots, v_{j+K}^n) \\ &= v_j^n + C_{y,j+1/2}^n \Delta v_{j+1/2}^n - D_{y,j-1/2}^n \Delta v_{j-1/2}^n, \end{aligned} \tag{3.19b}$$

where the coefficients  $C, D$  are defined by

$$C_{x,j+1/2}^n = C_x(v_{j-J+1}^n, \dots, v_{j+J}^n)$$

for some function  $C_x : \mathbb{R}^{2J} \rightarrow \mathbb{R}$  and so on. We can define an alternate-direction scheme by (3.18). Let us consider the simple linear constant coefficient case

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = 0$$

and assume moreover that the coefficients  $C_{x,j+1/2}^n, D_{x,j+1/2}^n, C_{y,j+1/2}^n, D_{y,j+1/2}^n$  do not depend on  $j, k, n$  and satisfy the TVD property (see G.R., Chapter 3, (3.24)), i.e.,

$$C_{x,j+1/2}^n = C_x, \quad C_{y,j+1/2}^n = C_y, \quad D_{x,j+1/2}^n = D_x, \quad D_{y,j+1/2}^n = D_y \quad (3.20a)$$

with

$$C \geq 0, \quad D \geq 0, \quad C + D = Q \leq 1. \quad (3.20b)$$

This occurs if schemes (3.17a) and (3.17b) coincide and are, for instance, three-point conservative consistent schemes (see Chap. IV, Sect. 1 and GR Chapter 3, formula (1.38)). Even in this simple case, we cannot prove that the resulting scheme is TVD, only that the total variation is bounded. More precisely, we obtain then the following estimate.

*Lemma 3.4*

Assume that the difference schemes (3.17) are TVD and that (3.20) holds. Then, scheme (3.18) satisfies

$$TV(v^{n+1}) \leq TV(v^n) + (\Delta x + \Delta y) \sum_{i,j} |\Delta_x \Delta_y v_{i,j}^n|.$$

*Proof.* Let us write the schemes (3.17) in incremental form,

$$v_{j,k}^{n,1} = H_x(v_{j-J,k}^n, \dots, v_{j+J,k}^n) = v_j^n + C_x \Delta_x v_{j,k}^n - D_x \Delta_x v_{j-1,k}^n$$

where we have used the notations (3.10), and

$$v_{j,k}^{n+1} = H_y(v_{j,k-K}^{n,1}, \dots, v_{j,k+K}^{n,1}) = v_{j,k}^{n,1} + C_y \Delta_y v_{j,k}^{n,1} - D_y \Delta_y v_{j,k-1}^{n,1}.$$

It is easy to prove that the operators  $\Delta_x$  and  $\Delta_y$  commute. Indeed, by (3.20), we have

$$\begin{aligned} \Delta_x \Delta_y v_{j,k} &= \Delta_x(v_{j,k+1} - v_{j,k}) = v_{j+1,k+1} - v_{j,k+1} - v_{j+1,k} + v_{j,k} \\ &= \Delta_y(v_{j+1,k} - v_{j,k}) = \Delta_y \Delta_x v_{j,k}. \end{aligned}$$

By assumption (3.20), this yields that the operators  $H_x$ ,  $H_y$ ,  $\Delta_x$ , and  $\Delta_y$  also commute; for instance, the two expressions

$$\begin{aligned} \Delta_y H_x v_{j,k} &= \Delta_y(v_{j,k} + C_x \Delta_x v_{j,k} - D_x \Delta_x v_{j-1,k}) \\ &= \Delta_y v_{j,k} + C_x \Delta_y \Delta_x v_{j,k} - D_x \Delta_y \Delta_x v_{j-1,k} \end{aligned} \quad (3.21a)$$

and

$$H_x \Delta_y u_{j,k} = \Delta_y v_{j,k} + C_x \Delta_x \Delta_y v_{j,k} - D_x \Delta_x \Delta_y v_{j-1,k} \quad (3.21b)$$

coincide. Then, by (3.8),

$$\begin{aligned} TV(v_{j,k}) &= \Delta y \sum_{j,k} |\Delta_x v_{j,k}| + \Delta x \sum_{j,k} |\Delta_y v_{j,k}| \\ &= \Delta y \sum_k TV_x(v_{\cdot,k}) + \Delta x \sum_j TV_y(v_{j,\cdot}), \end{aligned}$$

where  $TV_x$  or  $TV_y$  denotes the (one-dimensional) TV norm (see Chap. IV, Sect. 1). Thus, since the operators commute, we can write

$$TV(H_y H_x v_{j,k}) = \Delta y \sum_k TV_x(H_x H_y v)_{\cdot,k} + \Delta x \sum_j TV_y(H_y H_x v)_{j,\cdot},$$

and if the one-dimensional schemes are TVD,

$$\begin{aligned} TV(H_y H_x v_{j,k}) &\leq \Delta y \sum_k TV_x(H_y v)_{\cdot,k} + \Delta x \sum_j TV_y(H_x v)_{j,\cdot} \\ &= \Delta y \sum_{j,k} |\Delta_x H_y v_{j,k}| + \Delta x \sum_{j,k} |\Delta_y H_x v_{j,k}|. \end{aligned}$$

By (3.21) and (3.20b), we obtain, setting  $Q = C + D$

$$\begin{aligned} &\sum_{j,k} |\Delta_y H_x v_{j,k}| \\ &\leq \sum_{j,k} |\Delta_y v_{j,k}| + C_x \sum_{j,k} |\Delta_x \Delta_y v_{j,k}| + D_x \sum_{j,k} |\Delta_x \Delta_y v_{j-1,k}| \\ &\leq \sum_{j,k} |\Delta_y v_{j,k}| + Q_x \sum_{j,k} |\Delta_x \Delta_y v_{j,k}|. \end{aligned}$$

Thus

$$TV(H_y H_x v_{j,k}) \leq TV(v_{j,k}) + (Q_x \Delta x + Q_y \Delta y) \sum_{j,k} |\Delta_x \Delta_y v_{j,k}|,$$

which gives the result

$$TV(v^{n+1}) \leq TV(v^n) + (\Delta x + \Delta y) \sum_{j,k} |\Delta_x \Delta_y v_{j,k}^n|.$$

Now, using again the fact that both schemes are TVD and that the operators commute, we can write

$$\begin{aligned} \sum_{j,k} |\Delta_x \Delta_y v_{j,k}^n| &= \sum_{j,k} |\Delta_x \Delta_y H_y H_x v_{j,k}^{n-1}| = \sum_{j,k} |\Delta_x H_x \Delta_y H_y v_{j,k}^{n-1}| \\ &\leq \sum_{j,k} |\Delta_x \Delta_y H_y v_{j,k}^{n-1}| = \sum_{j,k} |\Delta_y H_y \Delta_x v_{j,k}^{n-1}| \leq \sum_{j,k} |\Delta_y \Delta_x v_{j,k}^{n-1}|. \end{aligned}$$

We obtain, then, that for  $n \leq \frac{T}{\Delta t}$ ,

$$TV(v^n) \leq TV(v^0) + T \left( \frac{Q_x}{\lambda_x} + \frac{Q_y}{\lambda_y} \right) \sum_{j,k} |\Delta_x \Delta_y v_{j,k}^0|,$$

and thus the scheme is TVB (the total variation is bounded).  $\square$

For instance, if we take the same three-point upwind scheme in each direction,

$$C_x + D_x = Q_x = \lambda_x |a|, \quad C_y + D_y = Q_y = \lambda_y |b|,$$

we get

$$TV(v^n) \leq TV(v^0) + T(|a| + |b|) \sum_{j,k} |\Delta_x \Delta_y v_{j,k}^0|.$$

*Remark 3.5.* In fact, following the ideas of Lemma 3.2, when using an alternating direction scheme, one can solve (3.13a) during a half timestep  $\frac{\Delta t}{2}$ , then (3.13b) during a timestep  $\Delta t$ , and finally (3.13a) during a half timestep  $\frac{\Delta t}{2}$ . It is then necessary to preserve the global order of discretization to use a second-order (in time) method such as a Runge–Kutta method or an implicit multistep method (see Beam and Warming [95]) instead of the backward Euler method. Anyway, the drawbacks of such alternating direction techniques are obvious since the grid directions play an overdetermined role. Stability is studied in Serre [1033].  $\square$

## 4 Finite-Volume Methods

Actually, the most usual extensions include a finite-element or a finite-volume formulation on structured or unstructured meshes. In structured (Cartesian or curvilinear) meshes, there exist locally two axes, the center of the cell admits a natural parametrization by  $(i, j)$ , and each cell is surrounded by a fixed number of neighboring cells. This allows ADI techniques and easy implementation on vector computers. Unstructured meshes, with the use of an automatic mesh generator and adaptive grid refinement [323], offer a greater flexibility when dealing with complex geometries and limit grid orientation effects. So let us consider now the general case of an unstructured mesh, which means that we do not consider a rectangular grid  $\Delta$  as in the previous section but a “triangulation”  $\mathcal{T}_h$  of the computational domain  $\mathcal{O}$  by triangles or quadrilaterals.

## 4.1 Definition of the Finite-Volume Method

### 4.1.1 General Principles

In a *finite-volume* method, the computational domain  $\mathcal{O}$  is composed of control cells, or control volumes,  $\Omega_i$  with center  $\mathbf{c}_i$ , which are either the elements of the triangulation or constructed from these elements (see Examples 4.1 and 4.2). On  $\Omega_i$ ,  $\mathbf{u}(\cdot, t)$  is approximated by a constant  $\mathbf{u}_i(t)$ , which should be considered as an approximation of the mean value of  $\mathbf{u}$  over the cell  $\Omega_i$  rather than of the value at point  $\mathbf{c}_i$ ,

$$\mathbf{u}_i(t) \cong \frac{1}{|\Omega_i|} \int_{\Omega_i} \mathbf{u}(\mathbf{x}, t) d\mathbf{x},$$

where  $|\Omega_i|$  denotes the area of  $\Omega_i$ . The differential system defining  $\mathbf{u}_i(t)$  is obtained as follows: first, integrating the system (3.2) over  $\Omega_i$ ,

$$\int_{\Omega_i} \left( \frac{\partial \mathbf{u}}{\partial t} + \operatorname{div} \mathbb{F}(\mathbf{u}) \right) d\mathbf{x} = 0, \quad \mathbb{F} = (\mathbf{f}, \mathbf{g})^T,$$

yields

$$\frac{\partial}{\partial t} \left( \int_{\Omega_i} \mathbf{u}(\mathbf{x}, t) d\mathbf{x} \right) + \int_{\partial\Omega_i} \mathbb{F}(\mathbf{u}(\cdot, t)) \cdot \mathbf{n}_i d\sigma = \mathbf{0}, \quad (4.1)$$

where  $\partial\Omega_i$  is the boundary of  $\Omega_i$  and  $\mathbf{n}_i$  the outward unit normal vector to  $\Omega_i$  (and the dot “.” stands for the product  $\mathbb{F} \cdot \mathbf{n} = \cos \theta \mathbf{f} + \sin \theta \mathbf{g}$  if  $\mathbf{n} = (\cos \theta, \sin \theta)^T$ ). The first term in (4.1) is naturally approximated by

$$\frac{\partial}{\partial t} \left( \int_{\Omega_i} \mathbf{u}(\mathbf{x}, t) d\mathbf{x} \right) \cong |\Omega_i| \frac{\partial \mathbf{u}_i(t)}{\partial t}.$$

Since the approximation is not continuous across  $\partial\Omega_i$ , we have to discretize the “residual”  $\int_{\partial\Omega_i} \mathbb{F}(u(\cdot, t_n)) \cdot \mathbf{n}_i d\sigma$ , which represents the flux across the boundary of the cell at time  $t_n$ . It can be written

$$\int_{\partial\Omega_i} \mathbb{F}(\mathbf{u}) \cdot \mathbf{n}_i d\sigma = \sum_{e \subset \partial\Omega_i, e = \Gamma_{ij}} \int_{\Gamma_{ij}} \mathbb{F}(\mathbf{u}) \cdot \mathbf{n}_i d\sigma,$$

where the sum is taken over all the edges  $e$  of the cell, and  $\partial\Omega_i = \cup \Gamma_{ij}$ , where  $\Gamma_{ij} = \Omega_i \cap \Omega_j$  is the face separating  $\Omega_i$  and  $\Omega_j$ . Note that the control cells are usually assumed to satisfy the properties of a finite-element triangulation: the  $\Omega_i$ s are nonoverlapping sets and, if  $e$  is a given edge of  $\partial\Omega_i$ , there exists a unique  $\Omega_j$  such that  $e = \Omega_i \cap \Omega_j$ .

The problem is then to define the numerical fluxes approximating the interface flux  $\int_{\Gamma_{ij}} \mathbb{F}(\mathbf{u}) \cdot \mathbf{n}_i d\sigma$ , using only the values  $\mathbf{u}_i(t)$ . In fact, we shall only detail here the case of internal fluxes and not the fluxes at the boundary

of  $\mathcal{O}$  (for instance, solid wall or inflow–outflow conditions will be considered later on in Chap. VI). The usual way consists in introducing a function  $\Phi$  such that for  $e = \Gamma_{ij} \subset \partial\Omega_i$

$$\int_{\Gamma_{ij}} \mathbb{F}(\mathbf{u}) \cdot \mathbf{n}_i \, d\sigma \cong |e| \Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}_e), \quad (4.2)$$

where  $\mathbf{n}_e$  denotes the unit normal to  $e$  pointing in the direction of  $\Omega_j$  (thus outward to  $\Omega_i$ ) and  $|e|$  the length of  $e$ . Though the notation for the numerical flux is not completely satisfying, it means that we have assumed that the numerical flux depends “only” on the values on each side of the edge and on the normal direction to the edge (it depends also on the continuous flux  $\mathbb{F}$ ). This yields a “method of lines”

$$|\Omega_i| \frac{\partial \mathbf{u}_i(t)}{\partial t} + \sum_{e \subset \partial\Omega_i, e=\Gamma_{ij}} |e| \Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}_e) = \mathbf{0}.$$

Eventually, we approximate this ordinary differential system by the explicit Euler scheme to obtain the formula

$$|\Omega_i| (\mathbf{u}_i^{n+1} - \mathbf{u}_i^n) + \Delta t \left\{ \sum_{e \subset \partial\Omega_i, e=\Gamma_{ij}} |e| \Phi(\mathbf{u}_i^n, \mathbf{u}_j^n, \mathbf{n}_e) \right\} = \mathbf{0}, \quad (4.3)$$

where  $\mathbf{u}_i^n \cong \mathbf{u}_i(t_n)$  and where  $\mathbf{u}_i^0$  is given.

### 4.1.2 Properties

In the general case, the numerical fluxes  $\Phi$  are assumed to be locally Lipschitz continuous and must satisfy some conditions:

*Conservation:*

$$\Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}) = -\Phi(\mathbf{u}_j, \mathbf{u}_i, -\mathbf{n}). \quad (4.4)$$

This property (which is directly inherited from the continuous flux (4.2)) means that, in the absence of source term, the approximate flux at the interface separating  $\Omega_i$  and  $\Omega_j$  is the *same* as the flux at this same interface separating  $\Omega_j$  and  $\Omega_i$  (since  $-\mathbf{n}_e$  is the unit normal to  $e$  pointing in the direction of  $\Omega_i$ ), i.e., there is only one exchange term per edge  $e$  separating  $\Omega_i$  and  $\Omega_j$ . In the one-dimensional case, it reduces to the fact that we can write the numerical flux at the interface  $x_{j+1/2}$  of the cells  $(x_{j-1/2}, x_{j+1/2})$  and  $(x_{j+1/2}, x_{j+3/2})$  in the form  $g_{j+1/2}$ . As previously, conservation ensures that when the scheme converges (in a strong way), the limit satisfies the Rankine–Hugoniot condition and is thus a weak solution of the conservation law (see Proposition 4.1).

*Consistency:*

$$\Phi(\mathbf{u}, \mathbf{u}, \mathbf{n}) = \mathbb{F}(\mathbf{u}) \cdot \mathbf{n}. \quad (4.5)$$

Again, this is natural from (4.2).

Usually, the flux  $\Phi(\mathbf{u}_i^n, \mathbf{u}_j^n, \mathbf{n})$  is defined by solving exactly or approximately a one-dimensional Riemann problem, in the direction  $\mathbf{n}$  normal to the edge  $e = \Gamma_{ij}$ , associated to the (continuous) flux  $\mathbb{F}(\mathbf{u}) \cdot \mathbf{n}$ . More precisely, we define as in Sect. 1.1 or 2.2 new variables  $\zeta$  (normal) and  $\tau$  (tangential) by

$$\zeta = \cos \theta x + \sin \theta y = \mathbf{X} \cdot \mathbf{n}, \quad \tau = -\sin \theta x + \cos \theta y = \mathbf{X} \cdot \mathbf{n}^\perp,$$

where  $\theta$  is the angle of the normal to an edge with the  $x$ -axis, i.e.,  $\mathbf{n} = (\cos \theta, \sin \theta)$ ,  $\mathbf{n}^\perp = (-\sin \theta, \cos \theta)$  is directly orthogonal to  $\mathbf{n}$ , and  $\mathbf{X} = (x, y)$ . The system (3.2) is transformed into

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial}{\partial \zeta} (\mathbb{F} \cdot \mathbf{n})(\mathbf{v}) + \frac{\partial}{\partial \tau} (\mathbb{F} \cdot \mathbf{n}^\perp)(\mathbf{v}) = \mathbf{0},$$

where

$$\mathbf{v}(\zeta, \tau, t) = \mathbf{u}(x(\zeta, \tau), y(\zeta, \tau), t)$$

and

$$\mathbb{F} \cdot \mathbf{n} = \cos \theta \mathbf{f} + \sin \theta \mathbf{g}, \quad \mathbb{F} \cdot \mathbf{n}^\perp = \sin \theta \mathbf{f} + \cos \theta \mathbf{g},$$

i.e.,

$$R^{-1}\mathbb{F} = (\mathbb{F} \cdot \mathbf{n}, \mathbb{F} \cdot \mathbf{n}^\perp)^T$$

if  $R$  is the rotation with angle  $\theta$  in  $\mathbb{R}^2$ .

Now, if  $\mathbf{u}$  is a given constant on each side of the line  $\zeta = 0$ , the associated Cauchy problem reduces to solving the one-dimensional Riemann projected problem in the direction  $\mathbf{n}$ ,

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial}{\partial \zeta} (\mathbb{F} \cdot \mathbf{n})(\mathbf{v}) = 0,$$

$$\mathbf{v}(\zeta, 0) = \begin{cases} \mathbf{u}_i, & \zeta < 0, \\ \mathbf{u}_j, & \zeta > 0, \end{cases}$$

and the flux through the (extended) edge is  $\mathbb{F} \cdot \mathbf{n}(\mathbf{w}_R(0; \mathbf{u}_i, \mathbf{u}_j))$ .

Thus, we shall take more generally a one-dimensional numerical flux  $\varphi(\mathbf{u}, \mathbf{v})$  associated to a three-point difference scheme. We introduce in the notations the dependence on the continuous flux (the conservative scheme with numerical flux  $\varphi(\mathbf{u}, \mathbf{v}) = \varphi(\mathbf{f}; \mathbf{u}, \mathbf{v})$  approximates the system  $\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = 0$ , and  $\varphi$  is consistent with  $\mathbf{f}$ ). We define  $\Phi$  by

$$\Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}) = \varphi(\mathbb{F} \cdot \mathbf{n}; \mathbf{u}_i, \mathbf{u}_j). \quad (4.6)$$

The finite-volume method is said to be *monotone* if this underlying flux is that of a three-point monotone scheme (see G.R., Chapter 3, Section 3). This gives a “first-order” method.

*Remark 4.1.* Some care must be taken in the notations. For instance, if we choose as above the Godunov flux

$$\varphi(\mathbf{f}; \mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{w}_R(0; \mathbf{u}, \mathbf{v}))$$

and then define

$$\Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}) = \mathbb{F} \cdot \mathbf{n}(\mathbf{w}_R(0; \mathbf{u}_i, \mathbf{u}_j)) = R\mathbf{f}(\mathbf{w}_R(0; \tilde{\mathbf{u}}_i, \tilde{\mathbf{u}}_j)),$$

the conservation property is not satisfied since in general

$$\mathbf{f}(\mathbf{w}_R(0; \mathbf{u}, \mathbf{v})) \neq \mathbf{f}(\mathbf{w}_R(0; \mathbf{v}, \mathbf{u})).$$

We must in fact introduce  $\mathbf{f}$  in the notations  $\mathbf{w}_R(0; \mathbf{u}, \mathbf{v}) = \mathbf{w}_R(\mathbf{f})(0; \mathbf{u}, \mathbf{v})$  for the solution of the Riemann problem associated to  $\mathbf{f}$  and the Riemann data  $\mathbf{u}$  and  $\mathbf{v}$ . Indeed, when we change  $\mathbf{n}$  to  $-\mathbf{n}$ , it corresponds after the rotation  $R$  to a change in the orientation of the  $x$ -axis or equivalently to changing  $\mathbf{f}$  to  $-\mathbf{f}$ . Now, we observe that

$$\mathbf{w}_R(-\mathbf{f})(0; \mathbf{v}, \mathbf{u}) = \mathbf{w}_R(\mathbf{f})(0; \mathbf{u}, \mathbf{v}),$$

and thus

$$\begin{aligned} \Phi(\mathbf{u}_j, \mathbf{u}_i, -\mathbf{n}) &= -\mathbb{F} \cdot \mathbf{n}(\mathbf{w}_R(-\mathbb{F} \cdot \mathbf{n})(0; \mathbf{u}_j, \mathbf{u}_i)) \\ &= -\mathbb{F} \cdot \mathbf{n}(\mathbf{w}_R(\mathbb{F} \cdot \mathbf{n})(0; \mathbf{u}_i, \mathbf{u}_j)) = -\Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}), \end{aligned}$$

so that the flux is indeed conservative.  $\square$

*Remark 4.2.* The work that is going on today concerning the study of the solution of the two-dimensional Riemann problem (see Remark 2.8 in this chapter) may lead to the derivation of truly bidimensional schemes in the spirit of Godunov’s scheme.  $\square$

If we want to apply the scheme to the gas dynamics equations, we can, moreover, require that  $\Phi$  be invariant under rotation. Setting for a rotation  $R$  in  $\mathbb{R}^2$ , as in (2.12),

$$R(a, \mathbf{b}, c)^T = (a, R\mathbf{b}, c)^T,$$

so that  $R$  denotes either the  $2 \times 2$  matrix (2.7) or the  $4 \times 4$  matrix

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta & 0 \\ 0 & \sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

depending on whether it acts on  $\mathbb{R}^2$  or  $\mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}$ , we have the obvious definition of *rotational invariance*:

$$\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) = R\Phi(R^{-1}\mathbf{U}_i, R^{-1}\mathbf{U}_j, R^{-1}\mathbf{n}), \quad (4.7)$$

for any rotation  $R$  in  $\mathbb{R}^2$ .

Recall that (see Sect. 2.2, Lemma 2.4, and Remark 2.3) if  $\mathbf{U} = (\rho, \rho\mathbf{u}, \rho e)^T$ ,

$$\tilde{\mathbf{U}} = R^{-1}\mathbf{U} = (\rho, \rho\tilde{\mathbf{u}}, \rho e)^T, \quad \text{where } \tilde{\mathbf{u}} = (u_n, u_\tau)^T,$$

and if  $\mathbb{F} = (\mathbf{f}, \mathbf{g})^T$

$$\begin{aligned} \mathbb{F} \cdot \mathbf{n} &= \cos \theta \mathbf{f} + \sin \theta \mathbf{g} \\ &= (\rho u_n, \rho u_n u + p \cos \theta, \rho u_n v + p \sin \theta, (\rho e + p) u_n)^T, \\ \mathbb{F} \cdot \mathbf{n}^\perp &= -\sin \theta \mathbf{f} + \cos \theta \mathbf{g} \\ &= (\rho u_\tau, \rho u_\tau u - p \sin \theta, \rho u_\tau v + p \cos \theta, (\rho e + p) u_\tau)^T. \end{aligned}$$

Let  $\mathbf{V}$  be defined as above by

$$\mathbf{V}(\zeta, \tau) = \mathbf{U}(x(\zeta, \tau), y(\zeta(\tau))),$$

so that

$$R^{-1}\mathbf{U}(x, y) = R^{-1}\mathbf{V}(\zeta, \tau) = \tilde{\mathbf{U}}(\zeta, \tau).$$

Letting  $R^{-1}$  act on  $\frac{\partial \mathbf{V}}{\partial t} + \frac{\partial}{\partial \zeta} \mathbb{F}(\mathbf{V}) \cdot \mathbf{n} + \frac{\partial}{\partial \tau} \mathbb{F}(\mathbf{V}) \cdot \mathbf{n}^\perp = \mathbf{0}$  (in fact, it acts only on the two components in the middle), we have seen that

$$\begin{aligned} R^{-1} \left\{ \frac{\partial \mathbf{V}}{\partial t} + \frac{\partial}{\partial \zeta} (\mathbb{F}(\mathbf{V}) \cdot \mathbf{n}) + \frac{\partial}{\partial \tau} (\mathbb{F}(\mathbf{V}) \cdot \mathbf{n}^\perp) \right\} \\ = \frac{\partial \tilde{\mathbf{U}}}{\partial t} + \frac{\partial}{\partial \zeta} \mathbf{f}(\tilde{\mathbf{U}}) + \frac{\partial}{\partial \tau} \mathbf{g}(\tilde{\mathbf{U}}) = \mathbf{0}, \end{aligned}$$

and the invariance of the Euler equations comes from the identities (2.14a) (see Remark 2.3)

$$R^{-1}((\mathbb{F} \cdot \mathbf{n})(\mathbf{U})) = \mathbf{f}(\tilde{\mathbf{U}}), \quad R^{-1}(\mathbb{F} \cdot \mathbf{n}^\perp(\mathbf{U})) = \mathbf{g}(\tilde{\mathbf{U}}).$$

Considering as above a rotation  $R$  with angle  $\theta$  such that  $R^{-1}\mathbf{n} = \mathbf{e}_1$ , where  $\mathbf{e}_1$  is the first basis vector  $\mathbf{e}_1 = (1, 0)$ , we have for a rotational invariant flux satisfying (4.7)

$$\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) = R\Phi(R^{-1}\mathbf{U}_i, R^{-1}\mathbf{U}_j, \mathbf{e}_1) = R\Phi(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{e}_1).$$

If we denote by  $(\Phi_1, \Phi_2, \Phi_3, \Phi_4)$  the four components of  $\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n})$  and by  $(\tilde{\Phi}_1, \tilde{\Phi}_2, \tilde{\Phi}_3, \tilde{\Phi}_4)$  those of  $\Phi(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{e}_1)$ , rotational invariance supposes  $\Phi_1 = \tilde{\Phi}_1, \Phi_4 = \tilde{\Phi}_4$ , while  $\Phi_2, \Phi_3$  are deduced from  $\tilde{\Phi}_2, \tilde{\Phi}_3$  by a rotation with angle  $\theta$ ,

$$(\Phi_2, \Phi_3)^T = R(\tilde{\Phi}_2, \tilde{\Phi}_3)^T = (\cos \tilde{\Phi}_2 + \sin \tilde{\Phi}_3, -\sin \theta \Phi_2 + \cos \theta \Phi_3)^T.$$

Now, if we take as above in (4.6) a flux  $\Phi$  of the form

$$\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) = \varphi(\mathbb{F} \cdot \mathbf{n}; \mathbf{U}_i, \mathbf{U}_j),$$

since

$$\Phi(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{e}_1) = \varphi(\mathbf{f}; \tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j),$$

rotational invariance is equivalent to

$$\varphi(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{e}_1) = R\varphi(\mathbf{f}; \tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j). \quad (4.8)$$

If we write the expressions of the usual three-point fluxes, we can see that the last equality is satisfied by many usual schemes. Indeed, from (2.14a),

$$\mathbb{F} \cdot \mathbf{n}(\mathbf{U}) = R\mathbf{f}(\tilde{\mathbf{U}}). \quad (4.9a)$$

We get by differentiation

$$\mathbf{A}(\mathbf{U}, \mathbf{n}) = (\mathbb{F} \cdot \mathbf{n})'(\mathbf{U}) = R\mathbf{f}'(\tilde{\mathbf{U}})R^{-1} = R\mathbf{A}(\tilde{\mathbf{U}})R^{-1}, \quad (4.9b)$$

so that for any  $\mathbf{V} = R\tilde{\mathbf{V}} \in \mathbb{R}^4$

$$\mathbf{A}(\mathbf{U}, \mathbf{n})\mathbf{V} = R\mathbf{A}(\tilde{\mathbf{U}})\tilde{\mathbf{V}}. \quad (4.9c)$$

We shall consider Roe's matrix in Sect. 4.3.1 and see that

$$\mathbf{A}_n(\mathbf{U}_L, \mathbf{U}_R)\mathbf{V} = R\mathbf{A}_n(\tilde{\mathbf{U}}_L, \tilde{\mathbf{U}}_R)\tilde{\mathbf{V}}. \quad (4.9d)$$

We shall also obtain a similar result for Osher's and Steger and Warming's fluxes. Substituting the identities (4.9) in the formula of a difference scheme proves indeed that (4.8) holds for most numerical fluxes. Hence, the flux can be equivalently defined by

$$\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) = R\varphi(\mathbf{f}; \tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j),$$

which yields much simpler computations (see, for instance, Mulder et al. [878], Selmin and Quartapelle [1029]). Note that the rotational invariance of the Euler equations is also used in more sophisticated schemes derived in order to prevent grid alignment problems. Indeed, in formulas (4.2) and (4.3), the interface normal is chosen as the direction for wave propagation. If the waves are not aligned with the grid, they may be misrepresented (see van

Leer [1157], Roe [981]; also Coirier and Powell [324], Noelle [904]). Many attempts have been made at adapting a computational grid or minimizing grid orientation effects (Bourgeat and Koebbe [185]), for instance, by taking into account the directions in which information is propagated and using a “rotated Riemann solver” (Davis [391], Levy, Powell and van Leer [787], Rumsey et al. [996], LeVeque and Walder [784], Fey [473, 474]) or including “tangential wave propagation” as well as normal wave propagation (LeVeque [773]). We shall proceed with the derivation of a “truly multidimensional” solver in Sect. 4.3.2.

### 4.1.3 Examples

Let us now give some examples of “finite-volume” methods.

*Example 4.1. “Cell-center” scheme.* Starting from a triangulation  $\mathcal{T}_h = \cup T_i$  of  $\mathcal{O} \subset \mathbb{R}^2$  (which we assume is regular enough), one defines a control cell as an element of  $\mathcal{T}_h$ , i.e., a triangle  $\Omega_i = T_i$  and the center of the cell is the centroid  $\mathbf{g}_i$  of the triangle  $T_i$  (see Fig. 4.1). Approximating the solution of (3.2) by a function that is piecewise constant on each triangle,  $\mathbf{w}_h(t)|_{T_i} = \mathbf{w}_i(t)$ , and integrating the equation over  $T_i$  gives

$$|T_i| \frac{\partial \mathbf{w}_i}{\partial t} + \int_{\partial T_i} (\mathbf{f}(\mathbf{w}_i) n_{ix} + \mathbf{g}(\mathbf{w}_i) n_{iy}) d\sigma = \mathbf{0}, \quad \forall i,$$

and the integral on the boundary is then discretized via a numerical flux function as we have already explained. One often interprets this “finite-volume method” as a “finite-element method,” which is sometimes called “cell centered”; indeed, since

$$\mathbf{u}_i \cong \frac{1}{|T_i|} \int_{T_i} \mathbf{u}(\mathbf{x}) d\mathbf{x} = \mathbf{u}(\mathbf{g}_i, t) + O(h^2),$$

one can say that the values  $\mathbf{u}_i$  are associated to the centroid  $\mathbf{g}_i$ . If we introduce the space  $\mathbf{W}_h = \{\mathbf{w}_h \in \mathbf{L}^2(\mathcal{O})^p; \forall T_i \in \mathcal{T}_h, \exists \mathbf{w}_i \in \mathbb{R}, \mathbf{w}_{h|T_i} = \mathbf{w}_i\}$ , we can associate a variational problem: find  $\mathbf{w}_h : (0, T) \rightarrow \mathbf{W}_h$ ,  $\mathbf{w}_h(t)|_{T_i} = \mathbf{w}_i(t)$ , such that

$$\int_{\Omega} \left\{ \frac{\partial \mathbf{w}_h}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{w}_h) + \frac{\partial}{\partial y} \mathbf{g}(\mathbf{w}_h) \right\} \cdot \boldsymbol{\varphi}_h dx dy = 0, \quad \forall \boldsymbol{\varphi}_h \in \mathbf{W}_h.$$

However, the above formula is not correct since the functions of  $\mathbf{W}_h$  are discontinuous. Instead, writing for any  $\boldsymbol{\varphi}_h \in \mathbf{W}_h$

$$\boldsymbol{\varphi}_h = \sum \boldsymbol{\varphi}_h(g_i) 1_{T_i},$$

we obtain

$$\int_{T_i} \left\{ \frac{\partial \mathbf{w}_i(t)}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{w}_i(t)) + \frac{\partial}{\partial y} \mathbf{g}(\mathbf{w}_i(t)) \right\} dx dy = \mathbf{0}, \quad \forall i,$$

and using Green's formula we find the finite-volume method (for the relation with the streamline diffusion finite-element method, see Hansbo [584]).

One can also consider unstructured quadrilaterals or a “triangulation” made of mixed triangular–quadrilateral elements.  $\square$

*Remark 4.3.* We will not consider discontinuous Galerkin methods (which follow the same idea, allowing to take higher-order polynomials, instead of constants in each cell, with possible discontinuities at the interface). They are receiving much interest because of the greater precision they offer (it is however difficult to prove general stability results). We refer to [321, 447, 613, 1057, 1084].  $\square$

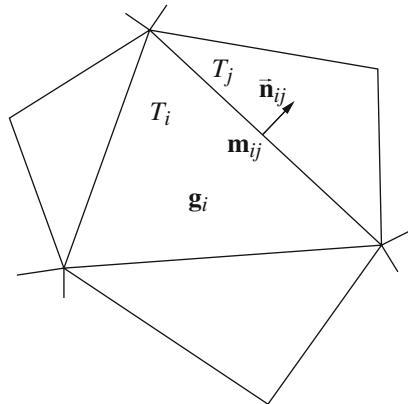
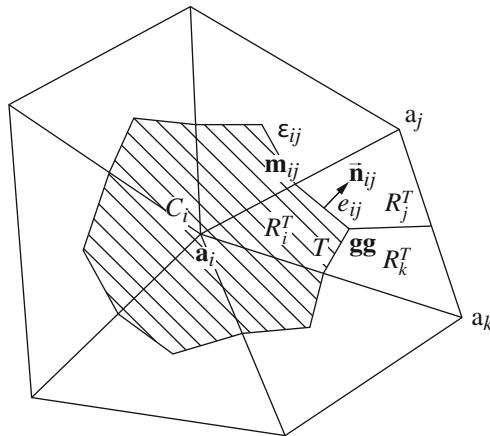


Fig. 4.1 “Cell center”

*Example 4.2. “Cell vertex” scheme.* In many situations, the quantities are defined at the vertices  $\mathbf{a}_i$  of the triangulation rather than at the barycenter  $\mathbf{g}_i$  of each triangle. Thus, starting from a triangulation  $\mathcal{T}_h$  of  $\mathcal{O} \subset \mathbb{R}^2$  (which again is regular enough), one defines the centers of the control cells as the vertices  $\mathbf{a}_i$  of the triangles (the nodes of the triangulation), and the control cell  $\Omega_i = C_i$  associated to  $\mathbf{a}_i$  as the “dual cell” of the node  $\mathbf{a}_i$ : the polygonal boundary of  $C_i$  is obtained by joining, for each triangle  $T$  having the vertex  $\mathbf{a}_i$  in common, the midpoint of each triangle edge issued from  $\mathbf{a}_i$  to the triangle barycenter  $\mathbf{g}$ . The boundary is thus composed of medians (Vijayasundaram [1169], Stoufflet [1075], Fezoui [476], Angrand and Lafon [54]). Note that  $\mathbf{a}_i$  is not necessarily the centroid of  $C_i$ . The cell  $C_i$  consists of the union of quadrilateral regions of type  $R_i : C_i = \cup R_i^T$  (where  $R_i^T$  is the

region  $R_i$  belonging to the triangle  $T$ ) for all triangles  $T$  having the vertex  $\mathbf{a}_i$  in common (see Fig. 4.2). The boundary between two neighboring cells  $C_i$  and  $C_j$  consists of segments  $e_{ij}$  and  $\varepsilon_{ij}$  crossing at triangle edge midpoints  $\mathbf{m}_{ij}$ . When integrating the system over  $C_i$ , on  $R_i$  Green's formula gives, in particular, a flux across the segment  $e_{ij}$ .

Some variants are found where the boundary of  $C_i$  is obtained by joining the barycenters of two neighboring triangles (in Perthame and Qiu [946], for instance). We can also consider the perpendicular bisectors of the edges of the triangles, which gives then for  $C_i$  the Voronoï cell associated to the Delaunay triangulation (see Mavriplis [857], W.K. Anderson [41]).  $\square$



**Fig. 4.2** “Cell vertex”

*Remark 4.4.* When starting from quadrilateral elements, with vertices characterized by  $(i, j)$ , the dual cell is obtained by taking the centers of the neighboring cells  $(i \pm \frac{1}{2}, j \pm \frac{1}{2})$  for the vertices of the dual cell, and the boundary of the dual cell is obtained by joining them to the middle  $(i \pm \frac{1}{2}, j)$  and  $(i, j \pm \frac{1}{2})$  of the edges (see Fig. 4.3). It is then easy to compute by interpolation the gradients at the centers  $(i, j)$  of the dual meshes.

Note also that when approximating a system, one sometimes uses the two meshes, the dual mesh is sometimes used for approximating one of the components of the unknown vector (typically for compressible Navier–Stokes equations, the pressure can be evaluated at the nodes  $i, j$ , while the velocity is evaluated at the centers) (staggered approach; see, for instance, [608]); it occurs also in codes for two-phase flows in reservoir simulation [534]). In structured meshes, a shifted mesh may be used for the computation of gradients (see Peyret and Taylor [951], Koren [707]).  $\square$

*Example 4.3. Weighted Finite Volume.* In this example, we start again from a triangulation of  $\mathbb{R}^2$ ,  $\mathcal{T}_h = \cup T$ , composed of triangles, and we introduce the finite element space of type (1)  $\mathbf{X}_h$  consisting of piecewise linear ( $\mathbf{P}_1$ ) continuous functions characterized by their values at the nodes  $\mathbf{a}_i, i \in I$ . A basis of  $\mathbf{X}_h$  consists of the shape functions  $\psi_j \in \mathbf{X}_h$  such that (using the Kronecker symbol)

$$\psi_j(a_i) = \delta_i^j, \quad i, j \in I$$

(for simplicity, we consider the scalar case). The support of  $\psi_i$ , which we denote by  $\text{supp}(\psi_i)$ , consists of the union of the triangles  $T$  having the vertex  $\mathbf{a}_i$  in common. The  $\mathbf{L}^2$ -projection  $\pi g$  of a function  $g \in \mathbf{L}^2(\mathcal{O})$  on  $\mathbf{X}_h$  is defined by

$$\int \pi g \varphi d\mathbf{x} = \int g \varphi d\mathbf{x}, \quad \forall \varphi \in \mathbf{X}_h,$$

and using a  $\mathbf{P}_1$  quadrature formula (which corresponds to the principle of “mass-lumping,” i.e., to diagonalize the “mass matrix” ( $\int \varphi_i \varphi_j d\mathbf{x}$ )),

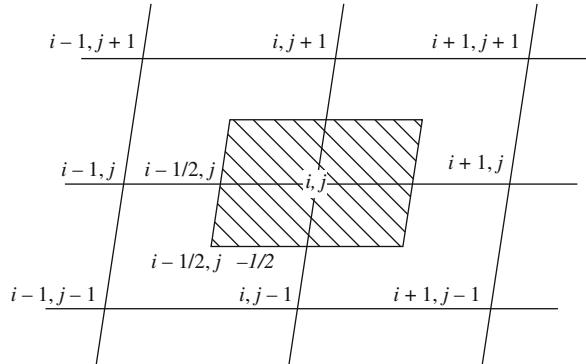


Fig. 4.3 “Dual cell” in the structured case

$$\int f d\mathbf{x} \cong \int \left( \sum_{i \in I} f(a_i) \psi_i \right) d\mathbf{x} = \sum_{i \in I} f(a_i) \int \psi_i d\mathbf{x}$$

gives “the value of  $g$ ” at point  $a_i$ ,

$$\pi g(a_i) = \frac{\left( \int g \psi_i d\mathbf{x} \right)}{\left( \int \psi_i d\mathbf{x} \right)}, \quad i \in I.$$

We associate a fictitious cell  $\Omega_i$  to a node  $\mathbf{a}_i$  as follows: we define  $\Omega_i$  as the volume (area) of  $\text{supp}(\psi_i)$  weighted by  $\psi_i$ , i.e., for any  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,

$$\int_{\Omega_i} g d\mathbf{x} = \int_{\text{supp}(\psi_i)} g \psi_i d\mathbf{x}.$$

Let us compute the “area” of  $\Omega_i$ ,

$$|\Omega_i| = \int_{\mathbb{R}^2} \psi_i d\mathbf{x}.$$

Since

$$\sum_i \psi_i(\mathbf{x}) = 1, \quad \forall \mathbf{x} \in \Omega,$$

we have

$$\int_{T \cap \Omega_i} d\mathbf{x} = \frac{1}{3} |T|,$$

and

$$|\Omega_i| = \frac{1}{3} \sum_{T/a_i \in T} |T|$$

is exactly the area of the dual cell  $C_i$  of Example 4.2. Formula (4.2) becomes

$$|\Omega_i|(u_i^{n+1} - u_i^n) - \Delta t \int_{\mathbb{R}^2} \mathbb{F}(u) \cdot \operatorname{grad} \psi_i d\mathbf{x} = 0.$$

Let us give an example of a situation where these weighted finite volumes are used. In problems from petroleum reservoir simulation, the continuous flux has a prescribed direction  $\mathbf{v}$ ,

$$\mathbb{F}(u) = \mathbf{v} f(u), \quad f : \mathbb{R} \rightarrow \mathbb{R},$$

with  $\mathbf{v} = \operatorname{grad} p \in \mathbb{R}^2$ . Let us replace  $p$  by its interpolate,

$$p \cong \pi p = \sum_j p(a_j) \psi_j.$$

Since  $\sum_j \operatorname{grad} \psi_j(x) = 0$ , we can write in cell  $i$

$$\operatorname{grad} p \cong \sum_j (p(a_j) - p(a_i)) \operatorname{grad} \psi_j,$$

and we obtain the scheme

$$|\Omega_i|(u_i^{n+1} - u_i^n) - \Delta t \sum_j (p(a_j) - p(a_i)) f(u_{ij}^n) \int_{\mathbb{R}^2} \operatorname{grad} \psi_i \cdot \operatorname{grad} \psi_j d\mathbf{x} = 0,$$

where the exchange term  $f(u_{ij}^n)$  can be defined in an upwind manner

$$f(u_{ij}^n) = \begin{cases} f(u_j^n) & \text{if } p(a_i) = p_i > p_j = p(a_j), \\ f(u_i^n) & \text{if } p(a_i) = p_i < p_j = p(a_j). \end{cases}$$

For details we refer to Eymard and Gallouët [464]. □

*Example 4.4. Rectangular mesh.* If we start from a rectangular mesh  $\Omega = \cup \Omega_{j,k}$ , where the rectangle  $\Omega_{j,k} = (x_{j-1/2}, x_{j+1/2}) \times (y_{k-1/2}, y_{k+1/2})$  has center  $\mathbf{a}_{jk} = (x_j, y_k)$ , we have a Cartesian grid and can find the classical finite difference schemes. Setting

$$\Delta x_j = x_{j+1/2} - x_{j-1/2}, \quad \Delta y_k = y_{k+1/2} - y_{k-1/2},$$

we have  $|\Omega_{j,k}| = \Delta x_j \Delta y_k$ , and formula (4.3) becomes

$$\mathbf{u}_{j,k}^{n+1} = \mathbf{u}_{j,k}^n - \frac{\Delta t}{\Delta x_j} (\mathbf{F}_{j+1/2,k}^n - \mathbf{F}_{j-1/2,k}^n) - \frac{\Delta t}{\Delta y_k} (\mathbf{G}_{j,k+1/2}^n - \mathbf{G}_{j,k-1/2}^n),$$

where  $\mathbf{F}_{j+1/2,k}^n$  denotes the flux at the edge  $x = x_{j+1/2}$  (with length  $\Delta y_k$ ) between  $\Omega_{j,k}$  and  $\Omega_{j+1,k}$  and so on. Defining

$$\mathbf{u}_{j,k}^0 = \frac{1}{\Delta x_j \Delta y_k} \int_{\Omega_{j,k}} \mathbf{u}_0(\mathbf{x}) d\mathbf{x}, \quad \lambda_{x_j} = \frac{\Delta t}{\Delta x_j}, \quad \lambda_{y_k} = \frac{\Delta t}{\Delta y_k},$$

we get exactly scheme (3.3) with variable mesh size,

$$\mathbf{u}_{j,k}^{n+1} = \mathbf{u}_{j,k}^n - \lambda_{x_j} (\mathbf{F}_{j+1/2,k}^n - \mathbf{F}_{j-1/2,k}^n) - \lambda_{y_k} (\mathbf{G}_{j,k+1/2}^n - \mathbf{G}_{j,k-1/2}^n).$$

We can take for  $\mathbf{F}$  and  $\mathbf{G}$  the numerical flux associated to any three-point scheme.  $\square$

## 4.2 General Results

Most stability and convergence results obtained in the one-dimensional case extend to a Cartesian grid (see Crandall and Majda [372, 373], Sanders [1005], Osher and Sanders [921]). Hence, we shall mainly consider in this section the case of an unstructured mesh and particularly the case of cell-centered triangles (Example 4.1). Let us first have a closer look at the order of accuracy.

### 4.2.1 Order

Let  $u$  be a smooth (scalar) solution of (3.6a)–(3.6b) and  $\Phi(u, v, n)$  a  $C^1$  (scalar) function of its arguments  $u$  and  $v$ ; the truncation error is obtained by substituting  $u$  in the formula (4.3) (divided by  $|K|$ ). Indeed, setting

$$u_K(t) = \frac{1}{|K|} \int_K u(\mathbf{x}, t) d\mathbf{x},$$

the truncation error in  $K$  is defined by  $\Delta t \varepsilon_K(t)$  where

$$\begin{aligned}\varepsilon_K(t) &= u_K(t + \Delta t) - u_K(t) \\ &+ \frac{\Delta t}{|K|} \sum_{e \subset \partial K, e = K \cap K'} |e| \Phi(u_K(t), u_{K'}(t), \mathbf{n}_e).\end{aligned}\quad (4.10)$$

We can write

$$\begin{aligned}u_K(t + \Delta t) - u_K(t) &= \frac{1}{|K|} \int_K (u(\mathbf{x}(t + \Delta t)) - u(\mathbf{x}, t)) d\mathbf{x} \\ &= \frac{1}{|K|} \int_K \int_0^{\Delta t} \frac{\partial u}{\partial t}(\mathbf{x}, t + s) d\mathbf{x} ds.\end{aligned}$$

Now, taking into account the fact that  $u$  is solution of (3.6a), we can replace  $\frac{\partial u}{\partial t}$  by  $-(\frac{\partial}{\partial x} f(u) + \frac{\partial}{\partial y} g(u))$  and use the divergence theorem

$$\begin{aligned}u_K(t + \Delta t) - u_K(t) &= -\frac{1}{|K|} \int_K \int_0^{\Delta t} \operatorname{div}(\mathbb{F}(u)) d\mathbf{x} ds \\ &= -\frac{1}{|K|} \int_0^{\Delta t} \int_{\partial K} \mathbb{F}(u) \cdot \mathbf{n} d\sigma ds,\end{aligned}$$

which yields

$$\begin{aligned}\varepsilon_K(t) &= -\frac{1}{|K|} \sum_e \left\{ \int_0^{\Delta t} \int_{\partial K} \mathbb{F}(u(\sigma, t + s)) \cdot \mathbf{n}_e d\sigma ds \right. \\ &\quad \left. + \Delta t |e| \Phi(u_K(t), u_{K'}(t), \mathbf{n}_e) \right\}.\end{aligned}$$

Using quadrature formulas, we write

$$\begin{aligned}\int_0^{\Delta t} \int_{\partial K} \mathbb{F}(u(\sigma, t + s)) \cdot \mathbf{n}_e d\sigma ds &= \Delta t \{ \mathbb{F}(u(\sigma, t)) \cdot \mathbf{n}_e + \tilde{\varepsilon}(\sigma, t) \}, \quad \forall \sigma \in e, \quad (4.11) \\ \int_e \int_{\partial K} \mathbb{F}(u(\sigma, t)) \cdot \mathbf{n}_e d\sigma &= |e| \{ \mathbb{F}(u(\mathbf{m}_e, t)) \cdot \mathbf{n}_e + O(h^2) \}, \\ u_K(t) &= \frac{1}{|K|} \int_K u(\mathbf{x}, t) d\mathbf{x} = u(\mathbf{g}, t) + O(h^2),\end{aligned}$$

where  $\mathbf{g}$  is the centroid of  $K$ ,  $\mathbf{m}_e$  is the midpoint of the edge  $e$  of  $K$ , and  $h$  is the diameter of  $K$ . In fact, some care must be taken in the evaluation of the  $O(\Delta t)$  term  $\tilde{\varepsilon}$  in (4.11). Since the sum comes from the integral  $\int_K \int_0^{\Delta t} \operatorname{div}(\mathbb{F}(u)) d\mathbf{x} ds$ , it is in fact  $O(\Delta t)O(h^2)$ . Indeed, by Taylor series expansion, we can write

$$\int_e \tilde{\varepsilon}(\sigma, t) d\sigma = |e| \left\{ \frac{\Delta t}{2} (\mathbb{F}'(u(\mathbf{g}, t)) \cdot \mathbf{n}_e) \frac{\partial u}{\partial t}(\mathbf{g}, t) + O(h) + O(\Delta t) \right\}.$$

Since  $\sum_{e \in K} n_e = 0$ , summing on all the edges  $e$  of  $K$ , we get

$$\sum_{e \subset K} \int_e \tilde{\varepsilon}(\sigma, t) d\sigma = |e| \Delta t \{O(h) + O(\Delta t)\}.$$

Now, setting

$$u_e = u(\mathbf{m}_e, t),$$

we have

$$\begin{aligned} \Phi(u_K(t), u_{K'}(t), \mathbf{n}_e) &= \Phi(u_e, u_e, \mathbf{n}_e) + \frac{\partial \Phi}{\partial u}(u_e, u_e, \mathbf{n}_e)(u_K(t) - u_e) \\ &\quad + \frac{\partial \Phi}{\partial v}(u_e, u_e, \mathbf{n}_e)(u_{K'}(t) - u_e) + O(h^2) \end{aligned}$$

and

$$u_K(t) - u_e = u_k(t) - u(\mathbf{m}_e, t) = \text{grad } u \cdot (\mathbf{g} - \mathbf{m}_e) + O(h^2).$$

Due to the consistency of  $\Phi$ , (4.10) becomes

$$\begin{aligned} \varepsilon_K(t) &= \frac{\Delta t}{|K|} \sum_{e \subset \partial K} |e| \left\{ \frac{\partial \Phi}{\partial u}(u_e, u_e, \mathbf{n}_e)(u_K(t) - u_e) \right. \\ &\quad \left. + \frac{\partial \Phi}{\partial v}(u_e, u_e, \mathbf{n}_e)(u_{K'}(t) - u_e) + O(h^2) + O(\Delta t) \right\}, \end{aligned}$$

and hence

$$\left\{ \begin{aligned} \varepsilon_K(t) &= \frac{\Delta t}{|K|} \text{grad } u \cdot \sum_{e \subset \partial K} |e| \left\{ \frac{\partial \Phi}{\partial u}(u_e, u_e, \mathbf{n}_e)(\mathbf{g} - \mathbf{m}_e) \right. \\ &\quad \left. + \frac{\partial \Phi}{\partial v}(u_e, u_e, \mathbf{n}_e)(\mathbf{g}' - \mathbf{m}_e) + O(h^2) + \Delta t O(h) \right\}. \end{aligned} \right. \quad (4.12)$$

*Lemma 4.1*

Assume that we are in the situation of Example 4.1, with, moreover, a uniform triangulation by equilateral triangles  $T$  with side  $h$ . If the numerical flux  $\Phi$  satisfies

$$\sum_{e \subset \partial T} \left\{ \frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}_e) - \frac{\partial \Phi}{\partial v}(u, u, \mathbf{n}_e) \right\} \mathbf{n}_e = 0, \quad \forall u \in \mathbb{R},$$

or

$$\sum_{e \subset \partial T} \left\{ \frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}_e) + \frac{\partial \Phi}{\partial u}(u, u, -\mathbf{n}_e) \right\} \mathbf{n}_e = 0, \quad \forall u \in \mathbb{R},$$

the scheme (4.3) is first-order accurate.

*Proof.* Starting from formula (4.12) now with  $K = T$ , we see that some terms simplify since

$$\mathbf{m}_e - g = |\mathbf{m}_e - \mathbf{g}| \mathbf{n}_e = -(\mathbf{m}_e - \mathbf{g}'), \quad |K| = |T| = 3|\mathbf{m}_e - \mathbf{g}| \frac{|e|}{2}.$$

Thus

$$\begin{aligned} \varepsilon_T = \frac{2\Delta t}{3} \operatorname{grad} u \cdot \sum_{e \subset \partial T} \left\{ \left( -\frac{\partial \Phi}{\partial u}(u_e, u_e, \mathbf{n}_e) + \frac{\partial \Phi}{\partial v}(u_e, u_e, \mathbf{n}_e) \right) \mathbf{n}_e \right. \\ \left. + O(\Delta t) + O(h) \right\}. \end{aligned}$$

We can write the same formula with the fluxes evaluated at the centroid: setting

$$u = u(\mathbf{g}, t),$$

we have

$$\begin{aligned} \varepsilon_K = \frac{2\Delta t}{3} \operatorname{grad} u \cdot \sum_{e \subset \partial T} \left\{ \left( -\frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}_e) + \frac{\partial \Phi}{\partial v}(u, u, \mathbf{n}_e) \right) \mathbf{n}_e \right. \\ \left. + O(\Delta t) + O(h) \right\}. \end{aligned}$$

Differentiating the conservation relation (4.4), we have

$$\frac{\partial \Phi}{\partial v}(u, u, \mathbf{n}) - \frac{\partial \Phi}{\partial u}(u, u, -\mathbf{n}) = 0$$

and

$$\frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}) - \frac{\partial \Phi}{\partial v}(u, u, \mathbf{n}) = \frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}) + \frac{\partial \Phi}{\partial u}(u, u, -\mathbf{n}),$$

so that we can also write  $\varepsilon_K(t)$  in the following form:

$$\begin{aligned} \varepsilon_K(t) = -\frac{2\Delta t}{3} \operatorname{grad} u \cdot \sum_{e \subset \partial T} \left\{ \left( \frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}_e) + \frac{\partial \Phi}{\partial u}(u, u, -\mathbf{n}_e) \right) \mathbf{n}_e \right. \\ \left. + O(\Delta t) + O(h) \right\}. \end{aligned}$$

Assuming that

$$\sum \left\{ \frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}_e) - \frac{\partial \Phi}{\partial v}(u, u, \mathbf{n}_e) \right\} \mathbf{n}_e = 0$$

or that

$$\sum \left\{ \frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}_e) - \frac{\partial \Phi}{\partial v}(u, u, -\mathbf{n}_e) \right\} \mathbf{n}_e = 0,$$

the first term in  $\varepsilon_K(t)$  vanishes, and we obtain a first-order accurate method

$$\varepsilon_K(t) = O(\Delta t) + O(h),$$

i.e.,  $\varepsilon_K(t) = O(h)$  under a CFL condition.  $\square$

The conditions of the lemma are satisfied by the usual schemes. Indeed, in the formula for the numerical flux  $\Phi$ , the terms that do not depend on  $F$  will be taken out of the sum “ $\Sigma$ ” so that we can use the identity  $\Sigma_e \mathbf{n}_e = 0$ , while those that do depend on  $F$  (and thus on  $\mathbf{n}_e$ ) are frequently symmetric in  $u$  and  $v$  or linear in  $\mathbf{n}$ . For instance, the Lax–Friedrichs scheme gives

$$\Phi(u, v, \mathbf{n}) = \mathbb{F}(u) \cdot \mathbf{n} + \mathbb{F}(v) \cdot \mathbf{n} - \frac{(v - u)}{2\lambda},$$

and hence

$$\begin{aligned}\frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}) &= \mathbb{F}'(u) \cdot \mathbf{n} + \frac{1}{2\lambda}, \\ \frac{\partial \Phi}{\partial v}(u, u, \mathbf{n}) &= \mathbb{F}'(u) \cdot \mathbf{n} - \frac{1}{2\lambda},\end{aligned}$$

and

$$\sum \left( \frac{\partial \Phi}{\partial u}(u(\mathbf{m}, t), u(\mathbf{m}, t), \mathbf{n}_e) - \frac{\partial \Phi}{\partial v}(u(\mathbf{m}, t), u(\mathbf{m}, t), \mathbf{n}_e) \right) \mathbf{n}_e = 0.$$

In other cases, such as the upwind scheme (Godunov’s),

$$\Phi(u, v, \mathbf{n}) = \begin{cases} \mathbb{F}(u) \cdot \mathbf{n} & \text{if } \mathbb{F}'(\cdot) \cdot \mathbf{n} > 0, \\ \mathbb{F}(v) \cdot \mathbf{n} & \text{if } \mathbb{F}'(\cdot) \cdot \mathbf{n} < 0. \end{cases}$$

We can see that away from “sonic” points,

$$\frac{\partial \Phi}{\partial u}(u, u, \mathbf{n}_e) + \frac{\partial \Phi}{\partial u}(u, u, -\mathbf{n}_e) = 0.$$

When the grid is not uniform, the scheme is not first-order accurate. Note that even in the one-dimensional case, when the length  $\Delta x_i$  of the mesh is not uniform, it is easy to see that conservation and consistency do not imply that the truncation error tends to zero, unless the ratio  $(\sup \Delta x_i) / (\inf \Delta x_i) \rightarrow 1$  as  $\sup \Delta x_i \rightarrow 0$  (and a second-order scheme gives a first-order scheme on an irregular grid) (see Pike [952]; see also Jeng and Chen [652], Turkel [1137, 1138], Wendroff and White [1186], Durlofsky et al. [452] for numerical experiments).

In any case, the conservation property implies somehow a cancellation of the errors in adjacent cells, and we shall prove the analog of the Lax–Wendroff theorem. It says that in the scalar case ( $p = 1$ ), when the scheme “converges” in some sense that will be specified below, the limit is a weak solution to the conservation law. Remember that such a solution exists (see Theorem 5.4 in the Chap. I and G.R., Chapter 2, Section 5 [539]).

### 4.2.2 The Lax–Wendroff Theorem

For the proof of the analog of the Lax–Wendroff theorem in the case of an unstructured mesh, we need the following classical approximation result:

*Lemma 4.2*

Let  $\mathcal{T}_h = \cup K$  be a countable family of triangulations of a bounded set  $\mathcal{O} \subset \mathbb{R}^2$ , and define for  $u \in \mathbf{L}^1(\mathcal{O})$

$$u_K = \frac{1}{|K|} \int_K u(\mathbf{x}) d\mathbf{x}$$

and

$$\pi_h u = \sum_k u_K 1_K.$$

Then  $\pi_h u \rightarrow u$  in  $\mathbf{L}^1(\mathcal{O})$  as  $h \rightarrow 0$ , where  $h = \sup |K|, K \in \mathcal{T}_h$ .

*Proof.* Let us first consider the case of a continuous function  $\varphi \in C^0(\mathcal{O})$ . Then

$$\pi_h \varphi \rightarrow \varphi \quad \text{a.e. as } h \rightarrow 0$$

(more precisely, outside the set of measure zero composed of the union of the boundaries of the sets  $K \in \mathcal{T}_h$  for the countable family  $\mathcal{T}_h$ ), and

$$|\pi_h \varphi| \leq \|\varphi\|_{\mathbf{L}^\infty(\mathcal{O})}.$$

Thus  $\pi_h \varphi \rightarrow \varphi$  in  $\mathbf{L}^1(\mathcal{O})$ .

Now, for  $u \in \mathbf{L}^1(\mathcal{O})$ ,

$$\begin{aligned} \|\pi_h u\|_{\mathbf{L}^1(\mathcal{O})} &= \int_{\mathcal{O}} |\pi_h u| d\mathbf{x} = \int_{\mathcal{O}} \left| \sum_K u_K 1_K \right| d\mathbf{x} \\ &\leq \sum_K |u_K| |K| = \sum_K \int_K |u(\mathbf{x})| d\mathbf{x} = \|u\|_{\mathbf{L}^1(\mathcal{O})}, \end{aligned}$$

and for any  $\varphi \in C^0(\mathcal{O})$ ,

$$\|\pi_h u - u\|_{\mathbf{L}^1(\mathcal{O})} \leq \|\pi_h u - \pi_h \varphi\|_{\mathbf{L}^1(\mathcal{O})} + \|\pi_h \varphi - \varphi\|_{\mathbf{L}^1(\mathcal{O})} + \|u - \varphi\|_{\mathbf{L}^1(\mathcal{O})}.$$

Given  $\varepsilon$ , we choose  $\varphi \in C^0(\mathcal{O})$  such that  $\|u - \varphi\|_{\mathbf{L}^1(\mathcal{O})} \leq \varepsilon$ . Then for  $h$  small enough, we have  $\|\pi_h \varphi - \varphi\|_{\mathbf{L}^1(\mathcal{O})} \leq \varepsilon$  and the result follows.  $\square$

Let us now introduce basis functions on each triangle, which will enable us to reconstruct conveniently an affine function from its values at the vertices.

*Lemma 4.3*

Let us consider a triangle  $T$  with vertices  $\mathbf{a}_i, 1 \leq i \leq 3$ . Let  $e_i$  denote the edge of  $T$  opposite to  $\mathbf{a}_i$ , with outside unit normal  $\mathbf{n}_i$ . There exists an affine function  $\mathbf{p}_i : T \rightarrow \mathbb{R}^2$ , for  $i = 1, 2, 3$ , such that

$$\mathbf{p}_i(\mathbf{x}) \cdot \mathbf{n}_i = 1, \quad \forall \mathbf{x} \in e_i; \quad \mathbf{p}_i(\mathbf{x}) \cdot \mathbf{n}_j = 0, \quad \forall \mathbf{x} \in e_j, \quad j \neq i. \quad (4.13a)$$

Moreover, the  $\mathbf{p}_i$  satisfy

$$\sum_{k=1}^3 \mathbf{n}_k \mathbf{p}_k(\mathbf{x})^T = \mathbf{I}, \quad \forall \mathbf{x} \in T \quad (4.13b)$$

and

$$\operatorname{div} \mathbf{p}_i = \frac{|e_i|}{|T|}, \quad i = 1, 2, 3. \quad (4.13c)$$

*Proof.* It is easy to construct the  $\mathbf{p}_i$  satisfying (4.13a) on a reference triangle with vertices  $\mathbf{a}_1 = \mathbf{x}_1 = (0, 0)^T, \mathbf{a}_2 = \mathbf{x}_2 = (0, 1)^T, \mathbf{a}_3 = \mathbf{x}_3 = (1, 0)^T$ . We take

$$\mathbf{p}_1(\mathbf{x}) = \sqrt{2}\mathbf{x}, \quad \mathbf{p}_2(\mathbf{x}) = \mathbf{x} - (0, 1)^T, \quad \mathbf{p}_3(\mathbf{x}) = \mathbf{x} - (1, 0)^T.$$

On any triangle  $T$ , we get for the function associated to the edge  $e_i$  opposite to the vertex  $\mathbf{a}_i$

$$\mathbf{p}_i(\mathbf{x}) = (\mathbf{x} - \mathbf{a}_i) \frac{|e_i|}{2|T|},$$

where the constants are chosen to satisfy (4.13c).

Let us check the property (4.13b); it means that for all  $\mathbf{v} \in \mathbb{R}^2$

$$\sum_k (\mathbf{p}_k(\mathbf{x}) \cdot \mathbf{v}) \mathbf{n}_k = \mathbf{v}.$$

Let us introduce the notation

$$\mathbf{B}(\mathbf{x}) = \sum_k \mathbf{n}_k \mathbf{p}_k(\mathbf{x})^T.$$

For any  $\mathbf{x} \in T$ ,  $\mathbf{B}(\mathbf{x})$  is a  $2 \times 2$  matrix, each component of which is an affine function of  $\mathbf{x}$ . On any edge  $e_j$ , we have

$$\mathbf{B}(\mathbf{x}) \mathbf{n}_j = \sum_k (\mathbf{p}_k(\mathbf{x}) \cdot \mathbf{n}_j) \mathbf{n}_k = \sum_k \delta_{j,k} \mathbf{n}_k = \mathbf{n}_j, \quad \mathbf{x} \in e_j.$$

Thus, at the vertex  $\mathbf{a}_k = e_i \cap e_j$  ( $i \neq k, j \neq k$ ),

$$\mathbf{B}(\mathbf{a}_k) \mathbf{n}_i = \mathbf{n}_i \quad \text{and} \quad \mathbf{B}(\mathbf{a}_k) \mathbf{n}_j = \mathbf{n}_j,$$

which proves that  $\mathbf{B}(\mathbf{a}_k) = \mathbf{I}$  since  $(\mathbf{n}_i, \mathbf{n}_j)$  form a basis of  $\mathbb{R}^2$ , for  $k = 1, 2, 3$ . Since the coefficients are affine functions, we obtain

$$(\mathbf{B}(\mathbf{a}_i) = \mathbf{I} \text{ and } \mathbf{B}(\mathbf{a}_j) = \mathbf{I}) \implies \mathbf{B}(\mathbf{x}) = \mathbf{I} \quad \text{on } e_k,$$

which yields

$$\mathbf{B}(\mathbf{x}) = \mathbf{I} \quad \text{on } e_k \implies \mathbf{B}(\mathbf{x}) = \mathbf{I} \quad \text{on } T.$$

Note that the above property is directly linked to the fact that

$$\sum_{i=1}^3 |e_i| \mathbf{n}_i = \mathbf{0}, \quad (4.14)$$

which we shall use again.  $\square$

Let us define the piecewise constant function associated to the finite-volume scheme (4.3),

$$v_\Delta(\mathbf{x}, t) = u_i^n, \quad \mathbf{x} \in \Omega_i, \quad t \in [n\Delta t, (n+1)\Delta t[, \quad \Delta = (h, \Delta t), \quad (4.15)$$

and set

$$u_i^0 = \frac{1}{|\Omega_i|} \int_{\Omega_i} u_0(\mathbf{x}) d\mathbf{x}. \quad (4.16)$$

We assume for simplicity that we are in the situation of Example 4.1, i.e., we are given a triangulation  $\mathcal{T}_h = \cup T_i$  of  $\Omega \subset \mathbb{R}^2$ , and  $\Omega_i = T_i$ . We introduce on the one hand regularity assumptions on the mesh: we assume that there exist positive constants  $a, b$  such that

$$ah^2 \leq |T_i| \leq bh^2, \quad \forall T_i \in \mathcal{T}_h \quad (4.17a)$$

and

$$ah \leq |e| \leq bh, \quad \forall e \in \mathcal{E} = \{e \subset \partial T_i, T_i \in \mathcal{T}_h\}. \quad (4.17b)$$

On the other hand, we assume a CFL-like stability condition

$$\Delta t \leq Ch. \quad (4.17c)$$

Let us now prove the analog of the Lax–Wendroff theorem. The result justifies the use of conservative numerical fluxes, and the proof presents some interesting arguments, though the assumptions of strong convergence are unrealistic. Indeed, at the present time, no general finite-volume scheme on triangles has yet been proven to converge strongly (only weakly in  $L^\infty$ ) (see Remark 4.5).

### *Proposition 4.1*

Let (4.3) be a conservative finite-volume scheme with Lipschitz continuous numerical flux  $\Phi$  consistent with Eq. (3.6a), and let  $u_i^0$  be given by (4.16). Assume that there exists a sequence  $\Delta_k$  which tends to 0 such that

$$\|v_{\Delta_k}\|_{L^\infty(\mathbb{R}^2 \times (0, +\infty))} \leq C, \quad (4.18)$$

$v_{\Delta_k}$  converges in  $\mathbf{L}^1_{\text{loc}}(\mathbb{R}^2 \times (0, +\infty))$  and a.e. to a function  $v$ . (4.19)

Then  $v$  is a weak solution of (3.6a)–(3.6b).

*Proof.* Let  $\varphi \in \mathbf{C}_0^1(\mathbb{R}^2 \times \mathbb{R}^+)$  be a test function with compact support, and set

$$\varphi_i^n = \frac{1}{|T_i|} \int_{T_i} \varphi(\mathbf{x}, t_n) d\mathbf{x}.$$

By multiplying Eq. (4.3) by  $\varphi_i^n$  and summing over  $i$  (such that  $T_i \in \mathcal{T}_h$ ) and  $n \geq 0$ , we obtain

$$\begin{aligned} & \Delta t \sum_{i,n \geq 0} \int_{T_i} \left\{ \varphi(\mathbf{x}, t_n) \frac{(u_i^{n+1} - u_i^n)}{\Delta t} \right\} d\mathbf{x} \\ & + \Delta t \sum_{i,n \geq 0} \varphi_i^n \sum_{e=\Gamma_{ij} \in \partial T_i} |e| \Phi(u_i^n, u_j^n, \mathbf{n}_e) \} = 0. \end{aligned}$$

The summation reduces, in fact, to the indices  $i, n$  such that  $T_i \times (0, t_n) \cap \text{supp } \varphi \neq \emptyset$ . A summation by parts gives for the first term

$$\begin{aligned} & \Delta t \sum_{i,n \geq 0} \int_{T_i} \left\{ \varphi(\mathbf{x}, t_n) \frac{(u_i^{n+1} - u_i^n)}{\Delta t} \right\} d\mathbf{x} \\ & = \Delta t \sum_{i,n \geq 1} \int_{T_i} u_i^n \left\{ \frac{\varphi(\mathbf{x}, t_{n-1}) - \varphi(\mathbf{x}, t_n)}{\Delta t} \right\} d\mathbf{x} - \sum_i \int_{T_i} \varphi(\mathbf{x}, 0) u_i^0 d\mathbf{x}. \end{aligned}$$

Consider the last term,

$$\sum_i \int_{T_i} \varphi(\mathbf{x}, 0) u_i^0 d\mathbf{x} = \int_{\mathbb{R}^2} \varphi(\mathbf{x}, 0) \sum_i u_i^0 \mathbf{1}_{T_i}(\mathbf{x}) d\mathbf{x}.$$

By (4.16)

$$u_i^0 = \frac{1}{|T_i|} \int_{T_i} u_0(\mathbf{x}) d\mathbf{x},$$

so that

$$\sum_i u_i^0 \mathbf{1}_{T_i}(\mathbf{x}) = v_\Delta(\mathbf{x}, 0) = \pi_h u_0,$$

and  $\pi_h u_0$  converges to  $u_0$  in  $\mathbf{L}^1$  as  $h \rightarrow 0$  by Lemma 4.2. Hence, we have

$$\sum_i \int_{T_i} \varphi(\mathbf{x}, 0) u_i^0 d\mathbf{x} \rightarrow \int_{\mathbb{R}^2} \varphi(\mathbf{x}, 0) u_0(\mathbf{x}) d\mathbf{x} \quad \text{as } h \rightarrow 0. \quad (4.20)$$

Now, setting for  $(\mathbf{x}, t) \in \mathbb{R}^2 \times \mathbb{R}^+$

$$\varphi_\Delta(\mathbf{x}, t) = \varphi(\mathbf{x}, t_n), \quad \text{if } t_n \leq t < t_{n+1}, \quad n \geq 0,$$

we have

$$\begin{aligned} v_\Delta(\mathbf{x}, t) &\{\varphi_\Delta(\mathbf{x}, t) - \varphi(\mathbf{x}, t - \Delta t)\} \\ &= u_i^n \{\varphi(\mathbf{x}, t_n) - \varphi(\mathbf{x}, t_{n-1})\}, \quad \mathbf{x} \in T_i, \quad t_n \leq t < t_{n+1}, \end{aligned}$$

and

$$\begin{aligned} \Delta t \sum_{i,n \geq 1} \int_{T_i} u_i^n \left\{ \frac{\varphi(\mathbf{x}, t_n) - \varphi(\mathbf{x}, t_{n-1})}{\Delta t} \right\} d\mathbf{x} \\ = \int_{\mathbb{R}^2} \int_{\Delta t}^{+\infty} v_\Delta(\mathbf{x}, t) \left\{ \frac{\varphi_\Delta(\mathbf{x}, t) - \varphi_\Delta(\mathbf{x}, t - \Delta t)}{\Delta t} \right\} d\mathbf{x} dt. \end{aligned}$$

The term  $1_{[\Delta t, +\infty)}(t) \frac{\varphi_\Delta(\mathbf{x}, t) - \varphi_\Delta(\mathbf{x}, t - \Delta t)}{\Delta t}$  converges to  $\frac{\partial \varphi}{\partial t}(\mathbf{x}, t)$  as  $\Delta t \rightarrow 0$  and, by assumption,  $v_\Delta$  is bounded. It yields

$$\begin{aligned} \int_{\mathbb{R}^2 \times (0, +\infty)} v_\Delta(\mathbf{x}, t) \left\{ 1_{(\Delta t, +\infty)}(t) \frac{\varphi_\Delta(\mathbf{x}, t) - \varphi_\Delta(\mathbf{x}, t - \Delta t)}{\Delta t} \right. \\ \left. - \frac{\partial \varphi}{\partial t}(\mathbf{x}, t) \right\} d\mathbf{x} dt \rightarrow 0 \end{aligned}$$

as  $\Delta t \rightarrow 0$ . Now, since  $v_\Delta$  converges in  $\mathbf{L}_{\text{loc}}^1$  toward  $v$ ,

$$\begin{aligned} \int_{\mathbb{R}^2 \times (0, +\infty)} v_\Delta(\mathbf{x}, t) \frac{\partial \varphi}{\partial t}(\mathbf{x}, t) d\mathbf{x} dt \\ \rightarrow \int_{\mathbb{R}^2 \times (0, +\infty)} v(\mathbf{x}, t) \frac{\partial \varphi}{\partial t}(\mathbf{x}, t) d\mathbf{x} dt \text{ as } \Delta \rightarrow 0. \end{aligned} \quad (4.21)$$

We have thus proven that

$$\begin{aligned} \int_{\mathbb{R}^2} \int_{\Delta t}^{+\infty} v_\Delta(\mathbf{x}, t) \left\{ \frac{\varphi_\Delta(\mathbf{x}, t) - \varphi_\Delta(\mathbf{x}, t - \Delta t)}{\Delta t} \right\} d\mathbf{x} dt \\ \rightarrow \int_{\mathbb{R}^2 \times (0, +\infty)} v(\mathbf{x}, t) \frac{\partial \varphi}{\partial t}(\mathbf{x}, t) d\mathbf{x} dt \text{ as } \Delta \rightarrow 0. \end{aligned}$$

Let us next study the term

$$\Delta t \sum_{i,n \geq 0} \left\{ \frac{1}{|T_i|} \int_{T_i} \varphi(\mathbf{x}, t_n) \sum_{e=\Gamma_{ij} \subset \partial T_i} |e| \Phi(u_i^n, u_j^n, \mathbf{n}_e) \right\} = 0.$$

First, we introduce the function  $\mathbf{p}_e(\mathbf{x})$  associated to an edge  $e$  by Lemma 4.3 (with slightly different notations). Then we define

$$\tilde{\mathbf{F}}^n(\mathbf{x}) = \sum_{e=\Gamma_{ij} \subset \partial T_i} \mathbf{p}_e(\mathbf{x}) \Phi(u_i^n, u_j^n, \mathbf{n}_e), \quad \mathbf{x} \in T_i$$

(the dependence of  $\mathbf{p}_e(\mathbf{x})$  on  $T_i$  does not appear in the notation but is effective). Thus, by (4.13),

$$\operatorname{div} \tilde{\mathbf{F}}^n(\mathbf{x}) = \sum_{e=\Gamma_{ij} \subset \partial T_i} \frac{|e|}{|\Omega_i|} \Phi(u_i^n, u_j^n, \mathbf{n}_e) \quad \mathbf{x} \in T_i$$

and

$$\begin{aligned} & \sum_i \int_{T_i} \varphi(\mathbf{x}, t_n) \left\{ \sum_e \frac{|e|}{|T_i|} \Phi(u_i^n, u_j^n, \mathbf{n}_e) \right\} d\mathbf{x} \\ &= \sum_i \left\{ \int_{T_i} \varphi(\mathbf{x}, t_n) \operatorname{div} \tilde{\mathbf{F}}^n(\mathbf{x}) d\mathbf{x} \right\} \\ &= - \sum_i \left\{ \int_{\Omega_i} \operatorname{grad} \varphi(\mathbf{x}, t_n) \tilde{\mathbf{F}}^n(\mathbf{x}) d\mathbf{x} \right\} + \sum_i \sum_e \int_e \varphi(\mathbf{x}, t_n) \tilde{\mathbf{F}}^n(\mathbf{x}) \cdot \mathbf{n}_e d\mathbf{x}. \end{aligned}$$

For each edge  $e = \Gamma_{ij} = T_i \cap T_j = \Gamma_{ji}$ , we get two terms: one coming from the integral on  $T_i$  and by (4.13),

$$\tilde{\mathbf{F}}^n(\mathbf{x})|_{\Gamma_{ij}} \cdot \mathbf{n}_e = \Phi(u_i^n, u_j^n, \mathbf{n}_e),$$

and the other from the integral on  $T_j$ , with the opposite normal,

$$\tilde{\mathbf{F}}^n(\mathbf{x})|_{\Gamma_{ij}} \cdot (-\mathbf{n}_e) = \Phi(u_j^n, u_i^n, -\mathbf{n}_e).$$

Thanks to the conservativity (4.4), these two terms cancel,

$$\Phi(u_j^n, u_i^n, -\mathbf{n}_e) = -\Phi(u_i^n, u_j^n, \mathbf{n}_e),$$

and

$$\sum_i \sum_e \int_e \varphi(\mathbf{x}, t_n) \tilde{\mathbf{F}}^n(\mathbf{x}) \cdot \mathbf{n}_e d\mathbf{x} = 0.$$

We are left with the term

$$\sum_i \int_{T_i} \operatorname{grad} \varphi(\mathbf{x}, t_n) \cdot \tilde{\mathbf{F}}^n(\mathbf{x}) d\mathbf{x}.$$

Using now the consistency property (4.5), we write

$$\Phi(u_i^n, u_j^n, \mathbf{n}_e) = \Phi(u_i^n, u_j^n, \mathbf{n}_e) - \Phi(u_i^n, u_i^n, \mathbf{n}_e) + \mathbb{F}(u_i^n) \cdot \mathbf{n}_e,$$

which leads us to define

$$\mathbf{F}^n(\mathbf{x}) = \sum_{e \in \partial \Omega_i} \mathbf{p}_e(\mathbf{x}) \mathbf{n}_e \cdot \mathbb{F}(u_i^n) \quad \mathbf{F}^n(\mathbf{x}) = \mathbb{F}(u_i^n), \quad \mathbf{x} \in T_i,$$

and, summing over  $n \geq 0$ , to consider the limit of the term

$$\Delta t \sum_{n,i} \int_{T_i} \operatorname{grad} \varphi(\mathbf{x}, t_n) \cdot \mathbf{F}^n(\mathbf{x}) d\mathbf{x} = \Delta t \sum_n \int_{\mathbb{R}^2} \operatorname{grad} \varphi(\mathbf{x}, t_n) \cdot \mathbf{F}^n(\mathbf{x}) d\mathbf{x}.$$

We see easily that as  $\Delta \rightarrow 0$ ,

$$\begin{aligned} & \Delta t \sum_n \int_{\mathbb{R}^2} \operatorname{grad} \varphi(\mathbf{x}, t_n) \cdot \mathbf{F}^n(\mathbf{x}) d\mathbf{x} \\ & \rightarrow \int_{\mathbb{R}^2 \times (0, +\infty)} \operatorname{grad} \varphi(\mathbf{x}, t) \cdot \mathbb{F}(v(\mathbf{x}, t)) d\mathbf{x} dt. \end{aligned} \quad (4.22)$$

There remains to study the term  $R$  corresponding to  $\Phi(u_i^n, u_j^n, \mathbf{n}_e) - \Phi(u_i^n, u_i^n, \mathbf{n}_e)$ ,

$$\begin{aligned} \frac{R}{\Delta t} &= \sum_{n,i} \int_{T_i} \operatorname{grad} \varphi(\mathbf{x}, t_n) \cdot \{\mathbf{F}^n(\mathbf{x}) - \tilde{\mathbf{F}}^n(\mathbf{x})\} d\mathbf{x} \\ &= \sum_{n,i} \int_{T_i} \operatorname{grad} \varphi(\mathbf{x}, t_n) \cdot \sum_{e=\Gamma_{ij}} \mathbf{p}_e(\Phi(u_i^n, u_j^n, \mathbf{n}_e) - \Phi(u_i^n, u_i^n, \mathbf{n}_e)) d\mathbf{x}. \end{aligned}$$

The regularity assumptions (4.17) imply that the  $\mathbf{p}_e$  are uniformly bounded, and hence

$$R \leq C \Delta t \sum_{n,i} \int_{T_i} |\operatorname{grad} \varphi(\mathbf{x}, t_n)| \sum_{j/\Gamma_{ij} \subset \partial T_i} |\Phi(u_i^n, u_j^n, \mathbf{n}_e) - \Phi(u_i^n, u_i^n, \mathbf{n}_e)| d\mathbf{x}.$$

Since  $\Phi$  is locally Lipschitz continuous, it remains only to estimate the right-hand side of

$$\begin{aligned} R &\leq C \Delta t \sum_{n,i} \int_{T_i} |\operatorname{grad} \varphi(\mathbf{x}, t_n)| \sum_{j/\Gamma_{ij} \subset \partial T_i} |u_i^n - u_j^n| d\mathbf{x} \\ &\leq C \Delta t \sum_{n,i} \|\operatorname{grad} \varphi\|_{\mathbf{L}^\infty} |\Omega_i| \sum_{j/\Gamma_{ij} \subset \partial T_i} |u_i^n - u_j^n|, \end{aligned}$$

where the summation is limited to those  $i$  such that  $T_i \cap \operatorname{supp} \varphi \neq \emptyset$ . Now, recall that

$$u_i^n = \frac{1}{|T_i| \Delta t} \int_{T_i \times (t_n, t_{n+1})} v_\Delta(\mathbf{x}, t) d\mathbf{x} dt.$$

Then, setting

$$v_i^n = \frac{1}{|T_i| \Delta t} \int_{T_i \times (t_n, t_{n+1})} v(\mathbf{x}, t) d\mathbf{x} dt$$

and writing

$$|u_i^n - u_j^n| \leq |u_i^n - v_i^n| + |v_i^n - v_j^n| + |u_j^n - v_j^n|,$$

we get first, since  $v_\Delta$  converges in  $\mathbf{L}_{\text{loc}}^1$  to  $v$ ,

$$\Delta t \sum_{i,n} |T_i| |u_i^n - v_j^n| = \int_{\text{supp}(\varphi)} |v_\Delta - v| d\mathbf{x} dt \rightarrow 0 \quad \text{as } \Delta \rightarrow 0.$$

For the term  $|u_j^n - v_j^n|$ , we can conclude in an analogous way since, thanks to the assumption of non-degeneracy of the mesh, we can write

$$\sum_i \sum_{j/\Gamma_{ij} \subset \partial T_i} |T_i| |u_j^n - v_j^n| \leq 3C \sum_j |T_j| |u_j^n - v_j^n|.$$

There remains to prove that the term

$$\Delta t \sum_{i,n} |T_i| \sum_{j/\Gamma_{ij} \subset \partial T_i} |v_i^n - v_j^n|$$

converges toward 0. Given  $\varepsilon$ , by the density of  $\mathbf{C}^1$  in  $\mathbf{L}^1$ , we can find a continuously differentiable function  $\psi$  with support in  $\text{supp}(\varphi)$ , such that

$$\|v - \psi\|_{\mathbf{L}^1(\text{supp}(\varphi))} \leq \varepsilon.$$

Setting

$$\psi_i^n = \frac{1}{\Delta t |T_i|} \int_{T_i \times (t_n, t_{n+1})} \psi(\mathbf{x}) d\mathbf{x},$$

by the Lipschitz continuity of  $\psi$ , we have

$$\Delta t \sum_i |T_i| \sum_{j/\Gamma_{ij} \in \partial T_i} |\psi_i^n - \psi_j^n| \rightarrow 0, \quad \text{as } \Delta \rightarrow 0.$$

Using again the assumption of non-degeneracy, we find

$$\Delta t \sum_{i,n} |T_i| \sum_{j/\Gamma_{ij} \subset \partial T_i} |\psi_j^n - v_j^n| \leq 3C \|v - \psi\|_{\mathbf{L}^1(\text{supp}(\varphi))} \leq \varepsilon.$$

Finally, combining (4.20)–(4.22), we conclude that the limit  $v$  satisfies

$$\begin{aligned} & \int_{\mathbb{R}^2 \times [0, +\infty)} \left\{ v(x, t) \frac{\partial \varphi}{\partial t}(\mathbf{x}, t) + \mathbb{F}(v(x, t)) \cdot \text{grad } \varphi(\mathbf{x}, t) \right\} d\mathbf{x} dt \\ & \quad + \int_{\mathbb{R}^2} u_0(\mathbf{x}) \varphi(\mathbf{x}, 0) d\mathbf{x} = 0. \end{aligned}$$

Hence  $v$  is a weak solution of (3.6a)–(3.6b).  $\square$

We now specify some of the main stability results that are available.

### 4.2.3 Stability

We can prove in the scalar case  $L^\infty$  stability, which is the first necessary step for the above mentioned convergence results.

*Proposition 4.2*

Assume that in scheme (4.3) the numerical flux  $\Phi$  is associated by (4.6) to a three-point one-dimensional monotone numerical flux  $\varphi$  (which is Lipschitz continuous). Then, the scheme (4.3) is  $L^\infty$ -stable and satisfies

$$\sum_i |\Omega_i| |u_i^{n+1}| \leq \sum_i |\Omega_i| |u_i^n|$$

under the CFL-like condition

$$\lambda C(\varphi) \leq 1,$$

where  $C(\varphi)$  is the Lipschitz constant of  $\varphi$  and  $\lambda = \Delta t \sup_i (\frac{|\partial\Omega_i|}{|\Omega_i|})$ .

In the formula for  $\lambda$ ,  $|\partial\Omega_i|$  denotes the perimeter of  $\Omega_i$ ; hence for a uniform triangulation by triangles with side  $h$ , we have  $\lambda = 4\sqrt{3}\frac{\Delta t}{h}$ . In the one-dimensional case, it corresponds to a CFL of  $1/2$ .

*Proof.* We can write, again using (4.14),

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\Delta t}{|\Omega_i|} \left\{ \sum_{j/e=\Gamma_{ij} \subset \partial\Omega_i} |e| \Phi(u_i^n, u_j^n, \mathbf{n}_e) \right\} \\ &= u_i^n - \frac{\Delta t}{|\Omega_i|} \left\{ \sum_{j/e=\Gamma_{ij} \subset \partial\Omega_i} |e| (\Phi(u_i^n, u_j^n, \mathbf{n}_e) - \mathbb{F} \cdot \mathbf{n}(u_i^n)) \right\} \end{aligned}$$

and for a consistent numerical flux

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{|\Omega_i|} \left\{ \sum |e| (\Phi(u_i^n, u_j^n, \mathbf{n}_e) - \Phi(u_i^n, u_i^n, \mathbf{n}_e)) \right\}.$$

Multiplying out and dividing each term of the sum by the factor  $(u_j^n - u_i^n)$ , we get

$$\begin{aligned} u_i^{n+1} &= u_i^n \left\{ 1 - \frac{\Delta t}{|\Omega_i|} \sum |\Gamma_{ij}| \frac{\Phi(u_i^n, u_i^n, \mathbf{n}_{ij}) - \Phi(u_i^n, u_j^n, \mathbf{n}_{ij})}{u_j^n - u_i^n} \right\} \\ &\quad + \frac{\Delta t}{|\Omega_i|} \left\{ \sum |\Gamma_{ij}| u_j^n \frac{\Phi(u_i^n, u_i^n, \mathbf{n}_{ij}) - \Phi(u_i^n, u_j^n, \mathbf{n}_{ij})}{u_j^n - u_i^n} \right\}, \end{aligned}$$

where we have denoted  $\mathbf{n}_e = \mathbf{n}_{ij}$  for  $e = \Gamma_{ij}$ . This corresponds, in fact, to expressing the one-dimensional three-point flux  $\varphi(\mathbb{F} \cdot \mathbf{n}; u, v)$  in terms of the incremental coefficients

$$\frac{C(u, v)}{\lambda} = \frac{\varphi(\mathbb{F} \cdot \mathbf{n}; u, u) - \varphi(\mathbb{F} \cdot \mathbf{n}; u, v)}{v - u},$$

$$\frac{D(u, v)}{\lambda} = \frac{\varphi(\mathbb{F} \cdot \mathbf{n}; v, v) - \varphi(\mathbb{F} \cdot \mathbf{n}; u, v)}{v - u}.$$

We then follow the lines of G.R., Chapter 3, Section 3.2 (Theorem 3.6), just replacing the 1d coefficient  $\lambda = \frac{\Delta t}{\Delta x}$  by  $\lambda_i = \frac{\Delta t |\partial\Omega_i|}{|\Omega_i|}$  ( $|\partial\Omega_i|$  denotes the perimeter of  $\Omega_i$ ). The condition in the 1d case,

$$0 \leq C_{j+1/2} \leq \frac{1}{2}, \quad 0 \leq D_{j-1/2} \leq \frac{1}{2},$$

which yields

$$C_{j+1/2} + D_{j-1/2} \leq 1,$$

is now replaced by

$$0 \leq C_{i,j} = \frac{\Delta t |\Gamma_{ij}|}{|\Omega_i|} \frac{\Phi(u_i^n, u_i^n, \mathbf{n}_{ij}) - \Phi(u_i^n, u_j^n, \mathbf{n}_{ij})}{u_j^n - u_i^n} \quad (4.23a)$$

and

$$\frac{C_{i,j} |\partial\Omega_i|}{|\Gamma_{ij}|} \leq 1. \quad (4.23b)$$

The coefficient  $C_{i,j}$  is indeed positive for a monotone scheme since

$$C_{i,j} = \frac{\lambda_i |\Gamma_{ij}|}{|\partial\Omega_i|} C(u_i^n, u_j^n)$$

and  $C(u, v) \leq 0$  ( $v \rightarrow \varphi(\cdot, v)$  is nonincreasing); (4.23a) also holds for other schemes such as  $E$ -schemes.

Inequality (4.23b), which is equivalent to

$$\lambda_i C(u_i^n, u_j^n) \leq 1,$$

holds if

$$\lambda_i \frac{\Phi(u_i^n, u_i^n, \mathbf{n}_{ij}) - \Phi(u_i^n, u_j^n, \mathbf{n}_{ij})}{u_j^n - u_i^n} \leq 1,$$

which appears as a CFL-like condition (linked to the modulus of continuity of  $\varphi$ ). This yields

$$\sum_{j, e=\Gamma_{ij} \subset \partial\Omega_i} C_{i,j} = \frac{\lambda_i}{|\partial\Omega_i|} \sum_j |\Gamma_{ij}| C(u_i^n, u_j^n) \leq 1$$

and proves in turn that  $u_i^{n+1}$  is a convex combination of  $u_i^n$  and the neighboring values  $u_j^n$ ,

$$u_i^{n+1} = u_i^n \left(1 - \sum_j C_{i,j}\right) + \sum_j C_{i,j} u_j^n = u_i^n + \sum_j C_{i,j} (u_j^n - u_i^n).$$

It implies first the local maximum principle

$$\min\left(u_i^n, \min_{e \subset \partial\Omega_i} u_j^n\right) \leq u_i^{n+1} \leq \max\left(u_i^n, \max_{e \subset \partial\Omega_i} u_j^n\right)$$

and also

$$|u_i^{n+1}| \leq \left(1 - \sum_j C_{i,j}\right) |u_i^n| + \sum_j C_{i,j} |u_j^n|.$$

Then, multiplying by  $|\Omega_i|$  and summing over a finite set, for instance, corresponding to the cells intersecting a given ball  $B_R$ , which we denote by  $i \in J(R)$ , we write

$$\sum_{i \in J(R)} |\Omega_i| |u_i^{n+1}| \leq \sum_{i \in J(R)} |\Omega_i| \left\{ \left(1 - \sum_j C_{i,j}\right) |u_i^n| + \sum_j C_{i,j} |u_j^n| \right\}. \quad (4.24)$$

We now observe that in the right-hand side of (4.24), we can reorder the terms and sum over the indices “ $j$ ,” provided we take care of the “boundary” elements  $\Omega_j$  which do not intersect  $B_R$  but then intersect  $B_{R+h}$  (if  $h$  is the maximum of the diameters of the  $\Omega_i$ ). Since by the conservation relation (4.4)

$$\begin{aligned} |\Omega_i| C_{i,j} &= \Delta t |\Gamma_{ij}| \frac{\mathbb{F} \cdot \mathbf{n}_{ij}(u_i^n) - \Phi(u_i^n, u_j^n, \mathbf{n}_{ij})}{u_j^n - u_i^n}, \\ |\Omega_j| C_{j,i} &= \Delta t |\Gamma_{ji}| \frac{-\mathbb{F} \cdot \mathbf{n}_{ij}(u_j^n) - \Phi(u_j^n, u_i^n, -\mathbf{n}_{ij})}{u_i^n - u_j^n} \\ &= \Delta t |\Gamma_{ij}| \frac{\mathbb{F} \cdot \mathbf{n}_{ij}(u_j^n) - \Phi(u_i^n, u_j^n, \mathbf{n}_{ij})}{u_j^n - u_i^n}, \end{aligned}$$

we get for an “interior” element  $\Omega_j$  that in (4.24),  $|u_j^n|$  is multiplied by

$$\sum_{i \text{ neighbors of } j} |\Omega_i| C_{i,j} + |\Omega_j| \left(1 - \sum_{i \text{ neighbors of } j} C_{j,i}\right) = |\Omega_j|.$$

Indeed, for an index  $i$  such that  $\Omega_j$  and  $\Omega_i$  are neighbors,

$$|\Omega_i| C_{i,j} - |\Omega_j| C_{j,i} = \Delta t |\Gamma_{ij}| \frac{\mathbb{F} \cdot \mathbf{n}_{ij}(u_i^n) - \mathbb{F} \cdot \mathbf{n}_{ij}(u_j^n)}{u_j^n - u_i^n},$$

so that by (4.14), summing up over  $i$ , the sum vanishes. Finally, we obtain as expected

$$\sum_{i \in J(R)} |\Omega_i| |u_i^{n+1}| \leq \sum_{i \in J(R+h)} |\Omega_i| |u_i^n|,$$

and then let  $R \rightarrow \infty$ . □

*Remark 4.5.* Note that we could have taken any convex function of  $u$  in place of  $|u|$  (in particular  $\eta(u) = |u|^p$ ), which yields

$$\sum_{i \in J(R)} |\Omega_i| \eta(u_i^{n+1}) \leq \sum_{i \in J(R+h)} |\Omega_i| \eta(u_i^n)$$

and provides other estimates.  $\square$

For what concerns BV stability, we have already observed that assumption (4.19) is unrealistic, and no (strong) BV-estimate is available in general. For the reader's convenience, we just state that the "natural" total variation of the sequence  $u = (u_i)$  is

$$TV(u) = \|u_\Delta\|_{BV(\mathbb{R}^2)},$$

where  $u_\Delta$  is the piecewise constant function associated to  $u$  and BV denotes the space of functions of bounded variation. An easy computation gives

$$TV(u) = \frac{1}{2} \sum_{ij} |\Gamma_{ij}| |u_i - u_j|,$$

which coincides with (3.8) in the case of rectangles.

Such BV-estimates, which hold in the one-dimensional case or on a rectangular 2d mesh, would provide relative compactness of the sequence  $(v_{\Delta_k})$  in  $\mathbf{L}_{loc}^1$  and imply the convergence of  $v_{\Delta_k}$  (or a subsequence) toward  $u$  and of  $\mathbb{F}(v_{\Delta_k})$  toward  $\mathbb{F}(u)$ .

Let us sketch the main ideas that are involved in the currently known convergence results initiated by Szepesty for a multidimensional scalar conservation law. The  $L^\infty$ -bound (4.18) is indeed satisfied by many schemes (Proposition 4.2) and yields convergence of  $(v_{\Delta_k})$  for the weak star topology in  $\mathbf{L}^\infty(\mathbb{R}^2 \times (0, +\infty))$  to some  $u$ . Instead of (strong) BV-estimates, "weak" BV-estimates are obtained, but they cannot allow this relative compactness, and it is well known that one cannot conclude that  $\mathbb{F}(v_{\Delta_k})$  tends to  $\mathbb{F}(u)$ .

Hence, one requires more powerful tools, namely, Young measure (Tartar [1107], Ball [74]) and entropy measure-valued solution (DiPerna [426]). A *Young measure* on  $\mathbb{R}^d \times (0, +\infty)$  is a parametrized family  $(\mathbf{x}, t) \in \mathbb{R}^d \times (0, +\infty) \rightarrow \nu_{\mathbf{x}, t}$  of probability measures on  $\mathbb{R}$ . For a continuous function  $g \in \mathbf{C}^0(\mathbb{R})$ ,

$$\langle \nu_{\mathbf{x}, t}, g(\cdot) \rangle = \int_{\mathbb{R}} g(\lambda) d\nu_{\mathbf{x}, t}(\lambda) = \mu_g(\mathbf{x}, t)$$

denotes the pairing between the probability  $\nu_{\mathbf{x}, t}$  and the function  $g$ ; for  $g(\lambda) = \lambda$ , we write  $\langle \nu_{\mathbf{x}, t}, \lambda \rangle = \int_{\mathbb{R}} \lambda d\nu_{\mathbf{x}, t}(\lambda)$ . In the particular case where  $\nu_{\mathbf{x}, t}$  is a Dirac measure  $\delta_u$ ,

$$\langle \delta_u, g(\cdot) \rangle = \int_{\mathbb{R}} g(\lambda) \delta_u(\lambda) = g(u).$$

In general,  $(\mathbf{x}, t) \in \mathbb{R}^d \times (0, \infty) \mapsto \mu_g(\mathbf{x}, t)$  is measurable, and we can assume for simplicity that it is in  $\mathbf{L}^\infty(\mathbb{R}^d \times (0, +\infty))$ .

*Definition 4.1*

A measure-valued (mv) solution of (3.6a) is a measurable map from  $\mathbb{R}^d \times (0, +\infty)$  to the space  $\text{Prob}(\mathbb{R})$  of nonnegative Radon measure with unit mass, such that for all  $\varphi \in \mathbf{C}_0^1(\mathbb{R}^d \times (0, +\infty))$

$$\int \left\{ \langle \nu_{\mathbf{x}, t}, \lambda \rangle \frac{\partial \varphi}{\partial t} + \sum_j \langle \nu_{\mathbf{x}, t}, f_j(\lambda) \rangle \frac{\partial \varphi}{\partial x_j} \right\} d\mathbf{x} dt = 0.$$

An entropy measure-valued solution satisfies, moreover, a weak entropy inequality (see DiPerna [426], Szepessy [1083]). The uniqueness theorem of DiPerna concerning entropy mv-solutions states that if at time zero an mv-entropy solution is a Dirac measure  $\delta_{u_0}(x)$ , then, for time  $t > 0$ , it remains a Dirac measure  $\delta_{u(\mathbf{x}, t)}$ , where  $u$  is the unique entropy weak solution given in Chap. III. This result has been generalized by Szepessy [1083]; see also [504].

Thanks to DiPerna's result (Theorem 2.1 DiPerna [426]), the uniform bound (4.18) ensures that a Young measure  $(\mathbf{x}, t) \in \mathbb{R}^d \times (0, +\infty) \mapsto \nu_{\mathbf{x}, t}$  can be constructed that represents all  $w^*$ -composite limits of  $(v_{\delta_k})$  for the weak star topology in  $\mathbf{L}^\infty(\mathbb{R}^2 \times (0, +\infty))$ . This means that for any  $g \in \mathbf{C}^0(\mathbb{R})$

$$g(v_{\delta_k}(\mathbf{x}, t)) \rightarrow \mu_g(\mathbf{x}, t) = \langle \nu_{\mathbf{x}, t}, g(\cdot) \rangle (\mathbf{L}^\infty \text{weak star limit}) \quad \text{as } \delta_k \rightarrow 0$$

i.e.,

$$\int g(v_{\delta_k}) \varphi(\mathbf{x}, t) d\mathbf{x} dt \rightarrow \int \langle \nu_{\mathbf{x}, t}, g(\cdot) \rangle \varphi(\mathbf{x}, t) d\mathbf{x} dt.$$

We can thus give a meaning to “ $\lim \mathbb{F}(v_{\delta_k})$ ,” which is defined by setting “ $\lim \mathbb{F}(v_{\delta_k}) = \mu_{\mathbb{F}}(\mathbf{x}, t) = \langle \nu_{\mathbf{x}, t}, \mathbb{F}(\cdot) \rangle$ ” and yields the existence of a measure-valued solution. The aim is to prove that  $\nu_{\mathbf{x}, t}$  is the Dirac measure  $\delta_{u(\mathbf{x}, t)}$

$$\langle \delta_{u(\mathbf{x}, t)}, g(\cdot) \rangle = \int_{\mathbb{R}} g(\lambda) \delta_u(\lambda) = g(u(x, t)).$$

The “weak” BV-estimates provide the extra information that enables one to show that the limit  $u$  of  $(v_{\delta_k})$  satisfies the weak entropy inequality; then, by Szepessy's uniqueness theorem,  $u$  is the unique entropy solution and  $(v_{\delta_k})$  converges strongly in  $\mathbf{L}^p$ ,  $1 \leq p < +\infty$ .

*Remark 4.6.* Let us cite some references with convergence results. In Champier, Gallouët, and Herbin [276], estimate (4.19), weak BV stability, and convergence are proven for an upstream finite-volume scheme in the case  $\mathbb{F}(u) = \mathbf{v}f(u)$ , with  $\text{div } \mathbf{v} = 0$  (see Example 4.3). Also convergence results and error estimates (following Kuznetsov's theorem; see G.R. [539], the Appendix to Chapter 3) for finite-volume monotone or  $E$ -schemes have been obtained by

Cockburn et al. [316, 317, 357] and Benharbit et al. [108]. Both approaches use in particular the relation (4.14), which yields

$$\sum_{e \subset \partial\Omega_i} |a| \mathbb{F}(u) \cdot \mathbf{n}_e = \mathbb{F}(\mathbf{u}) \cdot \sum_{e \subset \partial\Omega_i} |e| \mathbf{n}_e = 0$$

and induces the convex decomposition

$$u_i^{n+1} = \sum_{e \subset \partial\Omega_i} \frac{|e|}{|\partial\Omega_i|} u_{i,e}^{n+1},$$

where

$$u_{i,e}^{n+1} = u_i^n - \Delta t \frac{|\partial\Omega_i|}{|\Omega_i|} \{ \Phi(u_i^n, u_j^n, \mathbf{n}_e) - \mathbb{F}(u_i^n) \cdot \mathbf{n}_e \}$$

or, equivalently,

$$u_{i,e}^{n+1} = u_i^n - \Delta t \frac{|\partial\Omega_i|}{|\Omega_i|} \{ \Phi(u_i^n, u_j^n, \mathbf{n}_e) - \Phi(u_i^n, u_i^n, \mathbf{n}_e) \}$$

due to the consistency of  $\Phi$ . These last expressions are now purely one-dimensional, and one can prove for them a discrete maximum principle in the spirit of Proposition 4.2 and discrete entropy inequalities that lead rather simply to  $L^1$  stability. Weak BV-estimates are obtained by studying more sharply the local entropy production. We refer to the abovementioned papers for details; see also Szepessy [1084] and Perthame [943]. Perthame et al. [947, 948] prove the strong convergence in  $\mathbf{L}^2$  of the so-called  $N$ -scheme (Roe and Sidilkover [991], also [1061]) on general triangulations for a linear constant coefficient equation and also Noelle and coauthors [715, 905, 906, 1187].

Let us cite a few more recent references for stability results and error estimates: [255–257, 318, 343, 344, 465, 717, 828], also [415] (linear case), [672].  $\square$

We turn now to the description of some of the usual schemes in their two-dimensional finite-volume extensions, together with the application to gas dynamics.

### 4.3 Usual Schemes

Let us start with Roe's scheme, which is actually one of the most popular schemes. We first give some detailed computations of the straightforward finite-volume extension of the one-dimensional scheme, and then we shall present a recent attempt to derive a truly multidimensional scheme, which follows the same approach for finding an appropriate linearization.

### 4.3.1 Roe's Scheme

Following the ideas of the previous sections, we want to detail the standard two-dimensional Roe's scheme (see Chap. IV, Sect. 3.2). We first linearize the system (1.1) and then integrate the linearized system above the control cells. Let us set

$$\mathbf{A}_\omega(\mathbf{U}) = \mathbf{A}(\mathbf{U}, \omega),$$

where  $\mathbf{A}(\mathbf{U}, \omega)$  is defined by (1.3). We construct a Roe linearization in each direction  $\omega$ , i.e., a matrix  $A_\omega(\mathbf{U}_L, \mathbf{U}_R)$  such that

$$\begin{cases} (i) & \mathbf{A}_\omega(\mathbf{U}, \mathbf{U}) = \mathbf{A}_\omega(\mathbf{U}) = \mathbf{A}(\mathbf{U}, \omega), \\ (ii) & \mathbf{A}_\omega(\mathbf{U}, \mathbf{V})(\mathbf{U} - \mathbf{V}) = \mathbb{F} \cdot \omega(\mathbf{U}) - \mathbb{F} \cdot \omega(\mathbf{V}), \\ (iii) & \mathbf{A}_\omega(\mathbf{U}, \mathbf{V}) \text{ has real eigenvalues and a complete set of eigenvectors.} \end{cases}$$

If “parameter vectors”  $\mathbf{W}$  (see Chap. IV, Sect. 3) are available, we can define

$$\mathbf{A}_\omega(\mathbf{U}_L, \mathbf{U}_R) = \mathbf{A}(\bar{\mathbf{U}}, \omega),$$

where  $\bar{\mathbf{U}}$  is the Roe's averaged state  $\bar{\mathbf{U}} = \mathbf{U}(\mathbf{W}^*)$ ,  $\mathbf{W}^* = \frac{(\mathbf{w}_L + \mathbf{w}_R)}{2}$ .

When integrating the linearized system over  $C_i$ , Green's formula gives, in particular, a flux across the edge  $e = \Gamma_{ij}$ . We integrate the linearized system with matrix  $A_{\mathbf{n}_{ij}}(\mathbf{U}_i, \mathbf{U}_j)$  in the direction  $\mathbf{n}_{ij}$  normal to  $\Gamma_{ij}$  from data taken at each side of the edge. Then Roe's scheme is associated to the following flux (see Chap. IV (3.35) and Lemma 4.1):

$$\begin{aligned} \Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) &= \frac{1}{2}\{\mathbb{F} \cdot \mathbf{n}(\mathbf{U}_i) + \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_j)\} - \frac{1}{2}|\mathbf{A}_\mathbf{n}(\mathbf{U}_i, \mathbf{U}_j)|(\mathbf{U}_j - \mathbf{U}_i) \\ &= \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_j) - \mathbf{A}_\mathbf{n}^+(\mathbf{U}_i, \mathbf{U}_j)(\mathbf{U}_j - \mathbf{U}_i) \\ &= \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_i) + \mathbf{A}_\mathbf{n}^-(\mathbf{U}_i, \mathbf{U}_j)(\mathbf{U}_j - \mathbf{U}_i). \end{aligned}$$

It is well known that Roe's scheme may admit nonphysical (i.e., entropy violating) shocks at sonic points and, as in the one-dimensional case, some entropy correction is needed (see Chap. IV, Sect. 3.2.3).

In the case of the gas dynamics equations, the numerical flux is

$$\begin{aligned} \mathbb{F} \cdot \mathbf{n} &= (\rho \mathbf{u} \cdot \mathbf{n}, \rho \mathbf{u} \mathbf{u} \cdot \mathbf{n} + p \mathbf{n}, (\rho e + p) \mathbf{u} \cdot \mathbf{n})^T \\ &= (\rho u_n, \rho u_n u + \rho \cos \theta, \rho u_n v + p \sin \theta, (\rho e + p) u_n)^T. \end{aligned}$$

For an ideal gas  $p = (\gamma - 1)\rho\varepsilon$ , defining the enthalpy by

$$H = \frac{(E + p)}{\rho} = \varepsilon + \frac{(u^2 + v^2)}{2} + \frac{p}{\rho},$$

the “parameter vectors” are (see Chap. IV, Sect. 4.1.1)

$$w_1 = \rho^{1/2}, \quad w_2 = \rho^{1/2}u, \quad w_3 = \rho^{1/2}v, \quad w_4 = \rho^{1/2}H.$$

Following exactly the computations of Chap. IV, Sect. 4, we get  $\mathbf{A}_\omega(\mathbf{U}_L, \mathbf{U}_R) = \mathbf{A}_\omega(\bar{\mathbf{U}})$ , where  $\bar{u}, \bar{H}$  are given by Lemma 4.5 of Chap. IV, Sect. 4, and

$$\begin{aligned}\bar{v} &= \frac{(\rho_L^{1/2}v_L + \rho_R^{1/2}v_R)}{(\rho_L^{1/2}\rho_R^{1/2})} \\ \bar{c} &= \kappa\left(\bar{H} - \frac{(\bar{u}^2 - \bar{v}^2)}{2}\right) + \chi.\end{aligned}$$

We then compute the eigenvectors and the coefficients in the basis of eigenvectors (more details are given in Sect. 4.3.3 below). Note that for the numerical flux we have

$$\begin{aligned}\text{if } \bar{\mathbf{u}} \cdot \boldsymbol{\omega} - \bar{c} > 0, \quad &\mathbf{A}_\omega^-(\mathbf{U}_L, \mathbf{U}_R) = 0 \text{ and } \Phi(\mathbf{U}_L, \mathbf{U}_R, \boldsymbol{\omega}) = \mathbb{F} \cdot \boldsymbol{\omega}(\mathbf{U}_L), \\ \text{if } \bar{\mathbf{u}} \cdot \boldsymbol{\omega} + \bar{c} > 0, \quad &\mathbf{A}_\omega^+(\mathbf{U}_L, \mathbf{U}_R) = 0 \text{ and } \Phi(\mathbf{U}_L, \mathbf{U}_R, \boldsymbol{\omega}) = \mathbb{F} \cdot \boldsymbol{\omega}(\mathbf{U}_R).\end{aligned}$$

*Lemma 4.4*

*Roe’s scheme is invariant under rotation.*

*Proof.* Let us check that (4.8) holds. Indeed, consider

$$\mathbf{A}_{\mathbf{e}_1}(\tilde{\mathbf{U}}_L, \tilde{\mathbf{U}}_R) = \mathbf{A}(\bar{\tilde{\mathbf{U}}}, \mathbf{e}_1),$$

where  $\bar{\tilde{\mathbf{U}}}$  is the Roe’s averaged state  $\bar{\tilde{\mathbf{U}}} = \mathbf{U}(\tilde{\mathbf{W}}^*)$ ,  $\tilde{\mathbf{W}}^* = \frac{(\tilde{\mathbf{W}}_L + \tilde{\mathbf{W}}_R)}{2}$ . A simple computation using the definition of the parameter vectors  $\mathbf{W} = \rho^{1/2}(1, \mathbf{u}, H)^T$  yields

$$\tilde{\mathbf{W}}^* = \frac{\tilde{\mathbf{W}}_L + \tilde{\mathbf{W}}_R}{2} = R^{-1} \frac{\mathbf{W}_L + \mathbf{W}_R}{2} = R^{-1} \mathbf{W}^*$$

and

$$R^{-1}\bar{\tilde{\mathbf{U}}} = R^{-1}\mathbf{U}(\mathbf{W}^*) = \mathbf{U}(R^{-1}\mathbf{W}^*) = \mathbf{U}(\mathbf{W}^*).$$

Using (4.9b), it follows that if  $R$  is the rotation such that  $R^{-1}\mathbf{n} = \mathbf{e}_1$  and  $R^{-1}\mathbf{U} = \tilde{\mathbf{U}}$ , we have

$$\mathbf{A}_n(\mathbf{U}_L, \mathbf{U}_R) = \mathbf{A}(\bar{\tilde{\mathbf{U}}}, \mathbf{n}) = R\mathbf{A}(R^{-1}\bar{\tilde{\mathbf{U}}})R^{-1} = R \mathbf{A}_{\mathbf{e}_1}(\tilde{\mathbf{U}}_L, \tilde{\mathbf{U}}_R)R^{-1}$$

or (by (4.9c))

$$\mathbf{A}_n(\mathbf{U}_L, \mathbf{U}_R)\mathbf{V} = R \mathbf{A}_{\mathbf{e}_1}(\tilde{\mathbf{U}}_L, \tilde{\mathbf{U}}_R)\tilde{\mathbf{V}},$$

which implies that (4.8) holds for Roe’s scheme.  $\square$

Roe's scheme can be extended to the real gas case following the arguments of Chap. IV, Sect. 4.2, to reactive flows (Dubroca and Morreeuw [443] and also to computing interface motion (Mulder et al. [878]).

However the numerical solutions stay sensitive to the chosen local triangulation. One way to overcome this problem is to develop a “grid independent wave model” which introduces an additional shear wave propagating in the normal direction (see Rumsey et al. [996]). We shall now develop a two-dimensional Roe's linearization. Other approaches for deriving “trully multidimensional schemes” can be found in Roe and Sidilkover [991] and Tamura and Fujii [1096]; see van Leer [1157] for a survey and Abgrall [3] also [8, 17].

### 4.3.2 Fully Two-Dimensional Roe's Linearization

Recently, some attempts have been made to construct a fully 2d Roe linearization, which we present now (see Deconinck et al. [399, 400], Angrand and Lafon [54]). The point in presenting this particular method is to illustrate the fact that a 2d linearization that respects the ideas of the one-dimensional case is not so easy to construct. Given a triangle  $T$  with vertices  $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k$ , one constructs a Roe linearization matrix  $\mathbf{A}_\omega(T) = \mathbf{A}_\omega(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k)$  which depends on the three states. This method is not “consistent” (see [4] for a more precise statement) with the Euler system but guarantees hyperbolicity. Another approach which we shall not develop (see again [4]) derives jump relations and finds a linearization that respects these relations; the corresponding method is “consistent” but does not guarantee hyperbolicity.

One defines first for  $\omega = \mathbf{e}_1$ ,  $\mathbf{A}_{\mathbf{e}_1}(T) = \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k)$  as the Jacobian  $\mathbf{A} = \mathbf{f}'$  computed at the averaged state  $\bar{\mathbf{U}} = \mathbf{U}(\mathbf{W}^*)$ , where now

$$\mathbf{W}^* = \frac{1}{3}(\mathbf{W}_i + \mathbf{W}_j + \mathbf{W}_k)$$

is the parameter vector of the centroid of  $T$ ; ones makes a similar definition for  $\mathbf{B} = \mathbf{g}'$ .

Indeed, let us recall the situation in the one-dimensional case: given a parameter vector  $\mathbf{W}$  (such that  $\mathbf{U} = \mathbf{U}(\mathbf{W})$  and  $\mathbf{f}(\mathbf{U}) = \mathbf{f} \circ \mathbf{U}(\mathbf{W})$  are quadratic), we wrote

$$\begin{aligned} \mathbf{U}_R - \mathbf{U}_L &= \mathbf{U}(\mathbf{W}_R) - \mathbf{U}(\mathbf{W}_L) \\ &= \int_0^1 \mathbf{U}'(\mathbf{W}_L + \theta(\mathbf{W}_R - \mathbf{W}_L)) d\theta (\mathbf{W}_R - \mathbf{W}_L). \end{aligned}$$

Since  $\mathbf{U}'(\mathbf{W})$  varies linearly, we have

$$\int_0^1 \mathbf{U}'(\mathbf{W}_L + \theta(\mathbf{W}_R - \mathbf{W}_L)) d\theta = \mathbf{U}'(\mathbf{W}^*),$$

where

$$\mathbf{W}^* = \frac{(\mathbf{W}_R + \mathbf{W}_L)}{2};$$

similarly,

$$\begin{aligned} \mathbf{f}(\mathbf{U}_R) - \mathbf{f}(\mathbf{U}_L) &= \mathbf{f} \circ \mathbf{U}(\mathbf{W}_R) - \mathbf{f} \circ \mathbf{U}(\mathbf{W}_L) \\ &= \int_0^1 (\mathbf{f} \circ \mathbf{U})'(\mathbf{W}_L + \theta(\mathbf{W}_R - \mathbf{W}_L)) d\theta(\mathbf{W}_R - \mathbf{W}_L) \\ &= (\mathbf{f} \circ \mathbf{U})'(\mathbf{W}^*)(\mathbf{W}_R - \mathbf{W}_L) \\ &= \mathbf{f}'(\mathbf{U}(\mathbf{W}^*)) \mathbf{U}'(\mathbf{W}^*)(\mathbf{W}_R - \mathbf{W}_L), \end{aligned}$$

and hence

$$\mathbf{f}(\mathbf{U}_R) - \mathbf{f}(\mathbf{U}_L) = \mathbf{f}'(\mathbf{U}(\mathbf{W}^*))(\mathbf{U}_R - \mathbf{U}_L),$$

which led us to define

$$\mathbf{A}(\mathbf{U}_L, \mathbf{U}_R) = \mathbf{f}'(\mathbf{U}(\mathbf{W}^*)).$$

The analog of this property (or (4.4), Chap. IV, Sect. 4) in the two-dimensional case gives, if  $L(\mathbf{W})$  is a linear function of  $\mathbf{W}$  and  $\mathbf{W}$  is linear in  $\mathbf{x}$ ,

$$\frac{1}{|T|} \int_T L(\mathbf{W}(\mathbf{x})) d\mathbf{x} = L(\mathbf{W}^*) = \frac{L(\mathbf{W}_i + \mathbf{W}_j + \mathbf{W}_k)}{3},$$

and thus if  $\mathbf{W}$  is linear in  $\mathbf{x}$

$$\frac{1}{|T|} \int_T \mathbf{U}'(\mathbf{W}) d\mathbf{x} = \mathbf{U}'(\mathbf{W}^*) = \frac{\mathbf{U}'(\mathbf{W}_i + \mathbf{W}_j + \mathbf{W}_k)}{3}. \quad (4.25)$$

If we mimic the one-dimensional case, we are led to define

$$\mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) = \mathbf{f}'(\mathbf{U}(\mathbf{W}^*)), \quad (4.26a)$$

and similarly

$$\mathbf{A}_{\mathbf{e}_2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) = \mathbf{g}'(\mathbf{U}(\mathbf{W}^*)). \quad (4.26b)$$

Then, for any vector  $\boldsymbol{\omega}$  in  $\mathbb{R}^2$ ,

$$\begin{aligned} \mathbf{A}_{\boldsymbol{\omega}}(T) &= \mathbf{A}_{\boldsymbol{\omega}}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) \\ &= \omega_1 \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) + \omega_2 \mathbf{A}_{\mathbf{e}_2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) = \mathbf{A}(\bar{\mathbf{U}}, \boldsymbol{\omega}), \end{aligned} \quad (4.26c)$$

with the notation (1.3), where  $\bar{\mathbf{U}} = \mathbf{U}(\mathbf{W}^*)$ .

This linearization is consistent, since

$$\mathbf{A}_{\mathbf{e}_1}(\mathbf{U}, \mathbf{U}, \mathbf{U}) = \mathbf{A}(\mathbf{U}) = \mathbf{f}'(\mathbf{U}), \quad \mathbf{A}_{\mathbf{e}_2}(\mathbf{U}, \mathbf{U}, \mathbf{U}) = \mathbf{B}(\mathbf{U}) = \mathbf{g}'(\mathbf{U})$$

and for any  $\omega$ ,  $\mathbf{A}_\omega(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) = \mathbf{A}(\bar{\mathbf{U}}, \omega)$  has real eigenvalues and a complete set of eigenvectors. There remains to check the analog of the property (ii) of a Roe matrix. Since there is no obvious analog of the increment  $\Delta \mathbf{U} = \mathbf{U}_R - \mathbf{U}_L$ , we write instead

$$\begin{cases} \delta_{Tx}\mathbf{f} = \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_\kappa)\delta_{Tx}\mathbf{U}, \\ \delta_{Ty}\mathbf{g} = \mathbf{A}_{\mathbf{e}_2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_\kappa)\delta_{Ty}\mathbf{U}, \end{cases} \quad (4.27)$$

for some “derivation rule”  $\delta_T = (\delta_{Tx}, \delta_{Ty})$  directly inspired by the above computations and which we discuss now.

We can write the fact that  $\mathbf{U}$  and  $\mathbf{f}(\mathbf{U}), \mathbf{g}(\mathbf{U})$  are homogeneous quadratic functions of  $\mathbf{W}$  in the following form:

$$\begin{aligned} \mathbf{U}(\mathbf{W}) &= Q_{\mathbf{u}}(\mathbf{W}, \mathbf{W}) = \frac{1}{2}\mathbf{U}'(\mathbf{W})\mathbf{W}, \\ (\mathbf{f} \circ \mathbf{U})(\mathbf{W}) &= Q_{\mathbf{f}}(\mathbf{W}, \mathbf{W}) = \frac{1}{2}(\mathbf{f} \circ \mathbf{U})'(\mathbf{W})\mathbf{W}, \\ (\mathbf{g} \circ \mathbf{U})(\mathbf{W}) &= Q_{\mathbf{g}}(\mathbf{W}, \mathbf{W}) = \frac{1}{2}(\mathbf{g} \circ \mathbf{U})'(\mathbf{W})\mathbf{W}, \end{aligned}$$

where the  $Q$ 's are bilinearly symmetric and  $\mathbf{U}'$ ,  $(\mathbf{f} \circ \mathbf{U})'$ , and  $(\mathbf{g} \circ \mathbf{U})'$  are matrices depending linearly on  $\mathbf{W}$  (which are easily computed in the examples). Let us set for any quadratic function  $\mathbf{h}(\mathbf{W}) = Q_{\mathbf{h}}(\mathbf{W}, \mathbf{W})$

$$\hat{\mathbf{h}}_T = \hat{\mathbf{h}}_T(\mathbf{W}) = Q_{\mathbf{h}}(\mathbf{W}^*, \mathbf{W}). \quad (4.28)$$

Thus, freezing one of the variables, at the state  $\mathbf{W}^*$ , we introduce the linearized variables

$$\hat{\mathbf{U}}_T = \hat{\mathbf{U}}(\mathbf{W}) = Q_{\mathbf{U}}(\mathbf{W}^*, \mathbf{W}) = \frac{1}{2}\mathbf{U}'(\mathbf{W}^*)\mathbf{W}, \quad (4.28a)$$

$$\hat{\mathbf{f}}_T = Q_{\mathbf{f}}(\mathbf{W}^*, \mathbf{W}) = \frac{1}{2}(\mathbf{f} \circ \mathbf{U})'(\mathbf{W}^*)\mathbf{W}, \quad (4.28b)$$

$$\hat{\mathbf{g}}_T = Q_{\mathbf{g}}(\mathbf{W}^*, \mathbf{W}) = \frac{1}{2}(\mathbf{g} \circ \mathbf{U})'(\mathbf{W}^*)\mathbf{W}, \quad (4.28c)$$

and define  $\delta_T = (\delta_{Tx}, \delta_{Ty})$  by

$$\delta_T \mathbf{h} = 2\nabla \hat{\mathbf{h}}_T, \quad (4.29)$$

i.e.,

$$\delta_{Tx} \mathbf{h} = 2 \frac{\partial \hat{\mathbf{h}}_T}{\partial x}, \quad \delta_{Ty} \mathbf{h} = 2 \frac{\partial \hat{\mathbf{h}}_T}{\partial y}.$$

Thus

$$\begin{aligned}\delta_{Tx} \mathbf{U} &= 2 \frac{\partial \hat{\mathbf{h}}_T}{\partial x}, & \delta_{Ty} \mathbf{U} &= 2 \frac{\partial \hat{U}_T}{\partial y}, \\ \delta_{Tx} \mathbf{f}(\mathbf{U}) &= 2 \frac{\partial \hat{\mathbf{f}}}{\partial x}, & \delta_{Ty} \mathbf{g}(\mathbf{U}) &= 2 \frac{\partial \hat{\mathbf{g}}}{\partial y}.\end{aligned}$$

*Remark 4.7.* The “2” on the right-hand side of (4.29) may seem funny but is natural since the functions are quadratic; if we consider the simple example  $u(w) = w^2$ ,  $\hat{u}(w) = w^* w$ , then  $u'(w) = 2w$ , and  $2\hat{u}'(w) = 2w^*$  approximates  $u'(w)$  better than  $\hat{u}'(w)$  does!  $\square$

The following result, together with the desired property (4.27) (analog of the relation on the increments (ii) for a Roe matrix), shows that  $\delta_T$  appears as the mean of the gradient on the triangle.

*Lemma 4.5*

For a triangle  $T$  with vertices  $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k$ , define  $\delta_T = (\delta_{Tx}, \delta_{Ty})$  by formulas (4.28) and (4.29) and assume that  $\mathbf{W}$  is a linear function  $\mathbf{x}$  on  $T$ . Then, we have the identities

$$\begin{aligned}\delta_{Tx} \mathbf{U} &= \frac{1}{|T|} \int_T \frac{\partial \mathbf{U}}{\partial x} d\mathbf{x}, & \delta_{Ty} \mathbf{U} &= \frac{1}{|T|} \int_T \frac{\partial \mathbf{U}}{\partial y} d\mathbf{x}, \\ \delta_{Tx}(\mathbf{f}) \mathbf{U} &= \frac{1}{|T|} \int_T \frac{\partial}{\partial x} \mathbf{f}(\mathbf{U}) d\mathbf{x}, & \delta_{Ty}(\mathbf{g}) \mathbf{U} &= \frac{1}{|T|} \int_T \frac{\partial}{\partial y} \mathbf{g}(\mathbf{U}) d\mathbf{x},\end{aligned}\tag{4.30}$$

and (4.27) holds, i.e.,

$$\delta_{Tx} \mathbf{f}(\mathbf{U}) = \mathbf{A}_{\mathbf{e}1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) \delta_{Tx} \mathbf{U}, \quad \delta_{Ty} \mathbf{f}(\mathbf{U}) = \mathbf{A}_{\mathbf{e}2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) \delta_{Ty} \mathbf{U},$$

where  $A_{\mathbf{e}1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k)$  and  $A_{\mathbf{e}2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k)$  are defined by (4.26).

*Proof.* Using (4.25), we write

$$\frac{\partial \hat{\mathbf{U}}_T}{\partial x} = \mathbf{U}'(\mathbf{W}^*) \frac{\partial \mathbf{W}}{\partial x} = \frac{1}{|T|} \int_T \mathbf{U}'(\mathbf{W}) d\mathbf{x} \frac{\partial \mathbf{W}}{\partial x},$$

and if we assume that  $\mathbf{W}$  is a linear function of  $\mathbf{x}$  on  $T$ , then

$$\int_T \mathbf{U}'(\mathbf{W}) d\mathbf{x} \frac{\partial \mathbf{W}}{\partial x} = \int_T \mathbf{U}'(\mathbf{W}) \frac{\partial \mathbf{W}}{\partial x} d\mathbf{x} = \int_T \frac{\partial \mathbf{U}}{\partial x} d\mathbf{x};$$

thus

$$\delta_{Tx} \mathbf{U}(\mathbf{W}) = \frac{1}{|T|} \int_T \frac{\partial \mathbf{U}}{\partial x} d\mathbf{x}.$$

Similarly,

$$\delta_{Ty} \mathbf{U}(\mathbf{W}) = \frac{\partial \hat{\mathbf{U}}(\mathbf{W})}{\partial y} = \mathbf{U}'(\mathbf{W}^*) \frac{\partial \mathbf{W}}{\partial y} = \frac{1}{|T|} \int_T \frac{\partial \mathbf{U}}{\partial y} d\mathbf{x},$$

with analogous formulas for  $\mathbf{f}$  and  $\mathbf{g}$ , which proves (4.30).

Now, we have

$$\begin{aligned} \delta_{Tx} \mathbf{f}(\mathbf{U}) &= (\mathbf{f} \circ \mathbf{U})'(\mathbf{W}^*) \frac{\partial \mathbf{W}}{\partial x} = \mathbf{f}'(\mathbf{U}(\mathbf{W}^*)) \mathbf{U}'(\mathbf{W}^*) \frac{\partial \mathbf{W}}{\partial x} \\ &= \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) \mathbf{U}'(\mathbf{W}^*) \frac{\partial \mathbf{W}}{\partial \mathbf{x}} = \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) \frac{\partial \hat{\mathbf{U}}_T}{\partial x} \end{aligned}$$

and a similar formula for  $\mathbf{g}$ , which gives the desired result (4.27).  $\square$

Now, we use this linearization in the finite-volume method, where the cells are those of Example 4.2 (cell vertex). The triangle  $T = (\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k)$  is divided into three quadrangular regions  $R_i, R_j, R_k$ , each belonging to a different control cell  $C_i, C_j, C_k$  and having a common center of mass. Consider a segment  $e_{ij} = \mathbf{g}\mathbf{m}_{ij}$  ( $\mathbf{m}_{ij}$  is the midpoint of  $\mathbf{a}_i \mathbf{a}_j$ ,  $\mathbf{g}$  the barycenter of  $(\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k)$ ) separating  $R_i$  and  $R_j$ , which is thus part of the boundary between the dual cells  $C_i$  and  $C_j$ , and denote by  $\mathbf{n}_{ij} = (n_{ij,x}, n_{ij,y})^T$  the outward unit normal to  $e_{ij}$  (pointing in the direction of  $R_j$ ). Thus, we define the matrix  $\mathbf{A}_{ij}(T) = \mathbf{A}_{\mathbf{n}_{ij}}(T)$  by (4.26),

$$\begin{aligned} \mathbf{A}_{ij}(T) &= \mathbf{A}_{\mathbf{n}_{ij}}(T) \\ &= n_{ij,x} \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) + n_{ij,y} \mathbf{A}_{\mathbf{e}_2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) = \mathbf{A}(\bar{\mathbf{U}}, \mathbf{n}_{ij}), \end{aligned}$$

with the notation (1.3), where  $\bar{\mathbf{U}} = \mathbf{U}(\mathbf{W}^*)$ ,  $\mathbf{W}^* = \frac{(\mathbf{W}_i + \mathbf{W}_j + \mathbf{W}_k)}{3}$ .

Consider a cell  $C_i$  that is the union of quadrangular regions  $R_i^T$  of type  $R_i$ . As in the general finite-volume method of Example 3.2, we integrate over  $C_i = \cup R_i^T$  and, following Roe's method, we integrate, in fact, a linearized system. More precisely, here we integrate the linearized system in the “linearized” variables  $\hat{\mathbf{U}}$ , and on  $R_i^T = R_i$  (dropping the dependence on  $T$ ) Green's formula gives in particular a flux across the segment  $e_{ij}$ . Thus, as we have already explained for a general finite-volume method, we are led to approach the solution of the one-dimensional system

$$\frac{\partial \hat{\mathbf{U}}}{\partial t} + \mathbf{A}_{ij}(T) \frac{\partial \hat{\mathbf{U}}}{\partial \zeta} = 0, \quad \zeta = \mathbf{x} \cdot \mathbf{n}_{ij},$$

where

$$\tilde{\mathbf{U}} = \begin{cases} \hat{\mathbf{U}}_i & \zeta < 0, \\ \hat{\mathbf{U}}_j & \zeta > 0, \end{cases}$$

and with the linearized matrix  $\mathbf{A}_{ij}(T)$  in the direction  $\mathbf{n}_{ij}$  normal to  $e_{ij}$ . By Lemma 4.5 and the definition of  $\mathbf{A}_{ij}(T) = n_{ij,x} \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) + n_{ij,y} \mathbf{A}_{\mathbf{e}_2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) = \mathbf{A}(\bar{\mathbf{U}}, \mathbf{n}_{ij})$ , we have

$$\begin{aligned} \mathbf{A}_{\mathbf{e}_1}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) n_{ij,x} \frac{\partial \hat{\mathbf{U}}}{\partial x} + \mathbf{A}_{\mathbf{e}_2}(\mathbf{U}_i, \mathbf{U}_j, \mathbf{U}_k) n_{ij,x} \frac{\partial \hat{\mathbf{U}}}{\partial y} \\ = n_{ij,x} \frac{\partial \hat{\mathbf{f}}}{\partial x} + n_{ij,y} \frac{\partial \hat{\mathbf{g}}}{\partial y}, \end{aligned}$$

and we approximate in an upwind way the one-dimensional system relative to the flux  $\hat{\mathbb{F}} \cdot \mathbf{n}_{ij}$ , thus defining

$$\Phi_T(\hat{\mathbf{U}}_i, \hat{\mathbf{U}}_j, \mathbf{n}_{ij}) = \frac{1}{2} \{ \hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_i) + \hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_j) \} - \frac{1}{2} |\mathbf{A}_{ij}(T)| (\hat{\mathbf{U}}_j - \hat{\mathbf{U}}_i).$$

Thanks to Lemma 4.3,

$$\hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_i) + \hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_j) = \mathbf{A}_{ij}(T)(\hat{\mathbf{U}}_j - \hat{\mathbf{U}}_i).$$

Thus, we may also write

$$\begin{aligned} \Phi_T(\hat{\mathbf{U}}_i, \hat{\mathbf{U}}_j, \mathbf{n}_{ij}) &= \hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_j) - \mathbf{A}_{ij}^+(T)(\hat{\mathbf{U}}_j - \hat{\mathbf{U}}_i) \\ &= \hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_i) + \mathbf{A}_{ij}^-(T)(\hat{\mathbf{U}}_j - \hat{\mathbf{U}}_i). \end{aligned}$$

The local increments in the linearized conservative variables  $\hat{\mathbf{U}}$  are computed through

$$\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j = \mathbf{U}'(\mathbf{W}^*)(\mathbf{W}_i - \mathbf{W}_j). \quad (4.31)$$

Among the various “entropy fix,” we shall only detail the local Lax–Friedrichs scheme (at a sonic point, one turns back from the upwind flux to the local Lax–Friedrichs flux; see Chap. IV, Sect. 3.2.3). The procedure is applied to each characteristic field indexed by  $K$ . Thus, denote by  $a_K^{ij}$  the eigenvalues and  $\mathbf{r}_K^{ij}$  the eigenvectors of  $\mathbf{A}_{ij}(T) = \mathbf{A}(\bar{\mathbf{U}}, \mathbf{n}_{ij})$  and  $\mathbf{l}_K^{ij}$  the eigenvectors of  $\mathbf{A}_{ij}^T$ . We decompose the increment on the eigenbasis  $(\mathbf{r}_K^{ij})$ ,

$$\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j = \sum_K \alpha_K^{ij} \mathbf{r}_K^{ij},$$

where

$$\alpha_K^{ij} = \mathbf{l}_K^{ij T} (\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j) = \mathbf{l}_K^{ij T} (\mathbf{U}'(\mathbf{W}^*)(\mathbf{W}_i - \mathbf{W}_j)).$$

We shall also denote by a subscript  $K$  (i.e., by  $\mathbf{V}_K$ ) the coefficient of a vector  $\mathbf{V}$  on the eigenvector  $\mathbf{r}_K^{ij}$  in the decomposition on the eigenbasis  $(\mathbf{r}_K^{ij})$

$$\mathbf{V}_K = \mathbf{l}_K^{ij T} \mathbf{V}.$$

Then in the  $K$ th component  $\psi_K^{ij}$ , the LLF flux  $\psi^{ij}$  on  $\mathbf{r}_K^{ij}$  is

$$\psi_K^{ij} = \frac{1}{2} \{ \hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_i)_K + \hat{\mathbb{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_j)_K \} - \sigma_K^{ij} \alpha_K^{ij},$$

where the coefficients  $\sigma_K^{ij}$  are defined by

$$\sigma_K^{ij} = \max_{i,j,k} \{a_K^{ij}(A_i), a_K^{ij}(A_j), a_K^{ij}(A_k)\}.$$

The  $a_k^{ij}(A_i)$  denote the eigenvalues of the matrix  $\mathbf{A}(\mathbf{U}(\mathbf{a}_i), \mathbf{n}_{ij})$ , whereas the  $a_k^{ij}$  denote the eigenvalues of the matrix  $\mathbf{A}_{ij} = \mathbf{A}(\mathbf{U}(\mathbf{W}^*), \mathbf{n}_{ij})$ . If the three signs of  $a_k^{ij}(\mathbf{a}_i), a_k^{ij}(\mathbf{a}_j), a_k^{ij}(\mathbf{a}_k)$  relative to the three vertices of  $T$  are identical, we take for  $\Phi_K^{ij}$  the  $K$ th component of the upwind flow, i.e.,

$$\begin{aligned} (\hat{\mathbf{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_i))_K &\quad \text{if } a_K^{ij} \geq 0, \\ (\hat{\mathbf{F}} \cdot \mathbf{n}_{ij}(\mathbf{U}_j))_K &\quad \text{if } a_K^{ij} < 0. \end{aligned}$$

In the “sonic case,” for an index  $K$  such that the three signs of  $a_K^{ij}(\mathbf{a}_i), a_K^{ij}(\mathbf{a}_j), a_K^{ij}(\mathbf{a}_k)$  are not identical, we turn to the LLF flux and take for  $\Phi_K^{ij}$  the  $K$ th component  $\psi_K^{ij}$ ; finally,

$$\Phi_T(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{n}_{ij}) = \sum_K \Phi_K^{ij} \mathbf{r}_K^{ij}.$$

For an “optimal” choice of the coefficients  $\sigma_K^{ij}$ , we refer to Angrand and Lafon [54].

### 4.3.3 Application to Gas Dynamics

In the case of the Euler system, the parameter vectors are

$$\begin{aligned} w_1^* &= \frac{1}{3}(\sqrt{\rho_i} + \sqrt{\rho_j} + \sqrt{\rho_k}), \\ w_2^* &= \frac{1}{3}(\sqrt{\rho_i} u_i + \sqrt{\rho_j} u_j + \sqrt{\rho_k} u_k), \\ w_3^* &= \frac{1}{3}(\sqrt{\rho_i} v_i + \sqrt{\rho_j} v_j + \sqrt{\rho_k} v_k), \\ w_4^* &= \frac{1}{3}(\sqrt{\rho_i} H_i + \sqrt{\rho_j} H_j + \sqrt{\rho_k} H_k); \end{aligned}$$

moreover,

$$\bar{u} = \frac{w_2^*}{w_1^*}, \quad \bar{v} = \frac{w_3^*}{w_1^*}, \quad \bar{H} = \frac{w_4^*}{w_1^*}.$$

The matrix  $\mathbf{U}'(\mathbf{W})$  is easily computed for an ideal gas, following Lemma 4.2 in Chap. IV:

$$\mathbf{U}'(\mathbf{W}) = \begin{pmatrix} 2w_1 & 0 & 0 & 0 \\ w_2 & w_1 & 0 & 0 \\ w_0 & 0 & w_1 & 0 \\ w_4/\gamma & w_2(\gamma-1)/\gamma & w_3(\gamma-1)/\gamma & w_1/\gamma \end{pmatrix}.$$

Note that the above matrix is triangular and involves only the density on the diagonal. We get similarly  $(\mathbf{f} \circ \mathbf{U})'(\mathbf{W})$ ,  $(\mathbf{g} \circ \mathbf{U})'(\mathbf{W})$ , and

$$\mathbf{f}'(\mathbf{U}(\mathbf{W}^*)) = (\mathbf{f} \circ \mathbf{U})'(\mathbf{W}^*) \mathbf{U}'(\mathbf{W}^*)^{-1}.$$

As in the one-dimensional case, one can compute the Jacobian matrices  $\mathbf{A}$  and  $\mathbf{B}$  in terms of  $u$ ,  $v$ ,  $H$ , and  $c$ . Setting

$$K = \gamma - 1, \quad K = \frac{\kappa(u^2 + v^2)}{2}, \quad c^2 = \kappa h = \kappa \left( H - \frac{(u^2 + v^2)}{2} \right),$$

where  $H = \frac{(E+p)}{\rho} = \varepsilon + \frac{(u^2+v^2)}{2} + \frac{p}{\rho}$ , we get

$$\mathbf{A}(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ K - u^2 & (2 - \kappa)u & -\kappa v & \kappa \\ -uv & v & u & 0 \\ u(K - H) & H - \kappa u^2 & -\kappa uv & (1 + \kappa)u \end{pmatrix},$$

$$\mathbf{B}(\mathbf{U}) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ -uv & v & u & 0 \\ K - v^2 & -\kappa u & (2 - \kappa)v & \kappa \\ v(K - H) & -\kappa uv & H - \kappa v^2 & (1 + \kappa)v \end{pmatrix},$$

which gives the matrix  $\mathbf{A}_\omega$ , whose eigenvalues are  $\mathbf{u} \cdot \boldsymbol{\omega} \pm c$  and  $\mathbf{u} \cdot \boldsymbol{\omega}$  (double eigenvalue). The eigenvectors are (see Sect. 2.3)

$$\begin{cases} \mathbf{r}_1(\mathbf{U}, \boldsymbol{\omega}) = (1, u - \omega_1 c, v - \omega_2 c, H - \mathbf{u} \cdot \boldsymbol{\omega} c)^T, \\ \mathbf{r}_4(\mathbf{U}, \boldsymbol{\omega}) = (1, u + \omega_1 c, v + \omega_2 c, H + \mathbf{u} \cdot \boldsymbol{\omega} c)^T, \\ \mathbf{r}_2(\mathbf{U}, \boldsymbol{\omega}) = (1, u, v, |\mathbf{u}|^2/2)^T, \\ \mathbf{r}_3(\mathbf{U}, \boldsymbol{\omega}) = (0, -\omega_2, \omega_1, \mathbf{u} \cdot \boldsymbol{\omega}^\perp)^T. \end{cases} \quad (4.32)$$

In the decomposition of  $\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j = \mathbf{U}'(\mathbf{W}^*)(\mathbf{W}_i - \mathbf{W}_j)$ , we have

$$\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j = (\Delta\rho, \Delta\rho u, \Delta\rho v, \Delta\rho e) = \sum_k \alpha_k^{ij} \mathbf{r}_k^{ij},$$

where

$$\begin{aligned}
\Delta\rho &= 2w_1^* \Delta w_1 = 2\sqrt{\rho^*} \Delta \sqrt{\rho} \\
&= \alpha_1^{ij} + \alpha_2^{ij} + \alpha_4^{ij}, \\
\Delta\rho u &= w_2^* \Delta w_1 + w_1^* \Delta w_2 = (\sqrt{\rho u})^* \Delta \sqrt{\rho} + \sqrt{\rho^*} \Delta (\sqrt{\rho u}) \\
&= \alpha_1^{ij}(u - \omega_1 c) + \alpha_2^{ij}u - \alpha_3^{ij}\omega_2 + \alpha_4^{ij}(u - \omega_1 c), \\
\Delta\rho v &= w_3^* \Delta w_1 + w_1^* \Delta w_3 = (\sqrt{\rho v})^* \Delta \sqrt{\rho} + \sqrt{\rho^*} \Delta (\sqrt{\rho v}) \\
&= \alpha_1^{ij}(v - \omega_2 c) + \alpha_2^{ij}v + \alpha_3^{ij}\omega_1 + \alpha_4^{ij}(v - \omega_2 c), \\
\Delta\rho e &= \frac{w_4^*}{\gamma} \Delta w_1 + w_2^* \frac{(\gamma-1)}{\gamma} \Delta w_2 + w_3^* \frac{(\gamma-1)}{\gamma} \Delta w_3 + \frac{w_1^*}{\gamma} \Delta w_4 \\
&= \alpha_1^{ij}(H - \mathbf{u} \cdot \boldsymbol{\omega} c) + \alpha_2^{ij} \frac{|\mathbf{u}|^2}{2} + \alpha_3^{ij} \mathbf{u} \cdot \boldsymbol{\omega}^\perp + \alpha_4^{ij}(H + \mathbf{u} \cdot \boldsymbol{\omega} c),
\end{aligned}$$

which gives a system in the  $\alpha_k^{ij}$  (see Lemma 4.7, Chap. IV). We find (with the notations of Chap. IV, Sect. 4.1.2)

$$\begin{aligned}
\alpha_1^{ij} &= \frac{1}{2\bar{c}^2} (\Delta p - \bar{c} \hat{m}(\rho) \Delta(\mathbf{u} \cdot \boldsymbol{\omega})), \\
\alpha_4^{ij} &= \frac{1}{2\bar{c}^2} (\Delta p + \bar{c} \hat{m}(\rho) \Delta(\mathbf{u} \cdot \boldsymbol{\omega})), \\
\alpha_2^{ij} &= \Delta\rho - \frac{\Delta p}{\bar{c}^2}, \\
\alpha_3^{ij} &= \hat{m}(\rho) \Delta(\mathbf{u} \cdot \boldsymbol{\omega}^\perp).
\end{aligned}$$

The method can be extended to the real gas case following the arguments of Chap. IV, Sect. 4.2.

#### 4.3.4 The Osher Scheme

We now follow the derivation of Chap. IV, Sect. 3.4, and define the two-dimensional Osher flux by

$$\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) = \frac{1}{2} \{ \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_i) + \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_j) - \int_{\Gamma} |\mathbf{A}(\mathbf{w}, \mathbf{n})| d\mathbf{w} \}, \quad (4.33)$$

where  $\Gamma$  is a path consisting of portions  $\Gamma_1 = \mathbf{U}_i \mathbf{U}_1^{ij}, \Gamma_2 = \mathbf{U}_1^{ij} \mathbf{U}_2^{ij}, \dots$  of integral curves of the eigenvectors  $\mathbf{r}_k$  in the state space, connecting  $\mathbf{U}_i$  and  $\mathbf{U}_j$ , thus involving only two rarefaction or compression waves and a contact discontinuity. The intermediate states  $\mathbf{U}_1^{ij}, \mathbf{U}_2^{ij}, \dots$  can be computed by using the fact that the Riemann invariants are constant along the integral curves of  $\mathbf{r}_k$ .

In simple cases, we have a geometric interpretation (see G.R., Chapter 3, Example 2.5). First, in the linear scalar case,

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = 0,$$

if we set  $\mathbf{a} = (a, b)^T$ ,  $\mathbb{F}(u) = \mathbf{a}u = (au, bu)^T$ ,  $(\mathbb{F} \cdot \mathbf{n})(u) = \mathbf{a} \cdot \mathbf{n} u$ , then (4.33) gives

$$\Phi(u, v, \mathbf{n}) = f^+(u) + f^-(v), \quad (4.34)$$

where

$$f^+(u) = u \max(\mathbf{a} \cdot \mathbf{n}, 0), \quad f^-(u) = u \min(\mathbf{a} \cdot \mathbf{n}, 0).$$

On a given edge of the cell  $C_i$ , with outward normal  $\mathbf{n}$ , we get that

$$\begin{aligned} \text{if } \mathbf{a} \cdot \mathbf{n} > 0, & \text{ then } \Phi(u_i, u_j, \mathbf{n}) = f^+(u_i) = \mathbf{a} \cdot \mathbf{n} u_i, \\ \text{if } \mathbf{a} \cdot \mathbf{n} > 0, & \text{ then } \Phi(u_i, u_j, \mathbf{n}) = f^-(u_j) = \mathbf{a} \cdot \mathbf{n} u_j, \end{aligned}$$

which is the upwind flux applied to the function  $\mathbb{F} \cdot \mathbf{n}(u) = \mathbf{a} \cdot \mathbf{n} u$ .

Also, in the simple nonlinear scalar case  $f = g = \frac{u^2}{2}$ ,  $(\mathbb{F} \cdot \mathbf{n})'(u) = (n_1 + n_2)\frac{u}{2}$ , we can see that (4.34) holds, where

$$\begin{aligned} f^+(u) &= u \max((\mathbb{F} \cdot \mathbf{n})'(u), 0) = (\mathbb{F} \cdot \mathbf{n})(\max(u, 0)) \text{ if } (n_1 + n_2) > 0, \\ f^-(u) &= u \min((\mathbb{F} \cdot \mathbf{n})'(u), 0) = (\mathbb{F} \cdot \mathbf{n})(\min(u, 0)) \text{ if } (n_1 + n_2) < 0. \end{aligned}$$

More generally, we have similar results when  $f = g$  is strictly convex, since  $\mathbb{F} = f(\mathbf{e}_1 + \mathbf{e}_2)$ ,  $\mathbb{F} \cdot \mathbf{n} = f(n_1 + n_2)$ , and  $(\mathbb{F} \cdot \mathbf{n})'$  are monotone.

In the case of the gas dynamics equations, we have two intermediate states and, using the expressions for the Riemann invariants (see Sect. 2.3), we write that

$\mathbf{U}_1 = \mathbf{U}_L$  is connected to  $\mathbf{U}_2$ , where  $\mathbf{U}_2$  satisfies  $\mathbf{U}_L \cdot \mathbf{n} + \ell_L = \mathbf{u}_2 \cdot \mathbf{n} + \ell_2$ ,

$$\mathbf{u}_L \cdot \mathbf{n}^\perp = \mathbf{u}_2 \cdot \mathbf{n}^\perp, s_L = s_2,$$

$\mathbf{U}_2$  is connected to  $\mathbf{U}_3$  with  $\mathbf{u}_3 \cdot \mathbf{n} = \mathbf{u}_2 \cdot \mathbf{n}$ ,  $p_3 = p_2$ ,

$\mathbf{U}_3$  is connected to  $\mathbf{U}_4 = \mathbf{U}_R$  with  $\mathbf{u}_R \cdot \mathbf{n} - \ell_R = \mathbf{u}_3 \cdot \mathbf{n} - \ell_3$ ,  $\mathbf{u}_R \cdot \mathbf{n}^\perp = \mathbf{u}_3 \cdot \mathbf{n}^\perp$ ,  $s_R = s_3$ .

Note that the tangential velocity  $\mathbf{u} \cdot \mathbf{n}^\perp$  may jump across the contact discontinuity only. Then, the integral is split, so that

$$\begin{aligned} \Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) &= \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_i) + \int_{\Gamma} \mathbf{A}^-(\mathbf{w}, \mathbf{n}) d\mathbf{w} \\ &= \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_j) + \int_{\Gamma} \mathbf{A}^+(\mathbf{w}, \mathbf{n}) d\mathbf{w}, \end{aligned}$$

and one writes

$$\int_{\Gamma} \mathbf{A}^+(\mathbf{w}, \mathbf{n}) d\mathbf{w} = \sum_k \int_{\Gamma k} \mathbf{A}^+(\mathbf{w}, \mathbf{n}) d\mathbf{w}.$$

In fact, it is simpler to use (4.8) and write the flux for  $\mathbf{n} = \mathbf{e}_1, \mathbb{F} \cdot \mathbf{e}_1 = \mathbf{f}, \mathbf{A}(\mathbf{w}, \mathbf{e}_1) = \mathbf{A}(\mathbf{w})$  since the flux is invariant under rotation. Indeed, by definition (see (1.9), Chap. IV),

$$\mathbf{A}(\mathbf{U}, \mathbf{n})^\pm = (\mathbf{T} A^\pm \mathbf{T}^{-1})(\mathbf{U}, \mathbf{n}),$$

where  $\Lambda(\mathbf{U}, \mathbf{n})$  is the diagonal matrix of the eigenvalues of  $\mathbf{A}(\mathbf{U}, \mathbf{n})$ , which are

$$a_1(\mathbf{U}, \mathbf{n}) = \mathbf{u} \cdot \mathbf{n} - c, \quad a_2(\mathbf{U}, \mathbf{n}) = a_3(\mathbf{U}, \mathbf{n}) = \mathbf{u} \cdot \mathbf{n}, \quad a_4(\mathbf{U}, \mathbf{n}) = \mathbf{u} \cdot \mathbf{n} + c,$$

and  $\mathbf{T}(\mathbf{U}, \mathbf{n})$  is the matrix whose columns are the corresponding eigenvectors  $\mathbf{r}_k(\mathbf{U}, \mathbf{n})$  given by (4.32). From (4.9), if  $R$  is the rotation such that  $R^{-1}\mathbf{n} = \mathbf{e}_1$  and  $R^{-1}\mathbf{U} = \tilde{\mathbf{U}}$ , then

$$\mathbf{A}(\mathbf{U}, \mathbf{n})\mathbf{r}_k(\mathbf{U}, \mathbf{n}) = R\mathbf{A}(\tilde{\mathbf{U}})R^{-1}\mathbf{r}_k(\mathbf{U}, \mathbf{n}) = a_k(\mathbf{U}, \mathbf{n})\mathbf{r}_k(\mathbf{U}, \mathbf{n}),$$

and we obtain

$$\mathbf{A}(\tilde{\mathbf{U}})R^{-1}\mathbf{r}_k(\mathbf{U}, \mathbf{n}) = a_k(\mathbf{U}, \mathbf{n})R^{-1}\mathbf{r}_k(\mathbf{U}, \mathbf{n}),$$

and thus

$$a_k\tilde{\mathbf{U}} = a_k(\mathbf{U}, \mathbf{n}), \quad \mathbf{r}_k(\tilde{\mathbf{U}}) = R^{-1}\mathbf{r}_k(\mathbf{U}, \mathbf{n}),$$

which implies

$$\Lambda(\mathbf{U}, \mathbf{n}) = \Lambda(\tilde{\mathbf{U}}), \quad \mathbf{T}(\tilde{\mathbf{U}}) = R^{-1}\mathbf{T}(\mathbf{U}, \mathbf{n})$$

and

$$\mathbf{A}(\mathbf{U}, \mathbf{n})^\pm = R\mathbf{T}(\tilde{\mathbf{U}})\Lambda^\pm(\tilde{\mathbf{U}})\mathbf{T}^{-1}(\tilde{\mathbf{U}})R = R\mathbf{A}^\pm(\tilde{\mathbf{U}})R^{-1}, \quad (4.35)$$

i.e., for any vector  $\mathbf{w} \in \mathbb{R}^4$

$$\mathbf{A}(\mathbf{U}, \mathbf{n})^+ \mathbf{w} = R\mathbf{A}^+(\tilde{\mathbf{U}})R^{-1}\mathbf{w} = R\mathbf{A}^+(\tilde{\mathbf{U}})\tilde{\mathbf{w}}.$$

Thus

$$R\Phi(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{e}_1) = R\left\{ \mathbf{f}(\tilde{\mathbf{U}}_j) - \int_{\tilde{\Gamma}} \mathbf{A}^+(\mathbf{v})d\mathbf{v} \right\} = \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_j) - \int_{\tilde{\Gamma}} R\mathbf{A}^+(\mathbf{v})d\mathbf{v}.$$

Now, making a change of variables in the integral,  $\mathbf{v} = R^{-1}\mathbf{w}$ ,

$$\mathbf{A}(\mathbf{w}, \mathbf{n})^+ = R\mathbf{A}^+(\mathbf{v})R^{-1},$$

and noticing that the path  $\tilde{\Gamma}$ , consisting of portions  $\tilde{\Gamma}_1, \tilde{\Gamma}_2, \dots$  of integral curves of the eigenvectors  $\mathbf{r}_k$  in the state space, connecting  $\tilde{\mathbf{U}}_i$  and  $\tilde{\mathbf{U}}_j$  is transformed into the path  $\Gamma$  connecting  $\mathbf{U}_i$  and  $\mathbf{U}_j$ , we obtain

$$R\Phi(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{e}_1) = \mathbb{F} \cdot \mathbf{n}(\mathbf{U}_j) - \int_{\Gamma} \mathbf{A}^+(\mathbf{w}, \mathbf{n}) d\mathbf{w} = \Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}).$$

The remaining computations then follow exactly the lines of Sects. 3.4.2, 3.4.3 in Chap. IV. In particular, each integral  $\int_{\Gamma_k} \mathbf{A}^\pm(\mathbf{w}) d\mathbf{w}$  appears as a difference of fluxes computed at the points  $\mathbf{U}_L, \mathbf{U}_R, \mathbf{U}_2$ , and  $\mathbf{U}_3$ , and the numerical flux (4.33)  $\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n})$  can be written as the sum  $\sum(\varepsilon_i f(\mathbf{U}_i) + \bar{\varepsilon}_i f(\bar{\mathbf{U}}_i))$ , where  $\varepsilon$  takes the value 0 or  $\pm 1$  according to the sign of the eigenvalues (see (3.56) in Chap. IV, Sect. 3.4.2). The “sonic” states  $\bar{\mathbf{U}}_1, \bar{\mathbf{U}}_4$ , respectively, are such that

$$\begin{aligned}\lambda_1 &= \bar{u}_1 - c(u_1, s_L) = 0 & \text{and} & \quad \bar{u}_1 + \ell_1 = u_L + \ell_L, \quad \bar{v}_1 = v_L, \\ \lambda_4 &= \bar{u}_4 - c(u_4, s_R) = 0 & \text{and} & \quad \bar{u}_4 + \ell_4 = u_R + \ell_R, \quad \bar{v}_4 = v_R.\end{aligned}$$

One can similarly extend the method of Sect. 4.3 based on a shock-curve decomposition (for details, see Mehlman [860]).

### 4.3.5 Flux-Vector Splitting

We now extend Steger and Warming’s or van Leer’s flux splitting. The flux  $\mathbb{F} \cdot \boldsymbol{\omega}$  is split into

$$\mathbb{F} \cdot \boldsymbol{\omega} = \mathbf{F}_{\boldsymbol{\omega}}^+ + \mathbf{F}_{\boldsymbol{\omega}}^-,$$

in order that the split flux Jacobians  $(\mathbf{F}_{\boldsymbol{\omega}}^+)'$  (resp.  $(\mathbf{F}_{\boldsymbol{\omega}}^-)'$ ) have positive (resp. negative) eigenvalues. If the homogeneity property (see Chap. IV, Sect. 5)

$$\mathbb{F} \cdot \boldsymbol{\omega}(\mathbf{u}) = (\mathbb{F} \cdot \boldsymbol{\omega})'(\mathbf{u}) \mathbf{u} = \mathbf{A}_{\boldsymbol{\omega}}(\mathbf{u}) \mathbf{u},$$

where

$$\mathbf{A}_{\boldsymbol{\omega}}(\mathbf{u}) = (\mathbb{F} \cdot \boldsymbol{\omega})'(\mathbf{u}) = \mathbf{A}(\mathbf{u}, \boldsymbol{\omega}),$$

holds (which is easily verified for the ideal gas dynamics equations), following Steger and Warming, we can define

$$\mathbf{F}_{\boldsymbol{\omega}}^+(\mathbf{u}) = \mathbf{A}_{\boldsymbol{\omega}}^+(\mathbf{u}) \mathbf{u}, \quad \mathbf{F}_{\boldsymbol{\omega}}^-(\mathbf{u}) = \mathbf{A}_{\boldsymbol{\omega}}^-(\mathbf{u}) \mathbf{u}$$

and

$$\Phi(\mathbf{u}_i, \mathbf{u}_j, \boldsymbol{\omega}) = \mathbf{F}_{\boldsymbol{\omega}}^+(\mathbf{u}_i) + \mathbf{F}_{\boldsymbol{\omega}}^-(\mathbf{u}_j).$$

For the gas dynamics equations, as in the one-dimensional case, we can expand the flux  $\mathbb{F} \cdot \boldsymbol{\omega}(\mathbf{u})$  in “fictitious” fluxes  $\mathbf{F}_{\boldsymbol{\omega}i}$  associated to each distinct eigenvalue

$$\mathbb{F} \cdot \boldsymbol{\omega} = \sum \mathbf{F}_{\boldsymbol{\omega}i}$$

over  $i = 1, 2, 4$ , and then  $\mathbf{F}_{\boldsymbol{\omega}}^+(\mathbf{u})$  (resp.  $\mathbf{F}_{\boldsymbol{\omega}}^-(\mathbf{u})$ ) is the sum of the fluxes associated with the positive (resp. negative) eigenvalues (see Chap. IV, Sect. 5.2). We get

$$\begin{aligned}\mathbf{F}_{\omega i} &= \lambda_1 \frac{\rho}{2\gamma} \mathbf{r}_i, \quad i = 1, 4, \\ \mathbf{F}_{\omega 2} &= \lambda_2 \frac{\rho(\gamma - 1)}{\gamma} \mathbf{r}_2(\mathbf{U}, \boldsymbol{\omega}),\end{aligned}$$

where, with shorthand notations (see (4.32)),

$$\begin{aligned}\mathbf{r}_1(\mathbf{U}, \boldsymbol{\omega}) &= (1, \mathbf{u} - \boldsymbol{\omega}c, H - \mathbf{u} \cdot \boldsymbol{\omega}c)^T, \\ \mathbf{r}_4(\mathbf{U}, \boldsymbol{\omega}) &= (1, \mathbf{u} - \boldsymbol{\omega}c, H - \mathbf{u} \cdot \boldsymbol{\omega}c)^T, \\ \mathbf{r}_2(\mathbf{U}, \boldsymbol{\omega}) &= \left(1, \mathbf{u} - \frac{|\mathbf{u}|^2}{2}\right)^T.\end{aligned}$$

If the normal speed  $\mathbf{u} \cdot \boldsymbol{\omega}$  is supersonic, then

$$\begin{aligned}\mathbf{u} \cdot \boldsymbol{\omega} \geq c &\implies \mathbf{F}_{\boldsymbol{\omega}}^+(\mathbf{U}) = \mathbb{F} \cdot \boldsymbol{\omega}(\mathbf{U}), \quad \mathbf{F}_{\boldsymbol{\omega}}^-(\mathbf{U}) = 0, \\ \mathbf{u} \cdot \boldsymbol{\omega} \leq -c &\implies \mathbf{F}_{\boldsymbol{\omega}}^+(\mathbf{U}) = 0, \quad \mathbf{F}_{\boldsymbol{\omega}}^-(\mathbf{U}) = \mathbb{F} \cdot \boldsymbol{\omega}(\mathbf{U}).\end{aligned}$$

*Remark 4.8.* In the case of the gas dynamics equations, the flux of Steger and Warming is invariant under rotation. By (4.35)

$$\mathbf{A}_{\mathbf{n}}^{\pm}(\mathbf{U}) = R \mathbf{A}^{\pm}(\tilde{\mathbf{U}}) R^{-1},$$

i.e., for any vector  $\mathbf{V} \in \mathbb{R}^4$ ,

$$\mathbf{A}_{\mathbf{n}}^{\pm}(\mathbf{U}) \mathbf{V} = R \mathbf{A}^{\pm}(\tilde{\mathbf{U}}) R^{-1} \mathbf{V} = R \mathbf{A}^{\pm}(\tilde{\mathbf{U}}) \tilde{\mathbf{V}},$$

and

$$\mathbf{A}_{\mathbf{n}}^+(\mathbf{U}_i) \mathbf{U}_i + \mathbf{A}_{\mathbf{n}}^-(\mathbf{U}_j) \mathbf{U}_j = R \{ \mathbf{A}^+(\tilde{\mathbf{U}}_i) \tilde{\mathbf{U}}_i + \mathbf{A}^-(\tilde{\mathbf{U}}_j) \tilde{\mathbf{U}}_j \},$$

so that (4.8) holds,

$$\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}) = R \Phi(\tilde{\mathbf{U}}_i, \tilde{\mathbf{U}}_j, \mathbf{e}_1) = R(\mathbf{f}^+(\tilde{\mathbf{U}}_i) + (\mathbf{f}^-(\tilde{\mathbf{U}}_j)))$$

where  $\mathbf{f}^+$  is the one-dimensional split flux associated to  $\mathbf{f}$  ( $\mathbf{f}^+ = \mathbf{F}_{\mathbf{e}_1}^+$  in the case of Steger–Warming splitting). Again, this property yields shorter computations.  $\square$

In the case of van Leer's flux splitting (see Chap. IV, Sect. 5.3), we compute easily

$$\mathbf{f}^{\pm}(\mathbf{U}) = \left( f_1^{\pm}, f_1^{\pm} \frac{(\gamma - 1)u \pm 2c}{\gamma}, f_1^{\pm} v, f_1^{\pm} \frac{\{(\gamma - 1)u \pm 2c\}^2}{2(\gamma^2 - 1)} \right)^T,$$

where the mass flux  $f_1^\pm$  is given by (see Chap. IV, Sect. 5.3)

$$f_1^\pm = \pm \left( \frac{\rho}{4c} \right) (u \pm c)^2.$$

For  $\mathbf{f}^\pm(\tilde{\mathbf{U}})$ , we replace  $(u, v)$  by  $(u_n, u_\tau)$  (for details, we refer to Steger and Warming [1072], van Leer [1152], Liu and Vinokur [829], Fezoui and Steve [477], Fezoui et al. [478], Liou et al. [818]). Van Leer's flux cannot preserve a steady contact discontinuity (see Mulder et al. [878]) which is a drawback for the extension to the Navier–Stokes equations, since the shear layer may be interpreted as a layer of contact discontinuity. In that case, the Osher flux, which shares with van Leer's flux the property of being continuously differentiable (except at a stationary contact discontinuity (Larroutuou [738]); see also [434]), is to be preferred (see Koren [705, 707]). It is also an issue for multimaterial flows, for the computation of material interfaces since a scheme may produce an artificial mixing zone in which the computation of the thermodynamical variables (pressure, sound speed, temperature,  $\mathbf{E}$ ) is difficult to achieve correctly [15].

## 5 Second-Order Finite-Volume Schemes

### 5.1 MUSCL-Type Schemes

A “second-order” (in space) version can be obtained via a MUSCL (monotonic upstream schemes for conservation laws) approach introduced by van Leer [1158] that is now widely followed. It uses a piecewise linear reconstruction, instead of piecewise constant functions (Fezoui [476]; see [57, 479], Durlofsky et al. [452], Colella [327], Rostand and Stoufflet [992], Angrand and Lafon [54]) or piecewise constant functions in sub-cells (Perthame and Qiu [946]), [130], together with a limitation procedure (on conservative, characteristic, or physical variables according to the chosen approach, e.g., [804]). For the time discretization, Runge–Kutta schemes (Shu and Osher [1058], Mulder et al. [878], Durlofsky et al. [452], Lallemand [729], Angrand and Lafon [54]) allow a higher order of accuracy.

Let us sketch the main steps of the MUSCL method that was developed for one-dimensional systems in Chap. IV, Sect. 6.1 (for the scalar one-dimensional case, see G.R., Chapter 4, Section 3). In formula (4.3), the first-order flux  $\Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}_{ij})$ , where  $\mathbf{u}_i, \mathbf{u}_j$  are the constant values on each side of an edge  $\Gamma_{ij} = \Omega_i \cap \Omega_j$ , is now replaced by  $\Phi(\mathbf{u}_{ij}, \mathbf{u}_{ji}, \mathbf{n}_{ij})$ , where  $\mathbf{u}_{ij}, \mathbf{u}_{ji}$  are second-order approximations of the solution on each side of the edge  $e = \Gamma_{ij}$ . These second-order approximations are computed through the following prediction-correction steps:

- (i) prediction of the gradients  $\nabla \mathbf{u}_i = (\nabla_x \mathbf{u}_i, \nabla_y \mathbf{u}_i)$  in each cell (each  $\nabla_x \mathbf{u}_i$  has  $p$  components; we shall also use the notation  $\nabla \mathbf{u}_i = (\nabla u_i)$ , where  $u_i$  stands for some dependent variable and  $\nabla u_i = (\nabla_x u_i, \nabla_y u_i) \in \mathbb{R}^2$ );
- (ii) linear extrapolation to define values  $\mathbf{u}_{ij}, \mathbf{u}_{ji}$  on each side of the edge;
- (iii) limitation procedure.

For the gas dynamics equations, stability considerations, for instance, preservation of a maximum principle on the mass fractions (see Mehlmann [860], Larrouture [735]), the positivity of the pressure or temperature (see Perthame and Qiu [946]) ... may impose the choice of dependent variables whose gradient is computed and limited: conservative  $(\rho, \rho u, \rho v, \rho e)$ , primitive  $(\rho, u, v, p)$  or physical  $(\rho, u, v, T)$ , or entropy characteristic variables (see Mulder and van Leer [879], Arminjon et al. [58]). According to the scheme, the limitation procedure may thus be preceded by a computation of the components of the chosen vector and eventually followed by a return to the conservative variables. For instance, the velocity may be computed through  $u_i^n = (\rho u)_i^n / \rho_i^n, v_i^n = (\rho v)_i^n / \rho_i^n$ ; one computes  $(\rho, u, v)_i^{n+1}$  by a prediction-correction procedure and then sets  $(\rho u)_i^{n+1} = \rho_i^{n+1} u_i^{n+1}$  and so on.

*Remark 5.1.* In fact, in the van Leer-Hancock scheme (see Chap. IV, Sect. 6.1 and G.R., Chapter 4, Section 3.2), the second-order approximation in space is linked to a prediction of the variables at time  $t_n + \frac{\Delta t}{2}$ . In the two-dimensional case, the solution is first updated at time  $t_n + \frac{\Delta t}{2}$ ; first, one writes

$$\mathbf{u}_i^{n+1/2} \cong \mathbf{u}_i^n + \frac{\Delta t}{2} \left( \frac{\partial \mathbf{u}}{\partial t} \right)_i^n.$$

Then, one uses either the scheme

$$\left( \frac{\partial \mathbf{u}}{\partial t} \right)_i^n = \frac{1}{|\Omega_i|} \sum_j |\Gamma_{ij}| \Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}_{ij})$$

from Sect. 4.1.1, and thus

$$\mathbf{u}_i^{n+1/2} \cong \mathbf{u}_i^n - \frac{\Delta t}{2} |\Omega_i| \sum_j |\Gamma_{ij}| \Phi(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}_{ij}),$$

or else one can write (Dervieux and Vijayasundaram [410], Fezoui [476], Fezoui et al. [478]; see [57])

$$\left( \frac{\partial \mathbf{u}}{\partial t} \right)_i^n = -\{ \mathbf{A}(\mathbf{u}_i^n) \nabla_x \mathbf{u}_i + \mathbf{B}(\mathbf{u}_i^n) \nabla_y \mathbf{u}_i \}$$

which uses some prediction of the gradients.

Then, one defines  $\mathbf{u}_{ij}^{n+1/2}, \mathbf{u}_{ji}^{n+1/2}$  on each side of the edge by interpolation and limitation, as we shall describe more precisely in some examples.  $\square$

For simplicity, we skip this prediction step so that the scheme reads

$$|\Omega_i| \frac{(\mathbf{u}_i^{n+1} - \mathbf{u}_i^n)}{\Delta t} + \sum_{e \in \partial\Omega_i, e = \Gamma_{ij}} |e| \Phi(\mathbf{u}_{ij}, \mathbf{u}_{ji}, \mathbf{n}_e) = \mathbf{0}. \quad (5.1)$$

The resulting scheme is only first-order accurate in time.

*Example 5.1. Rectangular mesh.* (Example 4.4 revisited). We can define a linear function on  $\Omega_{j,k} = (x_{j-1/2}, x_{j+1/2}) \times (y_{k-1/2}, y_{k+1/2})$  for any component  $u$  of  $\mathbf{u}$ ,

$$u^n(x, y) = u_{j,k}^n + (x - x_j)\delta u_{j,k}^x + (y - y_k)\delta u_{j,k}^y, \quad (x, y) \in \Omega_{j,k}.$$

In the case of a rectangular mesh, the prediction of the gradients  $(\delta_{j,k}^x, \delta_{j,k}^y)$  is rather straightforward. We may define a centered approximation of the gradients as in the one-dimensional case (see G.R., Chapter 4, Section 3.1)

$$\hat{\delta}u_{i,j}^x = \frac{(u_{i+1,j} - u_{i-1,j})}{2\Delta x}, \quad \hat{\delta}u_{i,j}^y = \frac{(u_{i+1,j} - u_{i-1,j})}{2\Delta y},$$

where to simplify we have chosen a uniform grid and then limit component-wise (in  $x$  and  $y$ ), for instance,

$$\delta u_{i,j}^x = \text{minmod}\left(\hat{\delta}u_{i,j}^x, \frac{(u_{i+1,j} - u_{i,j})}{\Delta x}, \frac{(u_{i,j} - u_{i-1,j})}{\Delta x}\right)$$

(with the usual definition of the minmod function), and a similar formula for  $\delta_{i,j}^y$ . See also Jeng and Payne [653] for another limiting procedure and Coirier and Powell [324] for a related reconstruction.  $\square$

For a general unstructured mesh, step (i) is not so obvious. Let us give some examples of the prediction and limitation of the gradients.

### 5.1.1 MUSCL-Type Cell-Centered Schemes

We use the notations of Example 4.1. The control cell is a triangle  $T_i$ , and the center of the cell is the barycenter  $\mathbf{g}_i$  of the triangle  $T_i$ ;  $\mathbf{m}_{ij}$  denotes the middle of the edge  $\Gamma_{ij} = T_i \cap T_j$ . We present in each step some different approaches.

#### (i) Prediction of the gradients

Approach 1: One can predict the gradients from the discretization of Green's formula (Deconinck et al. [399, 400], Dubois and Michaux [439]; see also W.K. Anderson [41]),

$$|T| \nabla u_i \cong \int_T \nabla u \, d\mathbf{x} = \int_{\partial T} u \, \mathbf{n} d\sigma \cong \sum_{e \subset \partial T} |e| u_e \mathbf{n}_e,$$

and on an edge  $e = \Gamma_{ij}$  of  $\partial T$ , we take

$$u_e = \frac{u_i + u_j}{2}$$

or a convex combination of  $u_i, u_j$ ,

$$u_e = (1 - \theta)u_i + \theta u_j,$$

where, for instance,  $\theta = m_{ij}g_i/(m_{ij}g_i + m_{ij}g_j)$ , or even the barycentric coordinate of the intersection of  $\mathbf{g}_i, \mathbf{g}_j$  with  $e$ .

Approach 2: Given the value  $\mathbf{u}_i$  in  $T_i$  and the values  $\mathbf{u}_i, \mathbf{u}_k, \mathbf{u}_\ell$  in the neighboring cells  $T_j, T_k, T_\ell$ , we can use  $\mathbf{u}_j - \mathbf{u}_i$  in order to compute the gradient in the direction  $\mathbf{g}_i \mathbf{g}_j$ , which gives a system of three equations to determine the two components  $\nabla_x \mathbf{u}_i, \nabla_y \mathbf{u}_i$  solved by the least squares method. More precisely, let  $\mathbf{d}u_i = (d_x, d_y)$  be any vector in  $\mathbb{R}^2$ , and let  $\Delta_j^i$  denote the difference

$$\Delta_j^i = \Delta_j^i(d_x, d_y) \equiv u_j - u_i + \mathbf{d}u_i(\mathbf{g}_i - \mathbf{g}_j) = u_j - u_j^{\text{ext}}$$

between the mean value  $u_j$  in a neighboring mesh  $T_j$  and the extrapolated value

$$u_j^{\text{ext}} = u_i + \mathbf{d}u_i \cdot (\mathbf{g}_j - \mathbf{g}_i) = u_i + d_x(g_j - g_i)_x + d_y(g_j - g_i)_y$$

at the barycenter  $\mathbf{g}_j$  of  $T_j$ , taking for the gradient the value  $\mathbf{d}u_i = (d_x, d_y)$ .

Then, define a prediction of the slope  $\nabla u_i$  by solving the unconstrained (quadratic, convex) minimization problem (see Chevrier and Galley [300], Benharbit [107]).

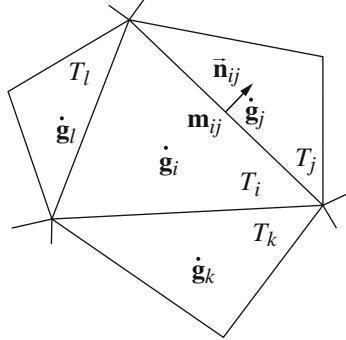
$$\nabla u_i \text{ is such that } \sum_j (\Delta_j^i)^2(d_x, d_y) \text{ is minimum,}$$

where the sum extends over all the indices  $j, k, \ell$  of the neighboring cells. The gradients  $\nabla_x u_i, \nabla_y u_i$ , minimizing  $\sum (\Delta_j^i)^2$ , are a solution of a  $2 \times 2$  linear system obtained by differentiating  $\sum (\Delta_j^i)^2$  w.r.t.  $(d_x, d_y)$  and setting to zero the two partial derivatives

$$\frac{\partial}{\partial d_x} \sum_j (\Delta_j^i)^2 = 0, \quad \frac{\partial}{\partial d_y} \sum_j (\Delta_j^i)^2 = 0.$$

As already noted, this is done separately for each component  $u$  of the vector  $\mathbf{u}$ .

Approach 3: One does not explicitly compute the gradient but uses linear interpolation from the values  $u_i$  at the barycenter  $\mathbf{g}_i$  of  $T_i$  and the values  $u_k, u_\ell$  at the barycenters  $\mathbf{g}_k, \mathbf{g}_\ell$  of two of the neighboring cells  $T_k, T_\ell$ , in order to compute the value at the middle of the edge between



**Fig. 5.1** Neighboring cells (approach 3)

$T_i$  and the third neighboring cell  $T_j$  (see Fig. 5.1). We get a piecewise linear function  $p_{ik\ell}$  in  $T_i$

$$p_{ik\ell} = u_i \psi_i + u_k \psi_k + u_\ell \psi_\ell,$$

where the  $\psi_k$  are the barycentric coordinates relative to the  $\mathbf{g}_k$  (see Lin et al. [804]).

Then, we have simply

$$(u_{ij})^{\text{pred}} = p_{ik\ell}(\mathbf{m}_{ij}) = u_i \Psi_i(\mathbf{m}_{ij}) + \mathbf{u}_k \psi_k(\mathbf{m}_{ij}) + \mathbf{u}_\ell \psi_\ell(\mathbf{m}_{ij}). \quad (5.2)$$

Approach 4: This approach uses one of the three preceding linear interpolants  $p_{ik\ell}, p_{ij\ell}, p_{ijk}$ , either the one that has a gradient with minimum norm or the one that has the greatest norm and that satisfies some limiting conditions as we shall discuss in (iii) (see Durlofsky et al. [452], X.-D. Liu [828]).

(ii) *Linear extrapolation*

In approaches 1 or 2, we define the value at the middle of an edge by

$$(u_{ij})^{\text{pred}} = u_i + \nabla u_i \cdot (\mathbf{m}_{ij} - \mathbf{g}_i) = u_i + \nabla_x u_i (\mathbf{m}_{ij} - g_i)_x + \nabla_y u_i (\mathbf{m}_{ij} - g_i)_y,$$

and similarly

$$(u_{ji})^{\text{pred}} = u_j + \nabla u_j \cdot (\mathbf{m}_{ij} - \mathbf{g}_j),$$

where  $\mathbf{m}_{ij} = (m_{ij\,x}, m_{ij\,y})$  denotes the coordinates of the middle of the edge  $\Gamma_{ij} = T_i \cap T_j$ .

It is known in the one-dimensional scalar case that the resulting scheme is not necessarily TVD, nor does it satisfy a discrete maximum principle.

(iii) *Limitation procedure*

Then  $\nabla \mathbf{u}_i$  is limited, which corresponds to correcting the preceding stage and setting

$$\mathbf{u}_{ij} = \mathbf{u}_i + \alpha_i \nabla \mathbf{u}_i \cdot (\mathbf{m}_{ij} - \mathbf{g}_i),$$

where the (diagonal matrix) limiter  $\alpha_i$  consists of a limiting factor  $\alpha_i$  for each component  $u$  of  $\mathbf{u}$ ,

$$u_{ij} = u_i + \alpha_i \nabla u_i \cdot (\mathbf{m}_{ij} - \mathbf{g}_i), \quad (5.3)$$

and  $\alpha$  depends on the chosen criteria. In particular, if  $u_i$  is already a local extremum relative to its neighbors, then usually  $\alpha_i = 0$ .

For instance, one can simply limit in such a way that the value  $u_{ij}$  at the middle of the edge belongs to the interval between  $u_i$  and  $u_j$ . In the case of Approach 4, the limiting procedure does select the gradient among the three gradients  $\nabla p_{ik\ell}, \nabla p_{ij\ell}, \nabla p_{ijk}$  as the one with greatest norm which meets this requirement (and if the gradient with second greatest norm does not satisfy this either, one takes the one with smallest norm).

Otherwise (see Mehlman [860] in the case of unstructured quadrilaterals, Dubois and Michaux [439], Dubois [436]) set

$$m_i = \min(u_j), \quad M_i = \max(u_j),$$

where the extrema are extended to the indices  $j$  of the neighboring cells,  $j \neq i$ . As already noted, if  $u_i$  is a local extremum, then  $\alpha_i = 0$ . If  $u_i$  belongs to  $[m_i, M_i]$  the procedure requires that the value  $u_{ij}$  at the middle of the edges should satisfy

$$K \max_{j, u_j \leq u_i} (u_j - u_i) \leq u_{ij} - u_i \leq K \min_{j, u_j \leq u_i} (u_j - u_i), \quad (5.4)$$

where  $0 \leq K \leq 1$ , and  $\alpha_i$  in (5.3) is the greatest possible value in  $[0, 1]$  for which (5.4) holds. These are two-dimensional versions of the one-dimensional limiter (3.9) or (3.11) in G.R., Chapter 4, Section 3.1, which limits the gradient and not each component in  $x$  and  $y$ .

In the case of Approach 2, another possibility considers a constrained minimization problem:  $\sum_j (\Delta_j^i)^2 (d_x, d_y)$  is minimized on the “limited” set  $(d_x, d_y)$  of vectors in  $\mathbb{R}^2$  satisfying

$$\begin{aligned} &\text{if } u_j - u_i \geq 0, \quad 0 \leq u_j^{\text{ext}} - u_i \leq u_j - u_i, \\ &\text{if } u_j - u_i \leq 0, \quad u_j - u_i \leq u_j^{\text{ext}} - u_i \leq 0. \end{aligned}$$

For details concerning the explicit computations, we refer to Buffard [212].

In the case of Approach 3, the limiting procedure defines the values at the middle of the edge by

$$u_{ij} - u_i = \text{minmod}((u_{ij})^{\text{pred}} - u_i, K(u_j - u_i)), \quad (5.5a)$$

$$u_{ji} - u_j = \text{minmod}((u_{ji})^{\text{pred}} - u_j, K(u_i - u_j)), \quad (5.5b)$$

where  $0 \leq K \leq \frac{1}{2}$ , and we shall see later that the resulting scheme satisfies the maximum principle.

*Remark 5.2.* Let us interpret these limitations in the one-dimensional case. The values on each side at the “middle of the edge” are simply the values

$$\begin{aligned} u_{i+1/2-} &= u_i + \frac{\delta_i}{2}, \\ u_{i+1/2+} &= u_{i+1} - \frac{\delta_{i+1}}{2}, \end{aligned}$$

(see G.R., Chapter 4, Sections 3.1, 3.2), where  $\delta_i$  is obtained from the prediction  $\hat{\delta}_i$  by some limiter, for instance, by the limiter minmod

$$\delta_i = \text{minmod}(|u_{i+1} - u_i|, \hat{\delta}_i, |u_i - u_{i-1}|).$$

Here, the function minmod is usually defined by (see G.R., Chapter 4, (2.28))

$$\text{minmod}(a_k) = \begin{cases} s \min(|a_k|) & \text{if the signs are identical, } s = \text{sgn}(a_k), \\ 0 & \text{otherwise.} \end{cases}$$

Since  $\delta_i = 2(u_{i+1/2-} - u_i)$ , we impose that the value  $u_{i+1/2-}$  satisfies

$$2(u_{i+1/2-} - u_i) = \begin{cases} \min(u_{i+1} - u_i, u_i - u_{i-1}) & \text{if both increments } > 0, \\ \max(u_i - u_{i+1}, u_{i-1} - u_i) & \text{if both increments } < 0, \\ 0 & \text{otherwise,} \end{cases}$$

which leads to the above formula (5.3) with  $K = \frac{1}{2}$  for limiting the 2d gradients.

Approach 3 corresponds in the one-dimensional case to predicting first  $u_{i+1/2,-}$  (resp.  $u_{i+1/2,+}$ ) by extrapolation from  $u_{i-1}, u_i$  (resp.  $u_{i+1}, u_{i+2}$ ) or equivalently to taking in G.R., formula (3.23), Chapter 4,

$$\begin{aligned} (u_{i+1/2,-})^{\text{pred}} - u_i &= \frac{u_i - u_{i-1}}{2} = \frac{\Delta u_{i-1/2}}{2}, \\ u_{i+1} - (u_{i+1/2,+})^{\text{pred}} &= \frac{u_{i+2} - u_{i+1}}{2} = \frac{\Delta u_{i+3/2}}{2}, \end{aligned}$$

and then limiting by setting

$$\begin{aligned} (u_{i+1/2-})^{\text{lim}} - u_i &= \text{minmod}((u_{i+1/2,-})^{\text{pred}} - u_i, K(u_{i+1} - u_i)) \\ &= \text{minmod}\left(\frac{\Delta u_{i-1/2}}{2}, K\Delta u_{i+1/2}\right), \end{aligned} \tag{5.6a}$$

$$\begin{aligned} u_{i+1} - (u_{i+1/2+})^{\text{lim}} &= \text{minmod}(u_{i+1} - (u_{i+1/2,+})^{\text{pred}}, K(u_{i+1} - u_i)) \\ &= \text{minmod}\left(\frac{\Delta u_{i+3/2}}{2}, K\Delta u_{i+1/2}\right) \end{aligned} \tag{5.6b}$$

for some constant  $K \leq 1$ . For  $K = \frac{1}{2}$ , we recover the minmod limiter (see G.R., Chapter 4, (3.8)) which ensures that the scheme is TVD and  $L^\infty$ -stable.

□

One can prove  $L^\infty$  stability for some of these two-dimensional schemes (see Benharbit [107] in the case of Approach 2, Lin et al. [804] for Approach 3, X.-D. Liu [828] for Approach 4). For instance, see the following lemma.

*Lemma 5.1*

(Approach 3) Let  $u_{ij}$ ,  $u_{ji}$  be defined by (5.2) and (5.5), and assume the following constraint on the triangulation:

$$\psi_q(\mathbf{m}_{ip}) \leq 0, \quad p, q \in \{j, k, \ell\}, \quad q \neq p.$$

Then, the scheme

$$|\Omega_i| \frac{(u_i^{n+1} - u_i^n)}{\Delta t} + \sum_{e \in \partial\Omega_j, e=\Gamma_{ij}} |e| \phi(u_{ij}, u_{ji}, \mathbf{n}_e) = 0,$$

where  $\phi$  is the numerical flux associated to a three-point monotone scheme by (4.6), satisfies

$$u_i^{n+1} = u_i^n + \{C_{ij}(u_j - u_i) + C_{ik}(u_k - u_i) + C_{i\ell}(u_\ell - u_i)\}$$

with nonnegative coefficients  $C_{ij}, C_{ik}, C_{i\ell}$ .

*Proof.* We consider first the one-dimensional case; we write

$$-\frac{(u_i^{n+1} - u_i^n)}{\lambda} = g_{i+1/2} - g_{i-1/2} = g_{i+1/2} - g(u_i, u_i) + g(u_i, u_i) - g_{i-1/2}$$

and

$$\begin{aligned} g_{i+1/2} - g(u_i, u_i) &= g(u_{i+1/2,-}, u_{i+1/2,+}) - g(u_i, u_i) \\ &= g(u_{i+1/2,-}, u_{i+1/2,+}) - g(u_i, u_{i+1/2,+}) + g(u_i, u_{i+1/2,+}) - g(u_i, u_i) \\ &= \frac{\partial g}{\partial u}(\xi_i)(u_{i+1/2,-} - u_i) + \frac{\partial g}{\partial v}(\eta_i)(u_{i+1/2,+} - u_i), \end{aligned}$$

where the derivatives are evaluated at some point and such that for a monotone scheme  $\frac{\partial g}{\partial u}(\xi_i) \geq 0$ ,  $\frac{\partial g}{\partial v}(\eta_i) \leq 0$ . We have the prediction

$$(u_{i+1/2,-})^{\text{pred}} - u_i = \frac{(u_i - u_{i-1})}{2} = \frac{\Delta u_{i-1/2}}{2},$$

which, by (5.6a), we limit in such a way that

$$u_{i+1/2,-} - u_i = -\alpha_i(u_{i-1} - u_i),$$

where  $\alpha_i = \min\left(\frac{1}{2}, \frac{k\Delta u_{i+1/2}}{\Delta u_{i-1/2}}\right)$  if  $\operatorname{sgn} \Delta u_{i-1/2} = \operatorname{sgn} \Delta u_{i+1/2}$ , and 0 otherwise. Also,

$$u_{i+1} - (u_{i+1/2,+})^{\text{pred}} = \frac{(u_{i+2} - u_{i+1})}{2} = \frac{\Delta u_{i+3/2}}{2},$$

and by (5.6b)

$$u_{i+1/2,+} - u_i = u_{i+1} - u_i - (u_{i+1} - u_{i+1/2,+}) = (1 - \beta_i)(u_{i+1} - u_i),$$

where  $\beta_i = \min(k, \frac{\Delta u_{i+3/2}}{2\Delta u_{i+1/2}})$  if  $\operatorname{sgn} \Delta u_{i+3/2} = \operatorname{sgn} \Delta u_{i+1/2}$ , and 0 otherwise.

Hence, we obtain

$$-\{g(u_{i+1/2,-}, u_{i+1/2,+}) - g(u_i, u_i)\} = c_i(u_{i-1} - u_i) + d_i(u_{i+1} - u_i)$$

with  $c_i = \alpha_i \frac{\partial g}{\partial u}(\xi_i) \geq 0$ ,  $d_i = -\frac{\partial g}{\partial v}(\eta_i)(1 - \beta_i) \geq 0$ .

Similarly,

$$\begin{aligned} g_{i-1/2} - g(u_i, u_i) &= \frac{\partial g}{\partial u}(\xi'_i)(u_{i-1/2,+} - u_i) + \frac{\partial g}{\partial v}(\eta'_i)(u_{i-1/2,-} - u_i) \\ &= c'_i(u_i - u_{i-1}) + d'_i(u_i - u_{i+1}), \end{aligned}$$

where

$$\begin{aligned} d'_i &= \alpha'_i \frac{\partial g}{\partial u}(\xi'_i) \geq 0, \\ \alpha'_i &= \min\left(\frac{1}{2}, \frac{K\Delta u_{i-1/2}}{\Delta u_{i+1/2}}\right) \text{ if } \operatorname{sgn} \Delta u_{i-1/2} = \operatorname{sgn} \Delta u_{i+1/2}, 0 \text{ otherwise,} \\ c'_i &= -\frac{\partial g}{\partial v}(\eta'_i)(1 - \beta'_i) \geq 0, \\ \beta'_i &= \min\left(K, \frac{\Delta u_{i-3/2}}{\Delta u_{i-1/2}}\right) \text{ if } \operatorname{sgn} \Delta u_{i-3/2} = \operatorname{sgn} \Delta u_{i-1/2}, 0 \text{ otherwise.} \end{aligned}$$

Thus, setting  $C_i = \lambda(c_i + c'_i)$ ,  $D_i = \lambda(d_i + d'_i)$ , we get

$$u_i^{n+1} - u_i^n = C_i(u_{i-1} - u_i) + D_i(u_{i+1} - u_i),$$

which can be written equivalently

$$u_i^{n+1} = u_i^n(1 - C_i - D_i)u_i + C_i u_{i-1} + D_i u_{i+1}, \quad \text{with } C_i \geq 0, D_i \geq 0.$$

Similarly, in the 2d case, using the consistency property (4.5) and (4.14), we can write

$$\begin{aligned}
\sum |e| \Phi(u_{ij}, u_{ji}, \mathbf{n}_e) &= \sum |e| \{ \Phi(u_{ij}, u_{ji}, \mathbf{n}_e) - \Phi(u_i, u_i, \mathbf{n}_e) \} \\
&= \sum |e| \{ \Phi(u_{ij}, u_{ji}, \mathbf{n}_e) - \Phi(u_i, u_{ji}, \mathbf{n}_e) \\
&\quad + \Phi(u_i, u_{ji}, \mathbf{n}_e) - \Phi(u_i, u_i, \mathbf{n}_e) \} \\
&= \sum |e| \frac{\partial \Phi}{\partial u}(\xi_i)(u_{ij} - u_i) + \sum |e| \frac{\partial \Phi}{\partial v}(\eta_i)(u_{ji} - u_j + u_j - u_i),
\end{aligned}$$

where  $\frac{\partial \Phi}{\partial u} \geq 0$ ,  $\frac{\partial \Phi}{\partial v} \leq 0$ . In the prediction step, we have

$$\begin{aligned}
(u_{ij})^{\text{pred}} - u_i &= u_i \psi_i(\mathbf{m}_{ij}) + u_k \psi_k(\mathbf{m}_{ij}) + u_\ell \psi_\ell(m_{ij}) - u_i \\
&= (u_k - u_i) \psi_k(\mathbf{m}_{ij}) + (u_\ell - u_i) \psi_\ell(\mathbf{m}_{ij}),
\end{aligned}$$

and we have assumed that the triangulation is such that the barycentric coordinates satisfy

$$\psi_k(\mathbf{m}_{ij}) \leq 0, \quad \psi_\ell(\mathbf{m}_{ij}) \leq 0.$$

We then use the formulas (5.5), and we conclude as in the one-dimensional case.  $\square$

*Remark 5.3.* The above decomposition of  $u_i^{n+1}$  is obviously to be compared with the incremental form of a numerical scheme (see Proposition 4.2 in Sect. 4.2.3), though the notations in the proof are slightly different. In the one-dimensional case, starting from a three-point monotone scheme, we know that there exist unique incremental coefficients (see (1.2d) in Chap. IV, Sect. 1)

$$\begin{aligned}
C_{i+1/2} &= C(u_i, u_{i+1}) = \lambda \frac{g(u_i, u_i) - g(u_i, u_{i+1})}{u_{i+1} - u_i} = -\lambda \frac{\partial g}{\partial v}(\zeta_i), \\
D_{i+1/2} &= D(u_i, u_{i+1}) = \lambda \frac{g(u_{i+1}, u_{i+1}) - g(u_i, u_{i+1})}{u_{i+1} - u_i} = \frac{\partial g}{\partial u}(v_i),
\end{aligned}$$

such that

$$u_i^{n+1} = u_i^n + C_{i+1/2} \Delta u_{i+1/2} - D_{i-1/2} \Delta u_{i-1/2},$$

and which, moreover, satisfy Harten's TVD criteria

$$C_{i+1/2} \geq 0, \quad D_{i+1/2} \geq 0, \quad C_{i+1/2} + D_{i+1/2} \leq 1.$$

However, when substituting  $u_{i+1/2,-}, u_{i+1/2,+}$  in the numerical flux  $g$ , the resulting scheme is five-point and is not written in the same incremental form. We can use the fact that it is “essentially three-point” to define incremental coefficients (given by G.R., formula (3.17) Chapter 3), but it is not straightforward that the coefficients satisfy Harten's criteria.  $\square$

Of course, other linear reconstructions and limitations are also possible.

### 5.1.2 MUSCL-Type Cell Vertex Schemes

We now come to Example 4.2

(i) *Prediction of the gradients  $\nabla u_i$*

Consider a triangle  $T$  with vertices  $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k$ . Given the values  $u_i, u_j, u_k$  of one component  $u$  of  $\mathbf{u}$  at the vertices, one usually takes the  $\mathbf{P}^1$ -interpolate (i.e., the linear function  $p_T$  on  $T$  which takes the same values), and  $p'_T$  yields a constant approximate value  $\nabla u_T$  of the gradient in  $T$ . More precisely, if the  $\psi_i$  are the  $\mathbf{P}^1$ -basis function (or barycentric coordinates),  $\psi_i(\mathbf{a}_j) = \delta_{ij}$ , then we take

$$\begin{aligned} p_T &= u_i \psi_i + u_j \psi_j + u_k \psi_k, \\ \nabla u_T &\equiv u_i \nabla \psi_i + u_j \nabla \psi_j + u_k \nabla \psi_k. \end{aligned}$$

For  $\nabla u_i$  we take a weighted average of the  $\nabla u_T$  for all the triangles  $T$  having  $\mathbf{a}_i$  as vertex. Hence (setting  $\nabla u = \nabla u_T$  on  $T$ ) we have

$$\nabla u_i = \frac{1}{|C_i|} \int_{C_i} \nabla u \, d\mathbf{x} = \frac{1}{|C_i|} \sum_{T/A_i \in T} \frac{|T|}{3} \nabla u_T.$$

(ii) *Extrapolation at the middle of the edge*

One simply sets

$$\begin{aligned} u_{ij} &= u_i + \frac{1}{2} \nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i), \\ u_{ji} &= u_j - \frac{1}{2} \nabla u_j \cdot (\mathbf{a}_j - \mathbf{a}_i). \end{aligned}$$

As earlier, the resulting scheme may generate oscillations, and one introduces some limitation.

(iii) *The limitation procedure*

The limitation can be achieved in different ways. By limiting the gradients directly and choosing the value of least modulus on all the triangles  $T$  with vertex  $\mathbf{a}_i$ : for each component, we define

$$\nabla_x u_i^{\ell im} = \min_{T/A_i \in T} \text{mod } \nabla_x u_T, \quad \nabla_y u_i^{\ell im} = \min_{T/\mathbf{a}_i \in T} \text{mod } \nabla_y u_T$$

( $\nabla_x, \nabla_y$  may be replaced by the derivatives in the direction of the local gradients and the orthogonal direction; see Arminjon et al. [58]). Then, the extrapolation at the middle of the edge yields

$$u_{ij}^{\ell im} = u_i + \frac{1}{2} \nabla u_i^{\ell im} \cdot (\mathbf{a}_j - \mathbf{a}_i),$$

$$u_{ji}^{\ell im} = u_j + \frac{1}{2} \nabla u_j^{\ell im} \cdot (\mathbf{a}_j - \mathbf{a}_i).$$

Otherwise using another limiter (van Leer I [1148]), Lallemand et al. [731], Fezoui et al. [478]),

$$\text{lim}^I(a, b) = \frac{(a+b)(ab + |ab| + \varepsilon)}{(a^2 + b^2 + 2\varepsilon)}$$

or (van Leer III [1150], Lallemand et al. [731])

$$\text{lim}^{III}(a, b) = \frac{(a+b+\varepsilon)(a+b)}{(a^2 + b^2 + \varepsilon^2)},$$

with  $\varepsilon > 0$  small enough, one can set

$$\begin{aligned} u_{ij}^{\ell im} &= u_i + \frac{1}{2} \text{lim}(2\nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i) - (u_j - u_i), u_j - u_i), \\ u_{ji}^{\ell im} &= u_i - \frac{1}{2} \text{lim}(2\nabla u_j \cdot (\mathbf{a}_j - \mathbf{a}_i) - (u_j - u_i), u_j - u_i). \end{aligned}$$

In the one-dimensional case, this limitation corresponds to replacing the increments  $\nabla u_{j+1/2}, \nabla u_{j-1/2}$ , which are involved in the averaging-limiting procedure (see the definition of  $\hat{\delta}_j$  and  $\delta_j$  in G.R., Chapter 4, Section 3.1), by  $2\nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i) - (u_j - u_i)$  and  $(u_j - u_i)$ .

Another procedure first defines gradients as follows (Billey et al. [151], Rostand and Stoufflet [992], Fezoui and Steve [477], Arminjon et al. [58]): one considers for a given edge  $\mathbf{a}_i\mathbf{a}_j$ , the “upstream” and “downstream” triangles  $T_{ij}$  and  $T_{ji}$  in which the line  $\mathbf{a}_i\mathbf{a}_j$  enters (i.e.,  $\mathbf{a}_i + \varepsilon(\mathbf{a}_i - \mathbf{a}_j) \in T_{ij}$  for small  $\varepsilon > 0$  small enough (see Fig. 5.2)). Then, one sets

$$\nabla u_i = \nabla u_{T_{ij}}, \quad \nabla u_j = \nabla u_{T_{ji}}$$

and

$$u_{ij} = u_i + \frac{1}{2} \nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i),$$

$$u_{ji} = u_i + \frac{1}{2} \nabla u_j \cdot (\mathbf{a}_j - \mathbf{a}_i),$$

or a variant

$$u_{ij} = u_i + \frac{1}{4} \{ \nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i) + u_j - u_i \},$$

$$u_{ji} = u_j - \frac{1}{4} \{ \nabla u_j \cdot (\mathbf{a}_j - \mathbf{a}_i) + u_j - u_i \},$$

or one can even introduce an “upwinding” parameter  $k$  (Billey et al. [151], Rostand and Stouflet [992])

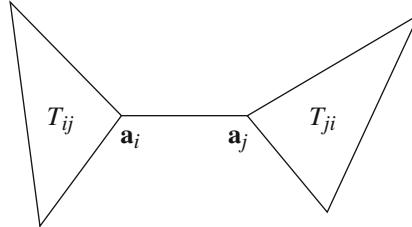
$$u_{ij} = u_i + \frac{1}{4} \{ (1-k) \nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i) + (1+k)(u_j - u_i) \},$$

$$u_{ji} = u_j - \frac{1}{4} \{ (1-k) \nabla u_j \cdot (\mathbf{a}_j - \mathbf{a}_i) + (1+k)(u_j - u_i) \}.$$

This is followed by a limitation procedure; for instance,

$$u_{ij}^{\ell im} = u_i + \frac{1}{2} \ell(\nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i), u_j - u_i),$$

where  $\ell(a, b)$  is some limiter, for instance,



**Fig. 5.2** “Upstream” and “downstream” triangles

$$\ell(a, b) = \frac{a(b^2 + \varepsilon) + b(a^2 + \varepsilon)}{a^2 + b^2 + 2\varepsilon} \text{ if } \operatorname{sgn} a = \operatorname{sgn} b, 0 \text{ otherwise.}$$

(see Dervieux and Vijayasundaram [410], Ciccoli et al. [308], Arminjon et al. [58], and Radespiel and Kroll [964], who precise the choice of  $\varepsilon$  for the van Albada limiter function). Or one can introduce a limiting factor and set

$$u_{ij}^{\ell im} = u_i + \frac{1}{4} s_{ij} \{ (u_j - u_i) + \nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i) \},$$

$$u_{ji}^{\ell im} = u_j - \frac{1}{4} s_{ji} \{ (u_j - u_i) + \nabla u_j \cdot (\mathbf{a}_j - \mathbf{a}_i) \},$$

where

$$s_{ij} = s(\nabla u_i \cdot (\mathbf{a}_j - \mathbf{a}_i), u_j - u_i),$$

and  $s(a, b)$  is van Albada's limiter (see Mulder and van Leer [879], Jorgenson and Turkel [671])

$$s(a, b) = \frac{2ab + \varepsilon^2}{(a^2 + b^2 + \varepsilon^2)}.$$

*Remark 5.4.* In the one-dimensional case, following the notations of G.R., Chapter 4, Remark 3.2, we had set

$$\delta_j = R(\theta_j)\hat{\delta}_j,$$

where  $R$  is a function of the ratio of consecutive increments  $\theta_j = \frac{\Delta u_{j+1/2}}{\Delta u_{j-1/2}}$  (and  $r_j = \frac{1}{\theta_j}$ ). To get the analogous formula in the 2d case, we can set  $a = \Delta u_{j-1/2}$ ,  $b = \Delta u_{j+1/2}$  and take a function such that  $R(\theta) = R(\frac{1}{\theta})$  and then set  $s(a, b) = R(\frac{a}{b})$ . For instance, van Albada's limiter  $R(\theta) = \frac{2\theta}{(1+\theta^2)}$  leads to the above formula for  $s$ . Note that in terms of flux limiters  $\varphi$  as defined in G.R., Chapter 4, Section 2, we have the relation  $\varphi(r) = \frac{R(\theta)(1+\theta)}{2\theta}$  with  $\theta = \frac{1}{r}$ ; then requiring that  $R$  satisfy  $R(\theta) = R(\frac{1}{\theta})$  is equivalent to the “symmetry property”  $\varphi(\frac{1}{r}) = \frac{\varphi(r)}{r}$  for  $\varphi$  (see again G.R., Chapter 4, Remark 3.2). The particular van Albada's limiter then gives  $\varphi(r) = \frac{r^2+r}{r^2+1}$ .  $\square$

Finally, one can also follow Davis' approach (see Arminjon et al. [58], Arminjon and Dervieux [56]).

Again, the limiting procedure is applied on each dependent variable, conservative, primitive  $\mathbf{W} = (\rho, u, v, p)$  (Fezoui et al. [478]), or characteristic. In this last case, the characteristic variables are usually obtained by linearizing at the midpoint (see Chap. II, Remark 2.2), i.e., the limited variable is  $T_{ij}^{-1}\mathbf{W}$ , where the transformation matrix  $\mathbf{T}$  that diagonalizes the Jacobian matrix is taken at the midvalue  $\frac{(\mathbf{W}_i + \mathbf{W}_j)}{2}$  (see Fezoui and Steve [477], Ro-stand and Stoufflet [992], Arminjon et al. [58], Mulder and van Leer [879]); see also MUSCL techniques.

## 5.2 Other Approaches

In fact, we shall only mention some of the most usual procedures.

Lax–Wendroff-type schemes were the first developed. These are usually centered predictor–corrector schemes on structured mesh such as the  $S_\alpha^\beta$ -schemes (Lerat and Sides [764]) or combined with a finite-element procedure such as the Richtmyer–Galerkin two-step schemes (Angrand et al. [53], Billey et al. [151], Arminjon and Dervieux [56]). Since their derivation is not so straightforward and not really linked to what precedes, we shall not detail it. Instead, we refer to the abovementioned papers for details.

One can also use an ENO scheme. These rely on a reconstruction procedure that gives a high-order accurate representation of the solution from given cell averages (see G.R., Chapter 4, Section 3.6) and usually involve Runge–Kutta methods for time integration. The reconstruction procedure is more easily achieved on structured mesh (Casper and Atkins [239]) but can also be on unstructured mesh ([258], Harten and Chakravarthy, Shu and Osher [1059] [1059], Shu et al. [1060], Abgrall [6], Angrand and Lafon [54]; see also [310]).

We can also mention different approaches; see Sanders and Li [1008], LeVeque [773], Beam and Warming [95], Cockburn et al. [319], Berde and Borrel [113] and Hansbo [583] and for schemes based on switching strategies Harabetian and Pego [588].

## 6 An Introduction to All-Mach Schemes for the System of Gas Dynamics

In this section, we consider again the system of gas dynamics, and we analyze the behavior of some numerical schemes as the Mach number tends to zero (low Mach limit). Roughly speaking, the Mach number of the fluid is the ratio of a typical (reference) fluid velocity to a typical (reference) sound speed. It is known that the singular limit gives the incompressible fluid equations. In particular, the limit velocity field  $\mathbf{u}$  is shown to satisfy the incompressible constraint  $\nabla \cdot \mathbf{u} = 0$ .

We first derive the dimensionless PDE system obtained by introducing reference quantities. Assuming asymptotic expansions in power of the reference Mach number, we will perform a *formal* asymptotic analysis to get the limit system (see [697], for rigorous proofs; see also [869, 1018] and [1212], also [1020], and references therein). To this end, we will need to introduce boundary conditions, a subject which we have not yet treated, for which we refer to the next chapter. Indeed we will consider some simple examples of boundary conditions that can be easily understood without much background.

It has been early noticed that standard upwind schemes do not behave correctly as the Mach number tends to zero (see [566, 567]). The issue is to design schemes for the Euler system which, as the Mach number goes to zero, give a scheme which is consistent with the limit system and provide an accurate approximation (on a fixed mesh). This property is referred to as *asymptotic preserving* (AP) and will be analyzed in other situations in the chapter concerning source terms.

We will perform the same formal asymptotic analysis on the semi-discrete (space discretization) scheme, focusing on the Roe and HLL solvers which fail to provide an accurate approximation of the limit system unless some correction is brought. We introduce a particular correction, for which the AP property can be proven.

## 6.1 The Low Mach Limit of the System of Gas Dynamics

Consider again the Euler system of gas dynamics in any space dimension that can be written

$$\left\{ \begin{array}{l} \frac{\partial \varrho}{\partial t} + \nabla \cdot (\varrho \mathbf{u}) = 0, \\ \frac{\partial(\varrho \mathbf{u})}{\partial t} + \nabla \cdot (\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla p = \mathbf{0}, \\ \frac{\partial(\varrho e)}{\partial t} + \nabla \cdot ((\varrho e + p)\mathbf{u}) = 0. \end{array} \right. \quad (6.1)$$

(see (2.1) in Example 2.1, in the Chap. I) where as usual  $\varrho$  denotes the mass density,  $\mathbf{u}$  the velocity,  $e$  the specific total energy, and  $p$  the pressure. We supplement (6.1) with a real gas equation of state

$$p = (\gamma - 1)\varrho\varepsilon + c_{\text{ref}}^2(\varrho - \varrho_{\text{ref}}),$$

which generalizes the usual ideal gas law (of Grüneisen type) or more generally

$$p = p(\varrho, \tilde{\varepsilon}), \quad \tilde{\varepsilon} = \varrho\varepsilon. \quad (6.2)$$

Let us now write the system (6.1) using dimensionless variables. We introduce the following quantities: a reference or characteristic length  $\hat{x}$ , a characteristic velocity  $\hat{u}$ , a characteristic density  $\hat{\varrho}$ , and a characteristic sound speed  $\hat{c}$ . Then we define a characteristic time  $\hat{t} = \frac{\hat{x}}{\hat{u}}$ , a characteristic internal energy  $\hat{\varepsilon} = \hat{c}^2$ , and a characteristic pressure  $\hat{p} = \hat{\varrho}\hat{c}^2$ . Setting  $\varphi = \hat{\varphi}\varphi'$  for any quantity  $\mathbf{x}, t, \varrho, \mathbf{u}, p, \varepsilon$ , the system (6.1) becomes

$$\left\{ \begin{array}{l} \frac{\partial \varrho'}{\partial t'} + \nabla' \cdot (\varrho' \mathbf{u}') = 0, \\ \frac{\partial(\varrho' \mathbf{u}')}{\partial t'} + \nabla' \cdot (\varrho' \mathbf{u}' \otimes \mathbf{u}') + \frac{\hat{c}^2}{\hat{u}^2} \nabla' p' = \mathbf{0}, \\ \frac{\partial(\varrho' \varepsilon')}{\partial t'} + \nabla' \cdot ((\varrho' \varepsilon' + p') \mathbf{u}') \\ \quad + \frac{\hat{u}^2}{2\hat{c}^2} \left( \frac{\partial(\varrho' |\mathbf{u}'|^2)}{\partial t'} + \nabla' \cdot (\varrho' |\mathbf{u}'|^2 \mathbf{u}') \right) = 0 \end{array} \right.$$

where  $\nabla'$  denotes the gradient operator with respect to the  $\mathbf{x}'$  variables. By introducing the Mach number

$$M = \frac{|\hat{u}|}{\hat{c}} \quad (6.3)$$

and suppressing the primes for simplicity, the nondimensional system of gas dynamics reads

$$\left\{ \begin{array}{l} \frac{\partial \varrho}{\partial t} + \nabla \cdot (\varrho \mathbf{u}) = 0, \\ \frac{\partial(\varrho \mathbf{u})}{\partial t} + \nabla \cdot (\varrho \mathbf{u} \otimes \mathbf{u}) + \frac{1}{M^2} \nabla p = \mathbf{0}, \\ \frac{\partial(\varrho \varepsilon)}{\partial t} + \nabla \cdot ((\varrho \varepsilon + p) \mathbf{u}) \\ \quad + \frac{M^2}{2} \left( \frac{\partial(\varrho |\mathbf{u}|^2)}{\partial t} + \nabla \cdot (\varrho |\mathbf{u}|^2 \mathbf{u}) \right) = 0. \end{array} \right. \quad (6.4)$$

Now, assuming that the solutions of (6.4) are smooth enough (which is indeed the case in the applications when  $M$  is sufficiently small), the momentum and energy conservation equations can be written, respectively,

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{1}{M^2 \varrho} \nabla p &= \mathbf{0}, \\ \frac{\partial(\varrho \varepsilon)}{\partial t} + \mathbf{u} \cdot \nabla(\varrho \varepsilon) + (p + \varrho \varepsilon) \nabla \cdot \mathbf{u} &= 0. \end{aligned}$$

Then, setting as usual (cf. Chap. IV, Lemma 4.3)

$$\kappa = \frac{\partial p(\varrho, \tilde{\varepsilon})}{\partial \tilde{\varepsilon}}, \quad \chi = \frac{\partial p(\varrho, \tilde{\varepsilon})}{\partial \varrho}$$

and observing that

$$\chi + \kappa h = c^2$$

where  $h$  denotes the specific enthalpy (recall that  $h = \varepsilon + p/\varrho$ ), we have

$$\begin{aligned} \frac{\partial p}{\partial t} + \mathbf{u} \cdot \nabla p &= \chi \left( \frac{\partial \varrho}{\partial t} + \mathbf{u} \cdot \nabla \varrho \right) + \kappa \left( \frac{\partial(\varrho \varepsilon)}{\partial t} + \mathbf{u} \cdot \nabla(\varrho \varepsilon) \right) \\ &= -\varrho(\chi + \kappa h) \nabla \cdot \mathbf{u} = -\varrho c^2 \nabla \cdot \mathbf{u}. \end{aligned}$$

Hence, we obtain the familiar nonconservative form of the nondimensional system of gas dynamics

$$\left\{ \begin{array}{l} \frac{\partial \varrho}{\partial t} + \nabla \cdot (\varrho \mathbf{u}) = 0, \\ \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \frac{1}{M^2 \varrho} \nabla p = \mathbf{0}, \\ \frac{\partial p}{\partial t} + \mathbf{u} \cdot \nabla p + \varrho c^2 \nabla \cdot \mathbf{u} = 0. \end{array} \right. \quad (6.5)$$

Again, we supplement the system (6.5) with a general equation of state of the form (6.2). Note that, in the particular case of an ideal polytropic gas for which  $\chi = 0$ ,  $\kappa = \gamma - 1$ , the last equation (6.5) reads also

$$\frac{\partial p}{\partial t} + \mathbf{u} \cdot \nabla p + \gamma p \nabla \cdot \mathbf{u} = 0.$$

In order to study the asymptotic behavior of the solutions of (6.5) as  $M$  tends to zero, we postulate *formal* expansions of  $\varrho$ ,  $\mathbf{u}$  and  $p$  in powers of  $M$

$$\begin{cases} \varrho = \varrho^{(0)} + M\varrho^{(1)} + \dots, \\ \mathbf{u} = \mathbf{u}^{(0)} + M\mathbf{u}^{(1)} + \dots, \\ p = p^{(0)} + Mp^{(1)} + M^2p^{(2)} \dots. \end{cases} \quad (6.6)$$

Then the second equation (6.5) gives at the orders  $-2$  and  $-1$  in  $M$

$$\nabla p^{(0)} = \mathbf{0}, \quad \nabla p^{(1)} = \mathbf{0}; \quad (6.7)$$

the pressure is constant in space (up to fluctuations of order 2), and therefore the system (6.5) yields at the order 0 in  $M$

$$\begin{cases} \frac{\partial \varrho^{(0)}}{\partial t} + \nabla \cdot (\varrho^{(0)} \mathbf{u}^{(0)}) = 0, \\ \frac{\partial \mathbf{u}^{(0)}}{\partial t} + (\mathbf{u}^{(0)} \cdot \nabla) \mathbf{u}^{(0)} + \frac{1}{\varrho^{(0)}} \nabla p^{(2)} = \mathbf{0}, \\ \frac{\partial p^{(0)}}{\partial t} + \varrho^{(0)} (c^{(0)})^2 \nabla \cdot \mathbf{u}^{(0)} = 0. \end{cases} \quad (6.8)$$

We need to study the system (6.8) in a bounded spatial domain  $\mathcal{O}$ . Assuming  $\varrho^{(0)}(c^{(0)})^2 > 0$  in  $\mathcal{O}$ , the third equation (6.8) yields

$$\nabla \cdot \mathbf{u}^{(0)} = -\frac{1}{\varrho^{(0)}(c^{(0)})^2} \frac{dp^{(0)}}{dt},$$

and since  $\frac{dp^{(0)}}{dt}$  does not depend on  $\mathbf{x}$ , we get

$$-\left(\int_{\mathcal{O}} \frac{1}{\varrho^{(0)}(c^{(0)})^2} d\mathbf{x}\right) \frac{dp^{(0)}}{dt} = \int_{\mathcal{O}} \nabla \cdot \mathbf{u}^{(0)}(\mathbf{x}) d\mathbf{x}.$$

We then consider several types of boundary conditions:

(i) a slip boundary condition

$$\mathbf{u} \cdot \mathbf{n} = 0 \text{ on } \partial\mathcal{O} \quad (6.9)$$

where  $\mathbf{n}$  stands as usual for the unit outward normal vector to the boundary  $\partial\mathcal{O}$  of  $\mathcal{O}$ ;

(ii) a periodicity condition;

(iii)  $\partial\mathcal{O}$  is an open boundary where a constant pressure is applied.

In both cases (i) and (ii), we have

$$\int_{\mathcal{O}} \nabla \cdot \mathbf{u}^{(0)} d\mathbf{x} = \int_{\mathcal{O}} \mathbf{u}^{(0)} \cdot \mathbf{n} dS = 0,$$

so that

$$\frac{dp^{(0)}}{dt} = 0, \quad (6.10)$$

and  $p^{(0)}$  is constant in time too. Besides (6.10) holds trivially in case (iii).

Hence, for each boundary condition of types (i), (ii), or (iii), we obtain

$$\nabla \cdot \mathbf{u}^{(0)} = 0,$$

and the *formal* limit of the system (6.4) as  $M$  tends to zero is the system of incompressible gas dynamics

$$\begin{cases} \frac{\partial \varrho^{(0)}}{\partial t} + \nabla \cdot (\varrho^{(0)} \mathbf{u}^{(0)}) = 0, \\ \frac{\partial \mathbf{u}^{(0)}}{\partial t} + (\mathbf{u}^{(0)} \cdot \nabla) \mathbf{u}^{(0)} + \frac{1}{\varrho^{(0)}} \nabla p^{(2)} = \mathbf{0}, \\ \nabla \cdot \mathbf{u}^{(0)} = 0. \end{cases} \quad (6.11)$$

In the sequel of this section, we will analyze the behavior of various semi-discrete Godunov-type methods in several space dimensions as the Mach number tends to zero. For the sake of simplicity, we will restrict ourselves to the two-dimensional case, but the results can be easily extended to the three-dimensional case. We then use the notations of Sect. 2: we write the system of gas dynamics in the form

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbb{F}(\mathbf{U}) = \mathbf{0}, \quad \mathbb{F} = (\mathbf{f}, \mathbf{g})^T$$

where  $\mathbf{U}$ ,  $\mathbf{f}(\mathbf{U})$  and  $\mathbf{g}(\mathbf{U})$  are defined as in (2.2a) (2.2b).

We introduce a numerical flux function  $\Phi(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n})$  which satisfies as usual

$$\Phi(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) + \Phi(\mathbf{U}_R, \mathbf{U}_L, -\mathbf{n}) = \mathbf{0}. \quad (6.12)$$

Assume on the other hand that  $\mathcal{O}$  is a polygonal domain or approximated by such a domain. As in Sect. 4, we are given a partition of  $\mathcal{O}$  with polygonal cells  $\Omega_i, i \in I$ . We denote by  $\Gamma_{ij}$  the common edge of two adjacent cells  $\Omega_i$  and  $\Omega_j$ , by  $\mathbf{n}_{ij}$  the unit normal vector to  $\Gamma_{ij}$  oriented from  $\Omega_i$  to  $\Omega_j$  and by  $\mathbf{t}_{ij} = \mathbf{n}_{ij}^\perp$  the associated unit tangent vector. Observe that

$$\sum_{\Gamma_{ij} \subset \partial \Omega_i} |\Gamma_{ij}| \mathbf{n}_{ij} = \mathbf{0}. \quad (6.13)$$

We then consider the following semi-discrete finite-volume method which amounts to look for functions  $t \mapsto \mathbf{U}_i(t)$ ,  $i \in I$  solutions for all  $i \in I$  of

$$\frac{d\mathbf{U}_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}_{ij}) = \mathbf{0}. \quad (6.14)$$

In order to obtain a closed problem, we need to supplement (6.14) with boundary conditions. Indeed, with any boundary cell  $\Omega_i$ , we associate an external fictitious cell  $\Omega_j$  and we assume that  $\mathbf{U}_j$  satisfies

$$\Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}_{ij}) + \Phi(\mathbf{U}_j, \mathbf{U}_i, -\mathbf{n}_{ij}) = \mathbf{0}. \quad (6.15)$$

This condition (6.15) clearly holds in the case of periodic boundary conditions. We will check later on that it also holds in the case of a slip boundary condition (6.9) provided we impose the conditions

$$\mathbf{u}_j \cdot \mathbf{n}_{ij} = -\mathbf{u}_i \cdot \mathbf{n}_{ij}, \quad \mathbf{u}_j \cdot \mathbf{t}_{ij} = \mathbf{u}_i \cdot \mathbf{t}_{ij} \quad (6.16)$$

and

$$\varrho_j = \varrho_i, \quad \varepsilon_j = \varepsilon_i. \quad (6.17)$$

Note that the conditions (6.16) (6.17) imply  $e_j = e_i$  and  $p_j = p_i$ . Observe that (6.16) is a “natural” discretization of (6.9) while (6.17) are “artificial” boundary conditions. As a consequence of (6.12) and (6.15), we obtain the conservation property of the finite-volume method

$$\frac{d}{dt} \left( \sum_{i \in I} |\Omega_i| \mathbf{U}_i \right) = \mathbf{0}.$$

In the sequel of this section, we will consider popular finite-volume methods of Godunov-type, namely, the Roe and the HLL methods. Using a *formal* asymptotic analysis, we will show that these finite-volume methods do not behave correctly at low Mach, i.e., are not asymptotic preserving as  $M$  tends to zero. Then, we will modify them in order to obtain asymptotic preserving methods or, in other words, all-Mach schemes.

## 6.2 Asymptotic Analysis of the Semi-Discrete Roe Scheme

Let us first write the semi-discrete Roe scheme for the projected Euler equations of gas dynamics (thus in one space dimension)

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{U}) = \mathbf{0} \quad (6.18)$$

where again  $\mathbf{U}$  and  $\mathbf{f}(\mathbf{U})$  (vectors of  $\mathbb{R}^4$ ) are defined as in (2.2a). As in Sect. 4.2, we supplement the system (6.18) with a real gas equation of

state (6.2). The Jacobian matrix  $\mathbf{A}(\mathbf{U})$  of the flux function  $\mathbf{f}(\mathbf{U})$  is given by (cf. Sect. 4.3.3)

$$\mathbf{A}(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ K - u^2 & (2 - \kappa)u & -\kappa v & \kappa \\ -uv & v & u & 0 \\ u(K - H) & H - \kappa u^2 & -\kappa uv & (1 + \kappa)v \end{pmatrix}$$

where

$$H = h + \frac{1}{2}(u^2 + v^2), \quad K = \frac{\kappa}{2}(u^2 + v^2).$$

Note that  $\mathbf{A}(\mathbf{U}) = \mathbf{A}(u, v, H, \chi, \kappa)$ , i.e., depends only on  $(u, v, H, \chi, \kappa)$ .

Now, given a pair of states  $(\mathbf{U}_L, \mathbf{U}_R)$ , we define as in (Chap. IV, Sect. 4.2) a Roe matrix

$$\mathbf{A}(\mathbf{U}_L, \mathbf{U}_R) = \mathbf{A}(\bar{u}, \bar{v}, \bar{H}, \bar{\chi}, \bar{\kappa}). \quad (6.19)$$

In (6.19) we have

$$\bar{u} = m(u), \bar{v} = m(v), \bar{H} = m(H)$$

where  $m$  is the averaging operator (cf. Chap. IV, Eq. (4.23))

$$m(a) = m_\varrho(a) = \frac{\sqrt{\varrho_L}a_L + \sqrt{\varrho_R}a_R}{\sqrt{\varrho_L} + \sqrt{\varrho_R}}$$

and the average values  $\bar{\chi}$  and  $\bar{\kappa}$  must satisfy (cf. Chap. IV, Eq. (4.36))

$$\Delta p = \bar{\chi}\Delta\varrho + \bar{\kappa}\Delta\tilde{\varepsilon}.$$

Let us set  $\mathbf{u} = (u, v)^T$ ,  $\mathbf{n} = (1, 0)^T$ ,  $\mathbf{t} = (0, 1)^T$ . Arguing as in Chap. IV, Sect. 4.2, one can easily check that

$$\Delta\mathbf{f} = \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)\Delta\mathbf{U}$$

and moreover the numerical flux of the Roe method

$$\Phi(\mathbf{U}_L, \mathbf{U}_R) = \frac{1}{2} \{ \mathbf{f}(\mathbf{U}_L) + \mathbf{f}(\mathbf{U}_R) - \mathbf{A}(\mathbf{U}_L, \mathbf{U}_R)\Delta\mathbf{U} \}$$

is easily seen to be given by

$$\left\{ \begin{array}{l} \Phi(\mathbf{U}_L, \mathbf{U}_R) = \frac{1}{2} \left\{ \mathbf{f}(\mathbf{U}_L) + \mathbf{f}(\mathbf{U}_R) - |\bar{u}| \Delta \varrho \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} - \frac{\bar{c}^2}{\bar{\kappa}} \end{pmatrix} \right. \\ - \frac{\bar{\rho}}{2\bar{c}} (|\bar{u} + \bar{c}| - |\bar{u} - \bar{c}|) \Delta u \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} \end{pmatrix} \\ - \frac{\bar{\varrho}}{2} (|\bar{u} - \bar{c}| + |\bar{u} + \bar{c}|) \Delta u \begin{pmatrix} 0 \\ \mathbf{n} \\ \bar{u} \end{pmatrix} - \bar{\varrho} |\bar{u}| \Delta v \begin{pmatrix} 0 \\ \mathbf{t} \\ \bar{v} \end{pmatrix} \\ - \frac{1}{2\bar{c}^2} (|\bar{u} - \bar{c}| + |\bar{u} + \bar{c}| - 2|\bar{u}|) \Delta p \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} \end{pmatrix} \\ \left. - \frac{1}{2\bar{c}} (|\bar{u} + \bar{c}| - |\bar{u} - \bar{c}|) \Delta p \begin{pmatrix} 0 \\ \mathbf{n} \\ \bar{H} \end{pmatrix} \right\} \end{array} \right. \quad (6.20)$$

where

$$\bar{\varrho} = \sqrt{\varrho_L \varrho_R}, \quad \bar{c}^2 = \bar{\chi} + \bar{\kappa} \bar{h}, \quad \bar{h} = \bar{H} - \frac{1}{2} (\bar{u}^2 + \bar{v}^2). \quad (6.21)$$

We now pass to the case of two space dimensions. Let  $\mathbf{U}_L$  and  $\mathbf{U}_R$  be two states separated by a straight interface; we denote by  $\mathbf{n}$  the unit normal vector to the interface oriented from  $\mathbf{U}_L$  to  $\mathbf{U}_R$  and by  $\zeta$  the coordinate along  $\mathbf{n}$ . We then introduce the system

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial}{\partial \zeta} (\mathbb{F} \cdot \mathbf{n})(\mathbf{U}) = \mathbf{0} \quad (6.22)$$

where

$$(\mathbb{F} \cdot \mathbf{n})(\mathbf{U}) = \begin{pmatrix} \varrho(\mathbf{u} \cdot \mathbf{n}) \\ \varrho(\mathbf{u} \cdot \mathbf{n})\mathbf{u} + p\mathbf{n} \\ (\varrho e + p)(\mathbf{u} \cdot \mathbf{n}) \end{pmatrix}. \quad (6.23)$$

Since the Euler system is invariant under rotation, starting from the one-dimensional flux (6.24), the numerical flux of the Roe solver associated with the system (6.22),(6.23) is then given by

$$\left\{ \begin{array}{l} \Phi(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) = \frac{1}{2} \left\{ (\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_L) + (\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_R) - |\bar{\mathbf{u}} \cdot \mathbf{n}| \Delta \varrho \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} - \frac{\bar{c}^2}{\bar{\kappa}} \end{pmatrix} \right. \\ \quad - \frac{\bar{\varrho}}{2\bar{c}} (|\bar{\mathbf{u}} \cdot \mathbf{n} + \bar{c}| - |\bar{\mathbf{u}} \cdot \mathbf{n} - \bar{c}|) \Delta(\mathbf{u} \cdot \mathbf{n}) \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} \end{pmatrix} \\ \quad - \frac{\bar{\varrho}}{2} (|\bar{\mathbf{u}} \cdot \mathbf{n} - \bar{c}| + |\bar{\mathbf{u}} \cdot \mathbf{n} + \bar{c}|) \Delta(\mathbf{u} \cdot \mathbf{n}) \begin{pmatrix} 0 \\ \mathbf{n} \\ \bar{\mathbf{u}} \cdot \mathbf{n} \end{pmatrix} - \bar{\varrho} |\bar{\mathbf{u}} \cdot \mathbf{n}| \Delta(\mathbf{u} \cdot \mathbf{t}) \begin{pmatrix} 0 \\ \mathbf{t} \\ \bar{\mathbf{u}} \cdot \mathbf{t} \end{pmatrix} \\ \quad - \frac{1}{2\bar{c}^2} (|\bar{\mathbf{u}} \cdot \mathbf{n} - \bar{c}| + |\bar{\mathbf{u}} \cdot \mathbf{n} + \bar{c}| - 2|\bar{\mathbf{u}} \cdot \mathbf{n}|) \Delta p \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} \end{pmatrix} \\ \quad \left. - \frac{1}{2\bar{c}} (|\bar{\mathbf{u}} \cdot \mathbf{n} + \bar{c}| - |\bar{\mathbf{u}} \cdot \mathbf{n} - \bar{c}|) \Delta p \begin{pmatrix} 0 \\ \mathbf{n} \\ \bar{\mathbf{u}} \cdot \mathbf{n} \end{pmatrix} \right\}. \end{array} \right. \quad (6.24)$$

One can easily check that the condition (6.12) holds while, in the case of a slip boundary condition (6.9), the discrete boundary conditions (6.16) and (6.17) imply (6.15).

Now, in order to analyze the behavior of the numerical scheme as  $M$  tends to zero, it is enough to assume that the flow is subsonic, i.e.,

$$|\bar{\mathbf{u}} \cdot \mathbf{n}| < \bar{c}. \quad (6.25)$$

Then the numerical flux (6.24) becomes

$$\left\{ \begin{array}{l} \Phi(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) = \frac{1}{2} \left\{ (\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_L) + (\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_R) - |\bar{\mathbf{u}} \cdot \mathbf{n}| \Delta \varrho \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} - \frac{\bar{c}^2}{\bar{\kappa}} \end{pmatrix} \right. \\ \quad - \frac{\bar{\varrho}}{\bar{c}} (\bar{\mathbf{u}} \cdot \mathbf{n}) \Delta(\mathbf{u} \cdot \mathbf{n}) \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} \end{pmatrix} - \bar{\varrho} \bar{c} \Delta(\mathbf{u} \cdot \mathbf{n}) \begin{pmatrix} 0 \\ \mathbf{n} \\ \bar{\mathbf{u}} \cdot \mathbf{n} \end{pmatrix} - \bar{\varrho} |\bar{\mathbf{u}} \cdot \mathbf{n}| \Delta(\mathbf{u} \cdot \mathbf{t}) \begin{pmatrix} 0 \\ \mathbf{t} \\ \bar{\mathbf{u}} \cdot \mathbf{t} \end{pmatrix} \\ \quad \left. - \frac{1}{\bar{c}^2} (\bar{c} - |\bar{\mathbf{u}} \cdot \mathbf{n}|) \Delta p \begin{pmatrix} 1 \\ \bar{\mathbf{u}} \\ \bar{H} \end{pmatrix} - \frac{1}{\bar{c}} (\bar{\mathbf{u}} \cdot \mathbf{n}) \Delta p \begin{pmatrix} 0 \\ \mathbf{n} \\ \bar{\mathbf{u}} \cdot \mathbf{n} \end{pmatrix} \right\}. \end{array} \right. \quad (6.26)$$

Hence, in the subsonic regime, using (6.14) and (6.26), we obtain that the mass, momentum, and energy semi-discrete conservation equations of the semi-discrete finite-volume Roe scheme read, respectively, as

$$\left\{ \begin{array}{l} \frac{d\varrho_i}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \{ \varrho_j \mathbf{u}_j \cdot \mathbf{n}_{ij} + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|(\varrho_i - \varrho_j) \\ + \frac{\varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij})(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} + \frac{1}{c_{ij}^2} (c_{ij} - |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|)(p_i - p_j) \} = 0, \end{array} \right. \quad (6.27)$$

$$\left\{ \begin{array}{l} \frac{d(\varrho_i \mathbf{u}_i)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \{ \varrho_j (\mathbf{u}_j \cdot \mathbf{n}_{ij}) \mathbf{u}_j + p_j \mathbf{n}_{ij} \\ + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|(\varrho_i - \varrho_j) \mathbf{u}_{ij} \\ + \varrho_{ij} ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij}) \left( \frac{(\mathbf{u}_{ij} \cdot \mathbf{n}_{ij})}{c_{ij}} \mathbf{u}_{ij} + c_{ij} \mathbf{n}_{ij} \right) \\ + \varrho_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij}) \mathbf{t}_{ij} \\ + \frac{1}{c_{ij}^2} (p_i - p_j) \left( (c_{ij} - |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|) \mathbf{u}_{ij} + \frac{(\mathbf{u}_{ij} \cdot \mathbf{n}_{ij})}{c_{ij}} \mathbf{n}_{ij} \right) \} = \mathbf{0}, \end{array} \right. \quad (6.28)$$

$$\left\{ \begin{array}{l} \frac{d(\varrho_i e_i)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \{ (\varrho_j e_j + p_j) (\mathbf{u}_j \cdot \mathbf{n}_{ij}) \\ + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \left( H_{ij} - \frac{c_{ij}^2}{\kappa_{ij}} \right) (\varrho_i - \varrho_j) \\ + \frac{\varrho_{ij} (H_{ij} + c_{ij}^2)}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \\ + \varrho_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| (\mathbf{u}_{ij} \cdot \mathbf{t}_{ij}) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij} \\ + \frac{1}{c_{ij}^2} \left( (c_{ij} - |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|) H_{ij} + c_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|^2 + \frac{c_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|}{\kappa_{ij}} \right) (p_i - p_j) \} = 0. \end{array} \right. \quad (6.29)$$

In (6.27)–(6.29), we have

$$\left\{ \begin{array}{l} \varrho_{ij} = \sqrt{\varrho_i \varrho_j}, \quad \mathbf{u}_{ij} = \frac{\sqrt{\varrho_i} \mathbf{u}_i + \sqrt{\varrho_j} \mathbf{u}_j}{\sqrt{\varrho_i} + \sqrt{\varrho_j}}, \quad H_{ij} = \frac{\sqrt{\varrho_i} H_i + \sqrt{\varrho_j} H_j}{\sqrt{\varrho_i} + \sqrt{\varrho_j}}, \\ \chi_{ij} = \bar{\chi}(\mathbf{U}_i, \mathbf{U}_j), \quad \kappa_{ij} = \bar{\kappa}(\mathbf{U}_i, \mathbf{U}_j), \\ c_{ij}^2 = \chi_{ij} + \kappa_{ij} h_{ij}, \quad h_{ij} = H_{ij} - \frac{1}{2} |\mathbf{u}_{ij}|^2. \end{array} \right.$$

Recall that we have assumed  $|\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \leq c_{ij}$  for all  $i, j \in I$ .

In order to analyze the asymptotic behavior of the semi-discrete Roe scheme, we begin by deriving the nondimensional form of Eqs. (6.27)–(6.29). Using the same scaling as previously (leading to (6.4)), they are obtained directly by replacing  $p$  by  $\frac{p}{M^2}$ ,  $c$  by  $\frac{c}{M}$ , and  $h$  by  $\frac{h}{M^2}$  (or equivalently  $H$  by  $\frac{h}{M^2} + \frac{1}{2} |\mathbf{u}|^2$ ). Then they become, respectively (writing the terms in increasing

order wrt.  $M$ ),

$$\left\{ \begin{array}{l} \frac{d\varrho_i}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{M} \frac{1}{c_{ij}} (p_i - p_j) + \varrho_j \mathbf{u}_j \cdot \mathbf{n}_{ij} \right. \\ \quad \left. + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \left( \varrho_i - \varrho_j - \frac{1}{c_{ij}^2} (p_i - p_j) \right) \right. \\ \quad \left. + M \frac{\varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \right\} = 0, \end{array} \right. \quad (6.30)$$

$$\left\{ \begin{array}{l} \frac{d(\varrho_i \mathbf{u}_i)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{M^2} p_j \mathbf{n}_{ij} + \right. \\ \quad \left. + \frac{1}{M} \left[ \varrho_{ij} c_{ij} ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij}) \mathbf{n}_{ij} \right. \right. \\ \quad \left. \left. + \frac{1}{c_{ij}} (\mathbf{u}_{ij} + (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) \mathbf{n}_{ij}) (p_i - p_j) \right] \right. \\ \quad \left. + \varrho_j (\mathbf{u}_j \cdot \mathbf{n}_{ij}) \mathbf{u}_j + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \left( \varrho_i - \varrho_j - \frac{1}{c_{ij}^2} (p_i - p_j) \right) \mathbf{u}_{ij} \right. \\ \quad \left. \left. - \varrho_{ij} (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij} \mathbf{t}_{ij} + M \frac{\varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \right) \mathbf{u}_{ij} \right\} = \mathbf{0}, \end{array} \right. \quad (6.31)$$

and

$$\left\{ \begin{array}{l} \frac{d(\varrho_i \varepsilon_i)}{dt} + \frac{M^2}{2} \frac{d(\varrho_i |\mathbf{u}_i|^2)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{M} \frac{h_{ij}}{c_{ij}} (p_i - p_j) \right. \\ \quad \left. + \varrho_j h_j (\mathbf{u}_j \cdot \mathbf{n}_{ij}) + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \left( h_{ij} (\varrho_i - \varrho_j - \frac{1}{c_{ij}^2} (p_i - p_j) - \frac{c_{ij}^2}{\kappa_{ij}} (\varrho_i - \varrho_j)) \right) + \right. \\ \quad \left. + M \left[ (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) \frac{\varrho_{ij}}{c_{ij}} (h_{ij} + c_{ij}^2) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} + \frac{1}{2} \frac{|\mathbf{u}_{ij}|^2}{c_{ij}} (p_i - p_j) \right] \right. \\ \quad \left. + M^2 \left[ \frac{1}{2} \varrho_i |\mathbf{u}_j|^2 (\mathbf{u}_j \cdot \mathbf{n}_{ij}) \right. \right. \\ \quad \left. \left. + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \left( \frac{1}{2} |\mathbf{u}_{ij}|^2 \left( \varrho_i - \varrho_j - \frac{1}{c_{ij}^2} (p_i - p_j) \right) + \varrho_{ij} (\mathbf{u}_{ij} \cdot \mathbf{t}_{ij}) ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij}) \right) \right] \right. \\ \quad \left. + \frac{M^3}{2} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) \frac{\varrho_{ij}}{c_{ij}} |\mathbf{u}_{ij}|^2 (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \right\} = 0. \end{array} \right. \quad (6.32)$$

Again, for each quantity  $\varphi$ , we postulate an asymptotic expansion of the form

$$\varphi = \varphi^{(0)} + M\varphi^{(1)} + M^2\varphi^{(2)} + \dots$$

Then Eq. (6.32) gives at the order  $-1$  in  $M$

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(0)} - p_j^{(0)}) = 0.$$

Multiplying the above equation by  $p_i$ , summing over  $i$ , and taking into account either the periodicity of  $p$  or the artificial boundary conditions (6.17) which imply  $p_j = p_i$  for any fictitious cell  $\Omega_j$  adjacent to a boundary cell  $\Omega_i$ , we obtain

$$\sum_{i \in I} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(0)} - p_j^{(0)}) p_i^{(0)} = \sum_{\Gamma_{ij}} |\Gamma_{ij}| \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(0)} - p_j^{(0)})^2 = 0$$

and therefore

$$p_i^{(0)} = p^{(0)} \text{ is independent of } i.$$

Next, (6.32) gives at the order 0 in  $M$

$$\left\{ \begin{array}{l} \frac{d(\varrho_i^{(0)} \varepsilon_i^{(0)})}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^1 - p_j^1) + \varrho_j^{(0)} h_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right. \right. \\ \left. \left. + |\mathbf{u}_{ij}^{(0)} \cdot \mathbf{n}_{ij}| \left( h_{ij}^{(0)} - \frac{(c_{ij}^{(0)})^2}{\kappa_{ij}^{(0)}} (\varrho_i^{(0)} - \varrho_j^{(0)}) \right) \right\} = 0. \end{array} \right. \quad (6.33)$$

Assume for simplicity an equation of state of a polytropic ideal gas for which

$$\varrho \varepsilon = \frac{p}{\gamma - 1} \quad \varrho h = \frac{\gamma p}{\gamma - 1}, \quad \bar{h} = \frac{\bar{c}^2}{\bar{\kappa}} = \frac{\bar{c}^2}{\gamma - 1}.$$

Then (6.33) reads

$$\frac{dp^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ c_{ij}^{(0)} (p_i^1 - p_j^1) + \gamma p^{(0)} \mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij} \right\} = 0.$$

Multiplying this equation by  $|\Omega_i|$ , summing with respect to  $i$ , and taking into account either the periodicity conditions or the boundary conditions (6.16) and (6.17), we obtain

$$(\sum_{i \in I} |\Omega_i|) \frac{dp^{(0)}}{dt} = 0.$$

Hence we have

$$p_i^{(0)} = p^{(0)} = \text{constant} \quad (6.34)$$

and (6.33) becomes

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ c_{ij}^{(0)} (p_i^1 - p_j^1) + \gamma p^{(0)} \mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij} \right\} = 0. \quad (6.35)$$

Equation (6.31) gives at the order  $-2$  in  $M$

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| p_j^{(0)} \mathbf{n}_{ij} = \mathbf{0}.$$

It follows from (6.13) and (6.34) that this relation is automatically satisfied. Next, (6.31) gives at the order  $-1$  in  $M$

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ p_j^{(1)} + \varrho_{ij}^{(0)} c_{ij}^{(0)} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij}) \right\} \mathbf{n}_{ij} = \mathbf{0}.$$

so that for a polytropic ideal gas

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ p_j^{(1)} + \frac{\gamma p^{(0)}}{c_{ij}^{(0)}} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij}) \right\} \mathbf{n}_{ij} = \mathbf{0}. \quad (6.36)$$

Let us then characterize the pairs  $(\mathbf{u}_i^{(0)}, p_i^{(1)})$ ,  $i \in I$ , which satisfy both conditions (6.35) and (6.36).

*Proposition 6.1*

The pairs  $(\mathbf{u}_i^{(0)}, p_i^{(1)})$ ,  $i \in I$  satisfy the conditions (6.35) and (6.36) if and only if

$$p_i^{(1)} = p^{(1)} \text{ is independent of } i, \quad (6.37)$$

$$(\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij} = 0 \text{ for all edge } \Gamma_{ij}. \quad (6.38)$$

*Proof.* Multiplying (6.35) by  $p_i^{(1)}$  and summing with respect to  $i$ , we obtain

$$\begin{aligned} & \sum_{i \in I} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ c_{ij}^{(0)} (p_i^1 - p_j^1) p_i^{(1)} + \gamma p^{(0)} p_i^{(1)} \mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij} \right\} \\ &= \sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ c_{ij}^{(0)} (p_i^1 - p_j^1)^2 + \gamma p^{(0)} (p_i^{(1)} \mathbf{u}_j^{(0)} - p_j^{(1)} \mathbf{u}_i^{(0)}) \cdot \mathbf{n}_{ij} \right\} = 0. \end{aligned}$$

Similarly, taking the inner product of (6.36) with  $\mathbf{u}_i^{(0)}$  and summing with respect to  $i$ , we find

$$\begin{aligned} & \sum_{i \in I} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ p_j^{(1)} \mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij} + \frac{\gamma p^{(0)}}{c_{ij}^{(0)}} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij}) (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) \right\} \\ &= \sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ (p_j^{(1)} \mathbf{u}_i^{(0)} - p_i^{(1)} \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij} + \frac{\gamma p^{(0)}}{c_{ij}^{(0)}} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij})^2 \right\} = 0. \end{aligned}$$

Then combining the above equations yields

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ c_{ij}^{(0)} (p_i^1 - p_j^1)^2 + \frac{(\gamma p^{(0)})^2}{c_{ij}^{(0)}} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij})^2 \right\} = 0,$$

and (6.37) and (6.38) follow. Conversely, if (6.37) and (6.38) hold, using (6.13), it is an obvious matter to check that the conditions (6.35) and (6.36) are satisfied.  $\square$

Note that (6.37) is the discrete analog (and a “perfect” approximation) of  $\nabla p^{(1)} = \mathbf{0}$  while (6.38) is the discrete analog of the incompressibility constraint  $\nabla \cdot \mathbf{u}^{(0)} = 0$ . In fact, (6.38) is not a good approximation of  $\nabla \cdot \mathbf{u}^{(0)} = 0$  unless the cells  $\Omega_i$  are triangles. Indeed, in the case of a uniform mesh where the cells are rectangles whose sides are parallel to the axes with unit vectors  $\mathbf{e}_k$ ,  $k = 1, 2$ , this condition (6.38) means that each component  $\mathbf{u}_i^{(0)} \cdot \mathbf{e}_k$  is constant along parallels to  $\mathbf{e}_k$ . In other words, (6.38) is an approximation of the conditions

$$\frac{\partial}{\partial x_k} (\mathbf{u}^{(0)} \cdot \mathbf{e}_k) = 0, \quad k = 1, 2,$$

which are far stronger than the incompressibility constraint.

On the other hand, when the cells  $\Omega_i$  are triangles, we define

$$\mathbf{u}^h = \sum_{i \in I} \mathbf{u}_i 1_{\Omega_i}$$

and

$$\Psi^h = \{ \psi^h \in \mathcal{C}^0(\bar{\Omega}); \psi^h|_{\Omega_i} \in P_1 \}$$

where  $P_1$  denotes the set of all polynomials of degree  $\leq 1$ . We can now state

*Proposition 6.2*

Assume that the cells  $\Omega_i$  are triangles. Then the condition

$$(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} = 0 \text{ for all side } \Gamma_{ij} \quad (6.39)$$

holds provided that there exists a function  $\psi^h \in \Psi^h$  such that

$$\mathbf{u}^h = \mathbf{curl} \psi^h, \quad \psi^h \in \Psi^h. \quad (6.40)$$

*Proof.* In each triangle  $\Omega_i$ , we can write

$$\mathbf{u}_i = \mathbf{u}^h|_{\Omega_i} = \mathbf{curl} \psi_i, \quad \psi_i \in P_1.$$

Setting

$$\psi^h = \sum_{i \in I} \psi_i 1_{\Omega_i},$$

the condition (6.39) expresses the continuity of the tangential derivative of  $\psi^h$  along the interface  $\Gamma_{ij}$  between two adjacent triangles  $\Omega_i$  and  $\Omega_j$ . Hence (6.40) follows.  $\square$

We thus obtain that the condition (6.38) is a “good” approximation of the incompressibility constraint  $\nabla \cdot \mathbf{u}^{(0)} = 0$  *in the case of a triangular mesh*.

*Remark 6.1.* Let us next derive the discrete analogs of the first two equations (6.11). Using (6.34), (6.37), Eqs. (6.30) and (6.31) give at the order 0 in  $M$

$$\frac{d\varrho_i^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \varrho_j^{(0)} \mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij} + |\mathbf{u}_{ij}^{(0)} \cdot \mathbf{n}_{ij}| (\varrho_i^{(0)} - \varrho_j^{(0)}) \right\} = 0,$$

and

$$\begin{cases} \frac{d(\varrho_i^{(0)} \mathbf{u}_i^{(0)})}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ (p_j^{(2)} \mathbf{n}_{ij} + \varrho_{ij}^{(0)} c_{ij}^{(0)} ((\mathbf{u}_i^{(1)} - \mathbf{u}_j^{(1)}) \cdot \mathbf{n}_{ij}) \mathbf{n}_{ij} + (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) \varrho_j^{(0)} \mathbf{u}_j^{(0)} + |\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}| [(\varrho_i^{(0)} - \varrho_j^{(0)}) \mathbf{u}_{ij}^{(0)} + \varrho_{ij}^{(0)} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{t}_{ij}) \mathbf{t}_{ij}] \right\} = \mathbf{0}. \end{cases}$$

The two above equations have to be supplemented with the discrete incompressibility constraint (6.38). Note however that this set of equations is not closed due to the presence of the term  $\varrho_{ij}^{(0)} c_{ij}^{(0)} ((\mathbf{u}_i^{(1)} - \mathbf{u}_j^{(1)}) \cdot \mathbf{n}_{ij}) \mathbf{n}_{ij}$  in the last equation.  $\square$

### 6.3 An All-Mach Semi-Discrete Roe Scheme

In Sect. 6.2, we have observed that, in general, the semi-discrete Roe scheme is not asymptotic preserving as the Mach number tends to zero. In order to guarantee a good behavior of this scheme as  $M \rightarrow 0$ , we slightly modify the conservation equations (6.27)–(6.29). Indeed, in any such equation, we replace each term

$$(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \text{ by } \theta_{ij}((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij})$$

where

$$\theta_{ij} = \theta \left( \frac{|\mathbf{u}_{ij}|}{c_{ij}} \right) \quad (6.41)$$

and where  $\theta : [0, +\infty) \rightarrow \mathbb{R}_+$  is a continuous nondecreasing function such that

$$\theta(\xi) = \begin{cases} \xi \text{ in a neighborhood of } 0, \\ 1 \text{ for } \xi \geq 1. \end{cases}$$

For instance, we may take

$$\theta(\xi) = \min(\xi, 1) \Rightarrow \theta_{ij} = \min\left(\frac{|\mathbf{u}_{ij}|}{c_{ij}}, 1\right).$$

Then, the modified semi-discrete Roe scheme reads in the subsonic case

$$\left\{ \begin{array}{l} \frac{d\varrho_i}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \{ \varrho_j \mathbf{u}_j \cdot \mathbf{n}_{ij} + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| (\varrho_i - \varrho_j) \\ \quad + \theta_{ij} \frac{\varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \\ \quad + \frac{1}{c_{ij}^2} (c_{ij} - |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|) (p_i - p_j) \} = 0, \\ \\ \frac{d(\varrho_i \mathbf{u}_i)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \{ \varrho_j (\mathbf{u}_j \cdot \mathbf{n}_{ij}) \mathbf{u}_j + p_j \mathbf{n}_{ij} \\ \quad + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| (\varrho_i - \varrho_j) \mathbf{u}_{ij} \\ \quad + \theta_{ij} \varrho_{ij} ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij}) \left( \frac{\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}}{c_{ij}} \mathbf{u}_{ij} + c_{ij} \mathbf{n}_{ij} \right) \\ \quad + \varrho_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij}) \mathbf{t}_{ij} \\ \quad + \frac{1}{c_{ij}^2} (p_i - p_j) \left( (c_{ij} - |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|) \mathbf{u}_{ij} + \frac{\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}}{c_{ij}} \mathbf{n}_{ij} \right) \} = \mathbf{0}, \end{array} \right.$$

and

$$\left\{ \begin{array}{l} \frac{d(\varrho_i e_i)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \{ (\varrho_j e_j + p_j) (\mathbf{u}_j \cdot \mathbf{n}_{ij}) \\ \quad + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \left( H_{ij} - \frac{c_{ij}^2}{\kappa_{ij}} \right) (\varrho_i - \varrho_j) \\ \quad + \frac{\theta_{ij} \varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) (H_{ij} + c_{ij}^2) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \\ \quad + \varrho_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| (\mathbf{u}_{ij} \cdot \mathbf{t}_{ij}) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij} \\ \\ \frac{1}{c_{ij}^2} \left( (c_{ij} - |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|) H_{ij} + c_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|^2 + \frac{c_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|}{\kappa_{ij}} \right) (p_i - p_j) \} = 0. \end{array} \right.$$

Since  $\theta_{ij} = \mathcal{O}(M)$  as  $M$  tends to zero, we obtain the nondimensional form of the modified conservation equations in the subsonic case by replacing in the above equations  $p$  by  $\frac{p}{M^2}$ ,  $c$  by  $\frac{c}{M}$ , and  $\theta$  by  $M\theta$  which yields, respectively,

$$\left\{ \begin{array}{l} \frac{d\varrho_i}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{M} \frac{1}{c_{ij}} (p_i - p_j) + \right. \\ \quad + \varrho_j \mathbf{u}_j \cdot \mathbf{n}_{ij} + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| (\varrho_i - \varrho_j) - \frac{1}{c_{ij}^2} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| (p_i - p_j) \\ \quad \left. + M \frac{\theta_{ij}\varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) (\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij} \right\} = 0, \end{array} \right. \quad (6.42)$$

$$\left\{ \begin{array}{l} \frac{d(\varrho_i \mathbf{u}_i)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{M^2} p_j \mathbf{n}_{ij} \right. \\ \quad + \frac{1}{M} \frac{1}{c_{ij}} (\mathbf{u}_{ij} + (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) \mathbf{n}_{ij}) (p_i - p_j) + \varrho_j ((\mathbf{u}_j \cdot \mathbf{n}_{ij}) \mathbf{u}_j \\ \quad + |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \mathbf{u}_{ij} (\varrho_i - \varrho_j) + \theta_{ij}\varrho_{ij} c_{ij} \mathbf{n}_{ij} ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij}) \\ \quad \left. + \varrho_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| \mathbf{t}_{ij} ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij}) - \frac{|\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|}{c_{ij}^2} \mathbf{u}_{ij} (p_i - p_j) \right. \\ \quad \left. + M^2 \frac{\theta_{ij}\varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) \mathbf{u}_{ij} ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij}) \right\} = \mathbf{0}, \end{array} \right. \quad (6.43)$$

and

$$\left\{ \begin{array}{l} \frac{d(\varrho_i \varepsilon_i)}{dt} + \frac{M^2}{2} \frac{d(\varrho_i |\mathbf{u}_i|^2)}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{M} \frac{h_{ij}}{c_{ij}} (p_i - p_j) \right. \\ \quad + \varrho_j h_j (\mathbf{u}_j \cdot \mathbf{n}_{ij}) + |\mathbf{u}_j \cdot \mathbf{n}_{ij}| \left( h_{ij} - \frac{c_{ij}^2}{\kappa_{ij}} \right) (\varrho_i - \varrho_j) \\ \quad - \frac{|\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|}{c_{ij}^2} h_{ij} (p_i - p_j) + M \frac{1}{2} \frac{|\mathbf{u}_{ij}|^2}{c_{ij}} (p_i - p_j) \\ \quad + M^2 \left[ \frac{1}{2} \varrho_j |\mathbf{u}_j|^2 (\mathbf{u}_j \cdot \mathbf{n}_{ij}) + \frac{1}{2} |\mathbf{u}_j \cdot \mathbf{n}_{ij}| |\mathbf{u}_{ij}|^2 (\varrho_i - \varrho_j) + \right. \\ \quad + \frac{\theta_{ij}\varrho_{ij}}{c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) (h_{ij} + c_{ij}^2) ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij}) \\ \quad \left. + \varrho_{ij} |\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}| (\mathbf{u}_{ij} \cdot \mathbf{t}_{ij}) ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{t}_{ij}) - \frac{1}{2} \frac{|\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}|}{c_{ij}^2} |\mathbf{u}_{ij}|^2 (p_i - p_j) \right] \\ \quad \left. + M^4 \frac{\theta_{ij}\varrho_{ij}}{2c_{ij}} (\mathbf{u}_{ij} \cdot \mathbf{n}_{ij}) |\mathbf{u}_{ij}|^2 ((\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{n}_{ij}) \right\} = 0. \end{array} \right. \quad (6.44)$$

First, at the order  $-1$  in  $M$ , (6.44) gives

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(0)} - p_j^{(0)}) = 0$$

which yields as usual

$$p_i^{(0)} = p^{(0)} \text{ is independent of } i. \quad (6.45)$$

Next, at the order 0 in  $M$ , (6.42) and (6.44) give, respectively,

$$\left\{ \begin{array}{l} \frac{d\varrho_i^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)}) + \varrho_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right. \\ \left. + |\mathbf{u}_{ij}^{(0)} \cdot \mathbf{n}_{ij}| (\varrho_i^{(0)} - \varrho_j^{(0)}) \right\} = 0 \end{array} \right. \quad (6.46)$$

and

$$\left\{ \begin{array}{l} \frac{d(\varrho_i^{(0)} \varepsilon_i^{(0)})}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)}) \right. \\ \left. + \varrho_j^{(0)} h_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right. \\ \left. + |\mathbf{u}_{ij}^{(0)} \cdot \mathbf{n}_{ij}| \left( h_{ij}^{(0)} - \frac{(c_{ij}^0)^2}{\kappa_{ij}^{(0)}} \right) (\varrho_i^{(0)} - \varrho_j^{(0)}) \right\} = 0. \end{array} \right. \quad (6.47)$$

Now, we distinguish several cases depending on the given equation of state.

(i) *The case  $\chi = 0$ ,  $\kappa = p'(\varrho\varepsilon) > 0$ .*

We have  $p = p(\varrho\varepsilon)$ ; in particular, this is the case of a polytropic ideal gas. Then we have  $c^2 = \kappa h$ . Hence, multiplying (6.47) by  $\kappa_i^{(0)}$  and using (6.13), we obtain

$$\left\{ \begin{array}{l} \frac{dp^{(0)}}{dt} + \frac{\kappa_i^{(0)}}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)}) \right. \\ \left. + \varrho_i^{(0)} h_i^{(0)} (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) + \varrho_j^{(0)} h_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \end{array} \right.$$

Next, multiplying the above equation by  $\frac{|\Omega_i|}{\kappa_i^{(0)}}$  and summing with respect to  $i$ , we find

$$\left( \sum_{i \in I} \frac{|\Omega_i|}{\kappa_i^{(0)}} \right) \frac{dp^{(0)}}{dt} = 0 \iff \frac{dp^{(0)}}{dt} = 0,$$

so that on the one hand

$$p^{(0)} = \text{constant}$$

and on the other hand

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)}) + \varrho_j^{(0)} h_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \quad (6.48)$$

Now, we observe that  $\varrho\varepsilon$  may be defined as a strictly increasing function of  $p$ , and we can write

$$\varrho h = \varrho\varepsilon + p = f(p)$$

for some strictly increasing function  $f$ . Hence (6.48) becomes

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)}) + f(p^{(0)}) (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \quad (6.49)$$

Next, (6.43) yields at the order  $-2$  in  $M$

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| p_j^{(0)} \mathbf{n}_{ij} = \mathbf{0}.$$

Again, due to (6.13), this condition is automatically satisfied. On the other hand, (6.43) yields at the order  $-1$  in  $M$

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| p_j^{(1)} \mathbf{n}_{ij} = \mathbf{0}. \quad (6.50)$$

It remains to characterize the pairs  $(\mathbf{u}_i^{(0)}, p_i^{(1)})$  which satisfy the conditions (6.49) and (6.50).

*Proposition 6.3*

The pairs  $(\mathbf{u}_i^{(0)}, p_i^{(1)})$ ,  $i \in I$ , satisfy (6.49) and (6.50) if and only if

$$p_i^{(1)} = p^{(1)} \text{ is independent of } i, \quad (6.51)$$

and

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij} = 0 \text{ for all } i \in I. \quad (6.52)$$

*Proof.* Multiplying (6.49) by  $p_i^1$  and summing with respect to  $i$ , we obtain

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)})^2 + f(p^{(0)}) (p_i^{(1)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) - p_j^{(1)} (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij})) \right\} = 0.$$

Similarly, taking the inner product of (6.50) with  $\mathbf{u}_i^{(0)}$  and summing with respect to  $i$ , we find

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ p_j^{(1)} (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) - p_i^{(1)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0.$$

Combining the above equations yields

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| \frac{h_{ij}^{(0)}}{c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)})^2 = 0,$$

and (6.49) and (6.50) follow. Conversely, if (6.49) and (6.50) hold, the conditions (6.47) and (6.48) are obviously satisfied.  $\square$

(ii) *The barotropic case*  $\kappa = 0$ ,  $\chi = p'(\varrho) > 0$ .

In this case, (6.45) reads equivalently

$$\varrho_i^{(0)} = \varrho^{(0)} \text{ is independent of } i,$$

while (6.46) becomes

$$\frac{d\varrho^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{c^{(0)}} (p_i^{(1)} - p_j^{(1)}) + \varrho^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0.$$

Multiplying this equation by  $2|\Omega_i|$  and summing with respect to  $i$ , we obtain

$$\varrho^{(0)} = \text{constant} \iff p^{(0)} = \text{constant},$$

and

$$\sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \frac{1}{c^{(0)}} (p_i^{(1)} - p_j^{(1)}) + \varrho^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \quad (6.53)$$

Again the condition (6.50) follows from the momentum conservation equation (6.42). Then, one can prove exactly as above the analog of Proposition 6.3: *the pairs  $(\mathbf{u}_i^{(0)}, p_i^{(1)})$ ,  $i \in I$ , satisfy (6.50) and (6.53) if and only if the conditions (6.51) and (6.52) hold.*

(iii) *Toward a general equation of state.*

Up to now, we have only considered the cases  $\chi = 0$  or  $\kappa = 0$ . It remains to analyze the general case where both  $\chi \neq 0$  and  $\kappa \neq 0$ . Since

$$\frac{dp_i^{(0)}}{dt} = \chi_i^{(0)} \frac{d\varrho_i^{(0)}}{dt} + \kappa_i^{(0)} \frac{d(\varrho_i^{(0)} \varepsilon_i^{(0)})}{dt},$$

(6.46) and (6.47) imply

$$\left\{ \begin{array}{l} \frac{dp^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} \left\{ \frac{1}{c_{ij}^{(0)}} (\chi_i^{(0)} + \kappa_i^{(0)} h_{ij}^{(0)}) (p_i^{(1)} - p_j^{(1)}) + \right. \\ \left. + (\chi_i^{(0)} + \kappa_i^{(0)} h_j^{(0)}) \varrho_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) + \right. \\ \left. + \left( \chi_i^{(0)} + \kappa_i^{(0)} \left( h_{ij}^{(0)} - \frac{(c_{ij}^{(0)})^2}{\kappa_{ij}^{(0)}} \right) \right) |\mathbf{u}_{ij}^{(0)} \cdot \mathbf{n}_{ij}| (\varrho_i^{(0)} - \varrho_j^{(0)}) \right\} = 0. \end{array} \right.$$

Using the relation  $c_{ij}^2 = \chi_{ij} + \kappa_{ij} h_{ij}$ , we have

$$h_{ij}^{(0)} - \frac{(c_{ij}^{(0)})^2}{\kappa_{ij}^{(0)}} = -\frac{\chi_{ij}^{(0)}}{\kappa_{ij}^{(0)}},$$

and the previous equation can be also written

$$\left\{ \begin{array}{l} \frac{dp^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial \Omega_i} \left\{ \frac{1}{c_{ij}^{(0)}} (\chi_i^{(0)} + \kappa_i^{(0)} h_{ij}^{(0)}) (p_i^{(1)} - p_j^{(1)}) + \right. \\ \quad + (\chi_i^{(0)} + \kappa_i^{(0)} h_i^{(0)}) \varrho_i^{(0)} (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) + (\chi_i^{(0)} + \kappa_i^{(0)} h_j^{(0)}) \varrho_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) + \\ \quad \left. + \left( \chi_i^{(0)} - \chi_{ij}^{(0)} \frac{\kappa_i^{(0)}}{c_{ij}^{(0)}} \right) |\mathbf{u}_{ij}^{(0)} \cdot \mathbf{n}_{ij}| (\varrho_i^{(0)} - \varrho_j^{(0)}) \right\} = 0. \end{array} \right. \quad (6.54)$$

Now, it is an easy matter to analyze the case where both  $\chi$  and  $\kappa$  are constants. Indeed (6.54) then reads

$$\left\{ \begin{array}{l} \frac{dp^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial \Omega_i} \left\{ \frac{1}{c_{ij}^{(0)}} (\chi + \kappa h_{ij}^{(0)}) (p_i^{(1)} - p_j^{(1)}) + \right. \\ \quad \left. + (\chi + \kappa h_i^{(0)}) \varrho_i^{(0)} (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) + (\chi + \kappa h_j^{(0)}) \varrho_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \end{array} \right.$$

Exactly as above, this implies that on the one hand

$$p^{(0)} = \text{constant}$$

and on the other hand

$$\sum_{\Gamma_{ij} \subset \partial \Omega_i} \left\{ \frac{1}{c_{ij}^{(0)}} (\chi + \kappa h_{ij}^{(0)}) (p_i^{(1)} - p_j^{(1)}) + (\chi + \kappa h_j^{(0)}) \varrho_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0.$$

Since

$$p = \chi \varrho + \kappa \varrho \varepsilon \iff p = \frac{(\chi + \kappa h) \varrho}{1 + \kappa},$$

the above condition becomes

$$\sum_{\Gamma_{ij} \subset \partial \Omega_i} \left\{ \frac{1}{\varrho_{ij}^{(0)} c_{ij}^{(0)}} (p_i^{(1)} - p_j^{(1)}) + (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \quad (6.55)$$

Again (6.50) holds and we have the following analog of Proposition 6.3: *the pairs  $(\mathbf{u}_i^{(0)}, p_i^{(1)})$ ,  $i \in I$ , satisfy the conditions (6.50) and (6.55) if and only if (6.51) and (6.52) hold.*

As a conclusion, we have obtained that the conditions (6.51) and (6.52) are valid in the three following cases

$$\left\{ \begin{array}{l} \text{(i)} \quad \chi = 0, \kappa > 0, \\ \text{(ii)} \quad \chi > 0 \text{ (barotropic case),} \\ \text{(iii)} \quad \chi \text{ and } \kappa \text{ are constant.} \end{array} \right. \quad (6.56)$$

It is worthwhile to notice that the assumptions (6.56) cover a number of cases of practical importance and in particular the case of a polytropic ideal gas

or a stiffened equation of state of Grüneisen type. We conjecture that the characterization (6.51), (6.52) still holds for any “reasonable” equation of state.

Observe that the condition (6.52) may be written, for instance,

$$\frac{1}{2} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| (\mathbf{u}_i^{(0)} + \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij} = 0$$

which is a discretization of

$$\int_{\partial\Omega_i} \mathbf{u}^{(0)} \cdot \mathbf{n} \, dS = \int_{\Omega_i} \nabla \cdot \mathbf{u}^{(0)} \, d\mathbf{x} = 0.$$

Hence (6.52) appears to be a “good” approximation of the incompressibility condition  $\nabla \cdot \mathbf{u}^{(0)} = 0$ , *independently of the geometry of the cell  $\Omega_i$* .

Let us now detail the limit of the all-Mach semi-discrete Roe scheme as the Mach number  $M$  tends to zero. Using (6.45) and (6.51), the nondimensional mass and momentum conservation equations (6.42) and (6.43) give, respectively, at the order 0 in  $M$

$$\frac{d\varrho_i^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \varrho_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) + |\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}| (\varrho_i^{(0)} - \varrho_j^{(0)}) \right\} = 0$$

and

$$\begin{cases} \frac{d(\varrho_i^{(0)} \mathbf{u}_i^{(0)})}{dt} + \frac{1}{2|\Omega_i|} \sum_{var\Gamma_{ij} \subset \partial\Omega_i} |\Gamma_{ij}| \left\{ \varrho_j^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \mathbf{u}_j^{(0)} + p_j^{(0)} \mathbf{n}_{ij} + \right. \\ \left. + |\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}| ((\varrho_i^{(0)} - \varrho_j^{(0)}) \mathbf{u}_{ij}^{(0)} + \theta_{ij}^{(0)} \varrho_{ij}^{(0)} c_{ij}^{(0)} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij}) \mathbf{n}_{ij} + \right. \\ \left. + \varrho_{ij}^{(0)} ((\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{t}_{ij}) \mathbf{t}_{ij} \right\} = 0. \end{cases}$$

If we supplement the two above equations with the discrete incompressibility condition (6.52), we obtain a closed system of equations which is the discrete analog of the limit system (6.11).

## 6.4 Asymptotic Analysis of the Semi-Discrete HLL Scheme

Again we first consider the system of gas dynamics in one space dimension (6.18). The numerical flux of the HLL method of approximation of (6.18) is given by

$$\Phi(\mathbf{U}_L, \mathbf{U}_R) = \frac{1}{a_R^+ - a_L^-} \{ a_R^+ \mathbf{f}(\mathbf{U}_L) - a_L^- \mathbf{f}(\mathbf{q}_R) + a_L^- a_R^+ (\mathbf{U}_R - \mathbf{U}_L) \}$$

where  $a_L$  and  $a_R$  are, respectively, a lower bound and an upper bound for the wave velocities of the solution of the Riemann problem connecting the two states  $\mathbf{U}_L$  and  $\mathbf{U}_R$ . We then know that the HLL scheme is entropy satisfying (cf. Chap. IV, Sect. 3.3). If we consider next the system (6.22) and (6.23), the numerical flux of the associated HLL method is easily seen to be given by

$$\Phi(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) = \frac{1}{a_R^+ - a_L^-} \{ a_R^+(\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_L) - a_L^-(\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_R) + a_L^- a_R^+(\mathbf{U}_R - \mathbf{U}_L) \}$$

where the wave velocities  $a_L$  and  $a_R$  remain to be specified. Since we are interested in the low Mach behavior of the method, we may suppose that  $a_L$  and  $a_R$  satisfy

$$a_L \leq (\mathbf{u}_L \cdot \mathbf{n}) - c_L < 0 < (\mathbf{u}_R \cdot \mathbf{n}) + c_R \leq a_R.$$

Hence we have

$$\Phi(\mathbf{U}_L, \mathbf{U}_R, \mathbf{n}) = \frac{1}{a_R - a_L} \{ a_R(\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_L) - a_L(\mathbb{F} \cdot \mathbf{n})(\mathbf{U}_R) + a_L^- a_R^+(\mathbf{U}_R - \mathbf{U}_L) \}.$$

Then it is a simple matter to check that, on the one hand, the condition (6.12) holds and, on the other hand, in the case of a slip boundary condition (6.9), the discrete boundary conditions (6.16) and (6.17) imply (6.15).

In the subsonic regime, the semi-discrete HLL method is therefore of the form (6.14) with

$$\left\{ \begin{array}{l} \Phi(\mathbf{U}_i, \mathbf{U}_j, \mathbf{n}_{ij}) = \frac{1}{a_{ij,R} - a_{ij,L}} \{ a_{ij,R}(\mathbb{F} \cdot \mathbf{n}_{ij})(\mathbf{U}_i) - a_{ij,L}(\mathbb{F} \cdot \mathbf{n}_{ij})(\mathbf{U}_j) \\ \quad + a_{ij,L} a_{ij,R}(\mathbf{U}_j - \mathbf{U}_i) \} \end{array} \right.$$

and

$$a_{ij,L} \leq (\mathbf{u}_i \cdot \mathbf{n}_{ij}) - c_i < 0 < (\mathbf{u}_j \cdot \mathbf{n}_{ij}) + c_j \leq a_{ij,R}. \quad (6.57)$$

The mass, momentum, and energy conservation equations of the scheme thus read, respectively,

$$\left\{ \begin{array}{l} \frac{d\rho_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \{ (a_{ij,R}(\varrho\mathbf{u})_i - a_{ij,L}(\varrho\mathbf{u})_j) \cdot \mathbf{n}_{ij} \\ \quad + a_{ij,L} a_{ij,R}(\varrho_j - \varrho_i) \} = 0, \end{array} \right. \quad (6.58)$$

$$\left\{ \begin{array}{l} \frac{d(\varrho\mathbf{u})_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \{ a_{ij,R}((\mathbf{u}_i \cdot \mathbf{n}_{ij})(\varrho\mathbf{u})_i + p_i \mathbf{n}_{ij}) \\ \quad - a_{ij,L}((\mathbf{u}_j \cdot \mathbf{n}_{ij})(\varrho\mathbf{u})_j + p_j \mathbf{n}_{ij}) + a_{ij,L} a_{ij,R}((\varrho\mathbf{u})_j - (\varrho\mathbf{u})_i) \} = \mathbf{0}, \end{array} \right. \quad (6.59)$$

$$\left\{ \begin{array}{l} \frac{d(\varrho e)_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \{ a_{ij,R}(\mathbf{u}_i \cdot \mathbf{n}_{ij})(\varrho e + p)_i \\ - a_{ij,L}(\mathbf{u}_j \cdot \mathbf{n}_{ij})(\varrho e + p)_j + a_{ij,L}a_{ij,R}((\varrho e)_j - (\varrho e)_i) \} = 0. \end{array} \right. \quad (6.60)$$

We pass to the study of the asymptotic behavior of the HLL scheme. We first consider the mass conservation equation (6.58). Since  $a_L, a_R = \mathcal{O}(\frac{1}{M})$ , its nondimensional form is obtained by replacing  $a$  by  $\frac{a}{M}$  which gives

$$\left\{ \begin{array}{l} \frac{d\varrho_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \left\{ \frac{1}{M} a_{ij,L} a_{ij,R} (\varrho_j - \varrho_i) \right. \\ \left. + (a_{ij,R}(\varrho \mathbf{u})_i - a_{ij,L}(\varrho \mathbf{u})_j) \cdot \mathbf{n}_{ij} \right\} = 0. \end{array} \right. \quad (6.61)$$

In all the sequel of this section, we will choose  $a_{ij,L}$  and  $a_{ij,R}$  so that

$$a_{ij,L}^{(0)} = -a_{ij,R}^{(0)} = -a_{ij}^{(0)} \quad (6.62)$$

and therefore

$$a_{ij}^{(0)} = a_{ji}^{(0)} > 0.$$

This is indeed the case if we take, for instance,

$$a_{ij,L} = -a_{ij,R} = \max(c_i - (\mathbf{u}_i \cdot \mathbf{n}_{ij}), c_j + (\mathbf{u}_j \cdot \mathbf{n}_{ij}))$$

which yields

$$a_{ij}^{(0)} = \max(c_i^{(0)}, c_j^{(0)}).$$

Then (6.61) gives at the order  $-1$  in  $M$

$$\sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| a_{ij}^{(0)} (\varrho_i^{(0)} - \varrho_j^{(0)}), = 0$$

and using a now usual argument,

$$\varrho_i^{(0)} = \varrho^{(0)} \text{ is independent of } i \in I.$$

Let us already notice that, except in the barotropic case, this constraint is far too strong!

Next, (6.61) gives at the order  $0$  in  $M$

$$\frac{d\varrho^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (\varrho_j^1 - \varrho_i^1) + \varrho^{(0)} (\mathbf{u}_i^{(0)} + \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij} \right\} = 0.$$

Using the conservation property

$$\frac{d}{dt} \left( \sum_{i \in I} |\Omega_i| \varrho_i^{(0)} \right) = 0,$$

we obtain

$$\varrho^{(0)} = \text{constant} \quad (6.63)$$

and

$$\sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (\varrho_j^1 - \varrho_i^1) + \varrho^{(0)} (\mathbf{u}_i^{(0)} + \mathbf{u}_j^{(0)}) \cdot \mathbf{n}_{ij} \right\} = 0. \quad (6.64)$$

We next consider the momentum conservation equation (6.59) whose nondimensional form is given by

$$\left\{ \begin{array}{l} \frac{d(\varrho \mathbf{u}_i)}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial \Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \left\{ \frac{1}{M^2} [a_{ij,R} p_i - a_{ij,L} p_j] \mathbf{n}_{ij} \right. \\ \quad \left. + \frac{1}{M} [a_{ij,L} a_{ij,R} ((\varrho \mathbf{u})_j - (\varrho \mathbf{u})_i)] \right. \\ \quad \left. + a_{ij,R} (\mathbf{u}_i \cdot \mathbf{n}_{ij}) (\varrho \mathbf{u})_i - a_{ij,L} ((\mathbf{u}_j \cdot \mathbf{n}_{ij}) (\varrho \mathbf{u})_j) \right\} = \mathbf{0}. \end{array} \right. \quad (6.65)$$

Then (6.65) gives at the order  $-2$  in  $M$

$$\sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| p_j^{(0)} \mathbf{n}_{ij} = \mathbf{0}. \quad (6.66)$$

Note that (6.66) is automatically satisfied if

$$p_i^{(0)} = p^{(0)} \text{ is independent of } i \in I. \quad (6.67)$$

Since in the barotropic case

$$p_i^{(0)} = p(\varrho_i^{(0)}) = p(\varrho^{(0)}) = p^{(0)},$$

this clearly occurs in this latter case. Next, under the assumption (6.67), (6.65) gives at the order  $-1$  in  $M$

$$\sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| \left\{ p_j^{(1)} \mathbf{n}_{ij} + \varrho^{(0)} a_{ij}^{(0)} (\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \right\} = \mathbf{0}. \quad (6.68)$$

In the barotropic case, since

$$p_i^{(1)} = (c^{(0)})^2 \rho_i^{(1)}, \quad a_{ij}^{(0)} = c^{(0)},$$

(6.68) becomes

$$\sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| \left\{ c^{(0)} \varrho_j^{(1)} \mathbf{n}_{ij} + \varrho^{(0)} (\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \right\} = \mathbf{0}. \quad (6.69)$$

Then we can state

*Proposition 6.4*

Assume a barotropic equation of state  $p = p(\varrho)$ . The pairs  $(\varrho_i^{(1)}, \mathbf{u}_i^{(0)}), i \in I$  satisfy the conditions (6.64) and (6.69) if and only if

$$\begin{cases} \varrho_i^{(1)} = \varrho^{(1)} \text{ is independent of } i \in I, \\ \mathbf{u}_i^{(0)} = \mathbf{u}^{(0)} \text{ is independent of } i \in I. \end{cases} \quad (6.70)$$

*Proof.* Multiplying (6.64) by  $\varrho_i^{(1)}$  and summing with respect to  $i$ , we obtain

$$\sum_{i \in I} \sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (\varrho_i^{(1)} - \varrho_j^{(1)}) \varrho_i^{(1)} + \varrho^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \varrho_i^{(1)} \right\} = \sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (\varrho_i^{(1)} - \varrho_j^{(1)})^2 + \varrho^{(0)} ((\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \varrho_i^{(1)} - (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) \varrho_j^{(1)}) \right\} = 0.$$

Next, taking the inner product of (6.69) with  $\mathbf{u}_i^{(0)}$  and summing with respect to  $i$  give

$$\sum_{i \in I} \sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| \left\{ (c^{(0)})^2 (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) \varrho_j^{(1)} + \varrho^{(0)} a_{ij}^{(0)} (\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}) \cdot \mathbf{u}_i^{(0)} \right\} = \sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ (c^{(0)})^2 ((\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) \varrho_j^{(1)} - (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \varrho_i^{(1)}) + \varrho^{(0)} a_{ij}^{(0)} |\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}|^2 \right\} = 0.$$

By combining the two above equations, we find

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| a_{ij}^{(0)} \left\{ (\varrho_i^{(1)} - \varrho_j^{(1)})^2 + \left( \frac{\varrho^{(0)}}{c^{(0)}} \right)^2 |\mathbf{u}_i^{(0)} - \mathbf{u}_j^{(0)}|^2 \right\} = 0,$$

and (6.70) follows.  $\square$

The second constraint (6.70) is obviously far too strong since it represents a fairly bad approximation of the incompressibility condition  $\nabla \cdot \mathbf{u}^{(0)} = 0$ .

At last, we consider the energy conservation equation (6.60) written in its nondimensional form

$$\begin{cases} \frac{d(\varrho\varepsilon)_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial \Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \left\{ \frac{1}{M} a_{ij,L} a_{ij,R} ((\varrho\varepsilon)_j - (\varrho\varepsilon)_i) \right. \\ \left. + a_{ij,R} (\mathbf{u}_i \cdot \mathbf{n}_{ij}) (\varrho\varepsilon + p)_i - a_{ij,L} (\mathbf{u}_j \cdot \mathbf{n}_{ij}) (\varrho\varepsilon + p)_j \right\} = \mathcal{O}(M). \end{cases} \quad (6.71)$$

At the order  $-1$  in  $M$ , (6.71) yields

$$\sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| a_{ij}^{(0)} ((\varrho\varepsilon)_j^{(0)}, -(\varrho\varepsilon)_i^{(0)}) = 0,$$

and therefore

$$(\varrho\varepsilon)_i^{(0)} = (\varrho\varepsilon)^{(0)} \text{ is independent of } i \in I. \quad (6.72)$$

At the order 0 in  $M$ , we obtain

$$\left\{ \begin{array}{l} \frac{d(\varrho\varepsilon)^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} ((\varrho\varepsilon)_i^{(1)} - (\varrho\varepsilon)_j^{(1)}) \right. \\ \left. + (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) ((\varrho\varepsilon)^{(0)} + p_i^{(0)}) + (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) ((\varrho\varepsilon)^{(0)} + p_j^{(0)}) \right\} = 0, \end{array} \right.$$

which yields

$$(\varrho\varepsilon)^{(0)} = \text{constant} \quad (6.73)$$

and

$$\left\{ \begin{array}{l} \sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} ((\varrho\varepsilon)_i^{(1)} - (\varrho\varepsilon)_j^{(1)}) + (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) ((\varrho\varepsilon)^{(0)} + p_i^{(0)}) \right. \\ \left. + (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) ((\varrho\varepsilon)^{(0)} + p_j^{(0)}) \right\} = 0. \end{array} \right. \quad (6.74)$$

If we assume for simplicity an equation of state of a polytropic ideal gas, (6.72) implies the property (6.67) while (6.74) becomes

$$\sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (p_i^{(1)} - p_j^{(1)}) + \gamma p^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \quad (6.75)$$

Then we can state the analog of Proposition 6.4 with a similar proof.

*Proposition 6.5*

Assume a polytropic ideal gas equation of state. The pairs  $(p_i^{(1)}, \mathbf{u}_i^{(0)})$ ,  $i \in I$ , satisfy the conditions (6.68) and (6.75) if and only if

$$\left\{ \begin{array}{l} p_i^{(1)} = p^{(1)} \text{ is independent of } i \in I, \\ \mathbf{u}_i^{(0)} = \mathbf{u}^{(0)} \text{ is independent of } i \in I. \end{array} \right. \quad (6.76)$$

Hence, again in this case, the semi-discrete HLL scheme does not behave correctly at low Mach since it verifies the constraints

$$\left\{ \begin{array}{l} \varrho_i^{(0)} = \varrho^{(0)} = \text{constant}, \\ \mathbf{u}_i^{(0)} = \mathbf{u}^{(0)} \text{ is independent of } i \in I, \end{array} \right. \quad (6.77)$$

which are not admissible.

## 6.5 An All-Mach Semi-Discrete HLL Scheme

Let us now construct a modified HLL scheme which behaves correctly at low Mach. Consider first the *barotropic case*. Here, we have to weaken the second condition (6.70):  $\mathbf{u}_i^{(0)}$  is independent of  $i \in I$ . Then we let the mass conservation equation (6.58) unchanged, but we modify the momentum conservation equation (6.59) in the following way: setting, for instance,

$$\theta_{ij} = \min \left( 1, \frac{|\mathbf{u}_i|}{c_i}, \frac{|\mathbf{u}_j|}{c_j} \right), \quad (6.78)$$

we replace (6.59) by

$$\begin{cases} \frac{d(\varrho\mathbf{u})_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \{ a_{ij,R}((\mathbf{u}_i \cdot \mathbf{n}_{ij})(\varrho\mathbf{u})_i + p_i \mathbf{n}_{ij}) \\ - a_{ij,L}((\mathbf{u}_j \cdot \mathbf{n}_{ij})(\varrho\mathbf{u})_j + p_j \mathbf{n}_{ij}) + \theta_{ij} a_{ij,L} a_{ij,R} ((\varrho\mathbf{u})_j - (\varrho\mathbf{u})_i) \} = \mathbf{0}, \end{cases} \quad (6.79)$$

whose nondimensional form reads

$$\begin{cases} \frac{d(\varrho\mathbf{u})_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \left\{ \frac{1}{M^2} [a_{ij,R} p_i - a_{ij,L} p_j] \mathbf{n}_{ij} \right. \\ \left. + a_{ij,R}(\mathbf{u}_i \cdot \mathbf{n}_{ij})(\varrho\mathbf{u})_i - a_{ij,L}((\mathbf{u}_j \cdot \mathbf{n}_{ij})(\varrho\mathbf{u})_j \\ + \theta_{ij} a_{ij,L} a_{ij,R} ((\varrho\mathbf{u})_j - (\varrho\mathbf{u})_i) \} = \mathbf{0}. \right. \end{cases} \quad (6.80)$$

At the order  $-2$  in  $M$ , we find again the condition (6.66) which is automatically satisfied in this barotropic case. On the other hand, (6.80) gives at the order  $-1$  in  $M$

$$\sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \varrho_j^{(1)} \mathbf{n}_{ij} = \mathbf{0}. \quad (6.81)$$

*Proposition 6.6*

Assume a barotropic equation of state. The pairs  $(\varrho_i^{(1)}, \mathbf{u}_i^{(0)})$ ,  $i \in I$ , satisfy conditions (6.64) and (6.81) if and only if

$$\begin{cases} \varrho_i^{(1)} = \varrho^{(1)} \text{ is independent of } i \in I, \\ \sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij} = 0 \text{ for all } i \in I. \end{cases} \quad (6.82)$$

*Proof.* As in the proof of Proposition 6.4, we obtain from (6.64)

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (\varrho_i^{(1)} - \varrho_j^{(1)})^2 + \varrho^{(0)} ((\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \varrho_i^{(1)} - (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) \varrho_j^{(1)}) \right\} = 0.$$

On the other hand, by taking the inner product of (6.81) with  $\mathbf{u}_i^{(0)}$  and summing with respect to  $i$ , we find

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| \left\{ (\mathbf{u}_i^{(0)} \cdot \mathbf{n}_{ij}) \varrho_j^{(1)} - (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \varrho_i^{(1)} \right\} = 0.$$

Together with the previous equation, it yields

$$\sum_{\Gamma_{ij}} |\Gamma_{ij}| (\varrho_i^{(1)} - \varrho_j^{(1)})^2 = 0,$$

and, using again (6.64), the conclusion follows.  $\square$

Hence, *in the barotropic case*, the modified semi-discrete HLL scheme (6.58), (6.79) has the good behavior at low Mach.

Let us pass to the case of a *general equation of state*. Here, we have to weaken the constraints (6.76). We then need to modify the mass conservation equation (6.58) that we replaced by

$$\begin{cases} \frac{d\varrho_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \{ (a_{ij,R}(\varrho\mathbf{u})_i - a_{ij,L}(\varrho\mathbf{u})_j) \cdot \mathbf{n}_{ij} \\ \quad + \theta_{ij} a_{ij,L} a_{ij,R} (\varrho_j - \varrho_i) \} = 0. \end{cases} \quad (6.83)$$

In fact, the corresponding nondimensional equation coincides with (6.83). It gives at the order 0 in  $M$

$$\frac{d\varrho_i^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \left\{ (\varrho\mathbf{u})_j^{(0)} \cdot \mathbf{n}_{ij} + \theta_{ij}^{(0)} a_{ij}^{(0)} (\varrho_i^{(0)} - \varrho_j^{(0)}) \right\} = 0. \quad (6.84)$$

We keep the modified momentum conservation equation (6.79). Its nondimensional form (6.80) gives (6.66) at the order  $-2$  in  $M$  and the condition

$$\sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| p_j^{(1)} \mathbf{n}_{ij} = \mathbf{0} \quad (6.85)$$

at the order  $-1$  in  $M$ .

Let us first restrict ourselves to the case of an equation of state of the form  $p = p(\varrho\varepsilon)$ . We then keep the energy conservation (6.60) unchanged so that conditions (6.72)–(6.74) hold. Since by (6.72) and (6.73)

$$p_i^{(0)} = p^{(0)} = \text{constant}, i \in I \quad (6.86)$$

and

$$p_i^{(1)} = \kappa^{(0)} (\varrho\varepsilon)_i^{(1)}, \quad \kappa^{(0)} = p'((\varrho\varepsilon)^{(0)}),$$

the condition (6.66) holds and (6.74) becomes

$$\sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (p_i^{(1)} - p_j^{(1)}) + \kappa^{(0)} (\varrho\varepsilon + p)^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\}. \quad (6.87)$$

*Proposition 6.7*

Assume an equation of state of the form  $p = p(\varrho\varepsilon)$ . Then, the pairs  $(p_i^{(1)}, \mathbf{u}_i^{(0)})$ ,  $i \in I$ , satisfy conditions (6.85) and (6.87) if and only if

$$\begin{cases} p_i^{(1)} = p^{(1)} \text{ is independent of } i \in I, \\ \sum_{\Gamma_{ij} \in \partial\Omega_i} |\Gamma_{ij}| \mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij} = 0 \text{ for all } i \in I. \end{cases} \quad (6.88)$$

The proof is similar to that of Proposition 6.6. Hence, if  $p = p(\varrho\varepsilon)$  the modified semi-discrete HLL scheme (6.60), (6.79), and (6.83) behaves nicely at low Mach.

So far, we have only modified both mass and momentum conservation equations. However, for a general equation of state  $p = p(\varrho, \varrho\varepsilon)$ , we need also to modify the energy conservation equation. This is clear if we observe that we have then  $p_i^{(0)} = p(\varrho_i^{(0)}, (\varrho\varepsilon)_i^{(0)})$ , and the desired property (6.86) holds if and only if  $\varrho_i^{(0)} = \varrho^{(0)} = \text{constant}$ , which is obviously undesirable except in the barotropic case. As in Sect. 6.2, we restrict ourselves to the simplest possible case where  $\chi$  and  $\kappa$  are constant so that

$$\Delta(\varrho e) = \frac{1}{\kappa}(\Delta p - \chi\Delta\varrho) + \frac{1}{2}\Delta(\varrho|\mathbf{u}|^2).$$

Then, we modify the energy conservation equation (6.60) as follows:

$$\left\{ \begin{array}{l} \frac{d(\varrho e)_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \{ a_{ij,R}(\mathbf{u}_i \cdot \mathbf{n}_{ij})(\varrho e + p)_i \\ \quad - a_{ij,L}(\mathbf{u}_j \cdot \mathbf{n}_{ij})(\varrho e + p)_j + a_{ij,L}a_{ij,R} \left( \frac{1}{\kappa}(p_j - p_i) \right. \\ \quad \left. + \frac{1}{2}((\varrho|\mathbf{u}|^2)_j - (\varrho|\mathbf{u}|^2)_i) - \frac{\chi}{\kappa}\theta_{ij}(\varrho_j - \varrho_i) \right) \} = 0. \end{array} \right. \quad (6.89)$$

The nondimensional energy conservation equation now reads

$$\left\{ \begin{array}{l} \frac{d(\varrho\varepsilon)_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial\Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \left\{ \frac{1}{M} \frac{a_{ij,L}a_{ij,R}}{\kappa} (p_j - p_i) \right. \\ \quad + a_{ij,R}(\mathbf{u}_i \cdot \mathbf{n}_{ij})(\varrho\varepsilon + p)_i - a_{ij,L}(\mathbf{u}_j \cdot \mathbf{n}_{ij})(\varrho\varepsilon + p)_j \\ \quad \left. - \frac{\chi}{\kappa}\theta_{ij}a_{ij,L}a_{ij,R}(\varrho_j - \varrho_i) \right\} = \mathcal{O}(M). \end{array} \right. \quad (6.90)$$

Combining (6.61) and (6.90) gives

$$\left\{ \begin{array}{l} \frac{dp_i}{dt} + \frac{1}{|\Omega_i|} \sum_{\Gamma_{ij} \in \partial \Omega_i} \frac{|\Gamma_{ij}|}{a_{ij,R} - a_{ij,L}} \left\{ \frac{1}{M} a_{ij,L} a_{ij,R} (p_j - p_i) \right. \\ \left. + (1 + \kappa) (a_{ij,R} (\mathbf{u}_i \cdot \mathbf{n}_{ij}) p_i - a_{ij,L} (\mathbf{u}_j \cdot \mathbf{n}_{ij}) p_j) \right\} = \mathcal{O}(M). \end{array} \right. \quad (6.91)$$

At the order  $-1$  in  $M$ , we obtain

$$\sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| a_{ij}^{(0)} (p_j^{(0)} - p_i^{(0)}) = 0,$$

which yields as usual

$$p_i^{(0)} = p^{(0)} \text{ is independent of } i \in I.$$

Next, we find at the order  $0$  in  $M$

$$\left\{ \begin{array}{l} \frac{dp^{(0)}}{dt} + \frac{1}{2|\Omega_i|} \sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (p_i^{(1)} - p_j^{(1)}) \right. \\ \left. + (1 + \kappa) p^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \end{array} \right.$$

By summing with respect to  $i$ , it yields

$$p^{(0)} = \text{constant}$$

and

$$\sum_{\Gamma_{ij} \in \partial \Omega_i} |\Gamma_{ij}| \left\{ a_{ij}^{(0)} (p_i^{(1)} - p_j^{(1)}) + (1 + \kappa) p^{(0)} (\mathbf{u}_j^{(0)} \cdot \mathbf{n}_{ij}) \right\} = 0. \quad (6.92)$$

Now, one can prove as above

*Proposition 6.8*

Assume an equation of state of the form  $p = \chi \rho + \kappa \rho \varepsilon$  where  $\chi$  and  $\kappa$  are constant. Then the pairs  $(p_i^{(1)}, \mathbf{u}_i^{(0)}), i \in I$ , satisfy conditions (6.85) and (6.87) if and only if (6.88) holds.

Hence, in this case the modified semi-discrete HLL scheme (6.79), (6.83), (6.89) has a good behavior at low Mach.

This section is indeed to be considered as an introduction to the low Mach behavior of finite-volume schemes of approximation to fluid dynamics systems. On the one hand, concerning the Euler system of gas dynamics, we have not studied the stability of the all-Mach Roe and HLL schemes nor studied the behavior of relaxation schemes. In this direction, we refer to [269–271]. On the other hand, we have not considered two-phase flows [120, 272] nor combustion [844] and neither MHD [701], although computing the respective low Mach number limits is of practical importance. Let us mention that other tools used for dealing with low Mach problems in flux dynamics are

preconditioning methods [1139] or splitting between stiff and non-stiff parts, with subsystems representing, respectively, fast acoustic wave propagation and slow advection velocity [270, 700, 907].

For schemes said in the low Mach regime, there are many interesting references among which we have mentioned only a few, and many other references can be found in these papers; see also [88, 89, 362, 403, 408, 409, 421, 575, 796, 797, 884, 885], and references therein.

Then, for the shallow water equations, the analogous situation to consider is the asymptotic behavior of a numerical scheme in the low Froude limit (see, for instance, [89, 196, 1165]).

## Notes

As we have already explained, there are few theoretical results concerning two-dimensional hyperbolic systems; besides the references already quoted in Remark 1.1, we refer essentially to the book of Majda [840]; see also the texts of Jeffrey [650], Hirsch [617], Richtmyer and Morton, Chapter 13 [974] and Chang and Hsiao [278] (which contains detailed computations concerning both the scalar two-dimensional case and steady gas dynamics), and [230]; see those of Courant and Friedrichs [371], J.D. Anderson [40], Whitham [1188], Taniuti and Nishihara [1105], for a better understanding of physical phenomena; see also the papers of Lytton [836], Métivier [866–868], Loh and Hui [830], Hui and Loh [641], Koren [705, 707], Yang and Hsu [1203], Glaz and Wardlaw [529] (for computational aspects) [565], and Glimm and Majda [535, 841].

For numerical approximation, this chapter is only a survey (by no means exhaustive) of a part of the huge literature on the subject. We mention the books of Godunov et al. [544] and Hirsch [617, 618] for the approximation of gas dynamics by finite difference schemes; for finite-volume methods, we refer to Eymard-Gallouët-Herbin [466]; for multidimensional upwinding, see, among other interesting many papers in [20], the recent paper of Roe [988] and also [970]. There are already so many references concerning numerical schemes in Sects. 4 and 3 that we dare not add more than a few: the review article of Woodward and Colella [1190] and that of Vinokur [1172], also [100, 395, 872, 982], Arminjon and Viallon [59, 60] for the extension of the Nessyahu–Tadmor scheme; Boukadia and Leroux [179] for that of the Lax–Friedrichs scheme; for comparison of several difference schemes [819]; for multigrid extensions, see Hemker and Spekreijse [604], Koren and Hemker [706], front-tracking methods Chern et al. [298], Mao [848], Davis [394]; for a Lagrangian approach, Loh and Liou [831], Dukowicz et al. [444]; positive schemes for linear advection can be found, for instance, in Hunsdorfer et al. [642] and more details concerning computational aspects in W.K. Anderson [41], Ajmani et al. [24], Jiang and Forsyth [655], [225] for adaptative mesh re-

finement; for WAF (weighted averaged flux)-type schemes, [150, 1124, 1125] and then wave propagation algorithms [775] and an analysis of the stability theory for three-dimensional algorithms [732].

We have mainly considered the application to the compressible Euler equations; these methods are also used for reactive flow [412], multimaterial flow [15, 81], incompressible flow (E and Shu [456]), Kelvin-Helmholtz instabilities (Munz and Schmidt [886]), relativistic hydrodynamics (Schneider et al. [1017], Balsara [76], Dolezal and Wong [428]), detonation waves (Clarke et al. [313]), Klein in [75], Quirk [963], [852, 853]) or solving other related problems such as magnetohydrodynamics (Brio and Wu [205], Tanaka [1101], Powell et al. [957], Dai and Woodward [386], Zachary and Colella [1214, 1215], Cargo and Gallice [235]), Balsara and coauthors [78–80]; shallow water equations (Alcrudo and Garcia-Navarro [26]), open-channel flows (Glaister [527]); petroleum reservoir simulation (Blunt and Rubin [152]); two-phase flow (Sainsaulieu [969, 1000], Abgrall and Saurel [18, 1010]) flow through porous media (Durlofsky [451, 452]).

#### *Note Added in the Second Edition*

A lot of work has been done in the last 20 years, and we only mention a few references, the book by Benzoni and Serre [112], works extending Kruzhkov's results [177, 834]; then concerning convergence or error estimates [159, 160, 229, 255–257, 275, 276, 343, 344, 465, 672, 751, 865, 1094, 1141, 1170, 1177, 1178], see [870] for a more recent survey; the related topic of a posteriori error estimate will not be considered in the present work [716, 913].

Let us also mention a few more or less original recent approaches: central upwind schemes [722], Lagrangian methods in [419], and [838] mixing cell-centered and an original solver located at the nodes for evaluating the fluxes; Multidimensional Optimal Order Detection (MOOD) [138, 311]; Discrete Duality Finite Volume Method (DDFV) [136]; Dual Mesh Gradient Reconstruction (DMGR) [137]; and Generalized Riemann Problem Solvers for scalar linear advection [143].



# VI

## An Introduction to Boundary Conditions

The aim of this chapter is to introduce the unfamiliar reader to the topic of boundary conditions: we just want to give some insight into this question and do not pretend to give an exhaustive study. We recall first the main features of the initial boundary value problem (I.B.V.P.) before we present the numerical treatment of the question.

Considerations on characteristics show that one must be cautious about prescribing the solution on the boundary. In some particular cases, the boundary conditions can be found by physical considerations (such as a solid wall), but their derivation in the general case is not obvious. The problem of finding the “correct” boundary conditions, i.e., which lead to a well-posed problem, is difficult in general from both the theoretical and practical points of view (proof of the well-posedness, choice of the physical variables that can be prescribed).

The implementation of these boundary conditions is crucial in practice; however, it depends very much on the problem, and we shall give only some examples of the most usual situations. Moreover, it mostly remains a matter for the expert, whose know-how is seldom described in detail.

### 1 The Initial Boundary Value Problem in the Linear Case

It is well known that even the simple scalar I.B.V.P. in the “quarter plane”

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, & x > 0, t > 0, \\ u(x, 0) = u_0(x), \\ u(0, t) = g(t), & t > 0, \end{cases}$$

is ill-posed in general (which means that there may be no solution or one that does not depend in a continuous way on the initial and boundary data or nonuniqueness). As soon as the initial condition  $u(x, 0) = u_0(x)$  is given, the function cannot be prescribed arbitrarily on the boundary, as we shall see below by considering the characteristics.

Many results concern the linear case and are then extended, mostly in a heuristic way, to the nonlinear case by “freezing” the Jacobian at a constant state, i.e., by linearizing around a constant state. That is why we begin by considering the linear case and even the simplest scalar linear case.

## 1.1 Scalar Advection Equations

### 1.1.1 One-Dimensional Scalar Advection Equation

We consider first the problem

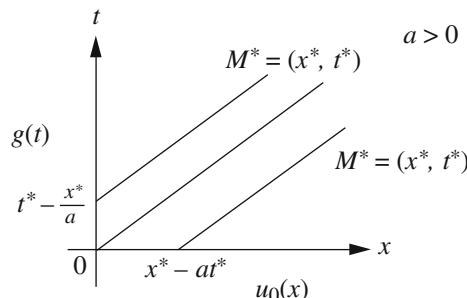
$$\begin{cases} \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0, & x > 0, t > 0, \\ u(x, 0) = u_0(x), & x \geq 0. \end{cases} \quad (1.1)$$

In this problem, one sees (Fig. 1.1) that if  $a > 0$ , the characteristics as  $t$  increases are leaving from the boundary  $x = 0$ , thus coming into the domain. Therefore, one needs to prescribe the solution on the boundary  $x = 0$ ,

$$u(0, t) = g(t), \quad t > 0, \quad (1.2)$$

where  $g$  is some given function. If  $M = (x, t)$  is any point in the domain  $\mathbb{R}_+^* \times \mathbb{R}_+^*$ , the value of  $u$  at  $M$  is then uniquely determined. The solution  $u$  of (1.1) and (1.2) is then given by

$$\begin{cases} u(x, t) = u_0(x - at) & \text{if } x - at > 0, \\ u(x, t) = g\left(t - \frac{x}{a}\right) & \text{if } x - at < 0. \end{cases}$$



**Fig. 1.1** I.B.V.P. for the one-dimensional scalar advection equation

The resulting solution is  $C^1$  if the initial and boundary data are  $C^1$  and satisfy the compatibility relations

$$u_0(0) = g(0), \quad u'_0(0) = -\frac{g'(0)}{a}.$$

If this does not hold, we have a weak solution satisfying the Rankine–Hugoniot jump condition on each side of the discontinuity  $x = at$ .

On the other hand, if  $a < 0$ , the characteristics are outgoing from the interior of the domain and impinging on the boundary; the information is thus carried from the given initial data  $u_0$ , and one cannot specify the solution on the boundary. The solution is

$$u(x, t) = u_0(x - at), \quad x \geq 0, \quad t \geq 0,$$

and in particular

$$u(0, t) = u_0(-at).$$

Note that in the trivial case  $a = 0$ , the characteristics are vertical and no boundary condition is needed, as for outgoing characteristics.

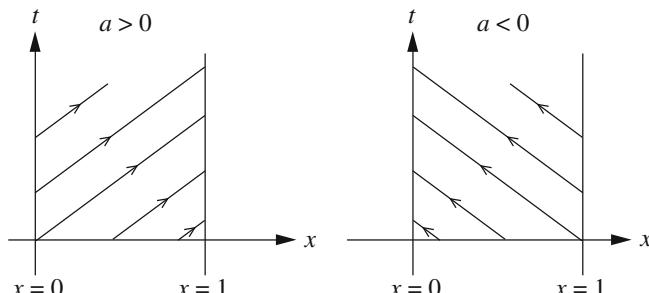
Combining both cases enables one to treat the case of a bounded interval in space, for instance, the strip  $0 < x < 1$ . The boundary condition that must be specified corresponds to incoming characteristics (see Fig. 1.2)

$$\begin{aligned} u(0, t) &= g(t), & t > 0, \text{ if } a > 0, \\ u(1, t) &= h(t), & t > 0, \text{ if } a < 0. \end{aligned}$$

### 1.1.2 Two-Dimensional Scalar Advection Equation

The solution of the pure Cauchy problem

$$\begin{cases} \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} + b \frac{\partial u}{\partial y} = 0, & (x, y, t) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+, \\ u(x, y, 0) = u_0(x, y), & x, y \in \mathbb{R} \end{cases} \quad (1.3)$$



**Fig. 1.2** One-dimensional scalar advection equation in a strip

is

$$u(x, y, t) = u_0(x - at, y - bt),$$

and is constant on the characteristic lines  $x - at = \text{const.}$ ,  $y - bt = \text{const.}$  (see Chap. V, Sect. 1.2, Remark 1.2). The advection direction is  $\mathbf{C} = (\mathbf{c}, 1)$ ,  $\mathbf{c} = (a, b)^T$ .

For an I.B.V.P., the independent variables  $(x, y, t)$  belong to a domain  $Q = \mathcal{O} \times \mathbb{R}_+^*$  of  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$ , with boundary  $\Sigma$ ;  $Q$  is a cylinder, and there are two different kinds of data given on the surface  $\Sigma$ :

- (i) initial data on the set  $\mathcal{O}$  of the plane  $t = 0$ ,
- (ii) boundary data on the remaining part  $\Gamma$  of  $\Sigma$  ( $\Gamma$  is the side of the cylinder):  
 $\Gamma = \partial\mathcal{O} \times \mathbb{R}_+^*$ . On this surface  $\Gamma$ ,  $n_t = 0$ , and one says that “the boundary of  $\mathcal{O}$  is characteristic” at a point if

$$an_x + bn_y = \mathbf{c} \cdot \mathbf{n} = 0$$

at this point, where  $\mathbf{n} = (n_x, n_y)^T$  is the outward normal to  $\partial\mathcal{O}$  in the plane  $t = 0$  (see Chap. V, Sect. 1.2).

We first detail the “half-space problem,” for which the computations are explicit. We have  $\mathcal{O} = \{(x, y)/x > 0, y \in \mathbb{R}\}$  and  $Q = \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}_+^*$ ; therefore,  $\Sigma$  is made of two half planes:  $t = 0$ ,  $x \geq 0$  on which the initial data are given and  $\Gamma = \{x = 0, t > 0\}$ . Let us follow the same arguments as in the one-dimensional case. Given any point  $M^* = (x^*, y^*, t^*)$  in  $Q$ , one introduces the characteristic line through  $M^*$ :  $t \rightarrow (x = x(M^*, t), y = y(M^*, t))$ , which is defined by

$$x - at = x^* - at^*, \quad y - bt = y^* - bt^*. \quad (1.4)$$

The solution is given by the value at the point where the characteristic line, on which  $u$  is constant, intersects the boundary  $\Sigma$ : if it intersects the plane  $t = 0$ , it is determined by the initial condition, whereas if it intersects the plane  $x = 0$ , it is given by the boundary data. The line (1.4) intersects the boundary  $x = 0$  at a point corresponding to the time

$$t_0 = t^* - \frac{x^*}{a}.$$

- (i) Assume first  $a > 0$ . If  $x^* > at^*$ , then  $t_0 < 0$ ; but the line (1.4) intersects the boundary  $t = 0$  at point  $(x_0, y_0, 0)$ , with  $x_0 = x^* - at^*$ ,  $y_0 = y^* - bt^*$ , and the solution is given by the initial data

$$u(x^*, y^*, t^*) = u(x^* - at^*, y^* - bt^*, 0) = u_0(x^* - at^*, y^* - bt^*).$$

Otherwise  $0 \leq t_0 < t^*$  if  $0 \leq x^* < at^*$ , and then the intersection point  $M_0 = (0, y_0, t_0)$  belongs to  $\Gamma = \{x = 0, t \geq 0\}$ , and  $u$  is given by

$$u(x^*, y^*, t^*) = u(0, y_0, t_0) = u\left(0, y^* - b(t^* - t_0), t^* - \frac{x^*}{a}\right),$$

which shows that one needs to prescribe the solution on the boundary  $x = 0$  where the characteristics are “incoming,”

$$u(0, y, t) = g(y, t), \quad t > 0,$$

and

$$u(x^*, y^*, t^*) = g\left(y^* - b\frac{x^*}{a}, t^* - \frac{x^*}{a}\right).$$

The solution  $u$  is then uniquely determined in the whole domain (see Fig. 1.2).

- (ii) If  $a < 0$ , the characteristic line intersects the boundary  $x = 0$  at time  $t_0 > T$  and the boundary  $t = 0$  at  $x_0 = x^* - at^*$ ,  $y_0 = y^* - bt^*$ . Therefore, one cannot specify the solution on the boundary  $x = 0$ ; it is thoroughly determined by the initial data. In fact, the same thing occurs if  $a = 0$ , which means that if the boundary is characteristic, we consider it as part of the “outgoing” boundary.

In short, one has to prescribe the boundary data on the “incoming” part of  $\Gamma : \Gamma_- = \partial\mathcal{O}_- \times \mathbb{R}^+$ , where  $\partial\mathcal{O}_- = \{(x, y) \in \partial\mathcal{O}, \mathbf{c} \cdot \mathbf{n} < 0\}$ . In this particular case, the outward normal to  $\partial\mathcal{O}$  is  $\mathbf{n} = (-1, 0)^T$ , and  $\Gamma_- = \Gamma = \{x = 0, t > 0\}$  if  $a > 0$  and is empty if  $a < 0$ .

More generally, consider now a bounded domain  $\mathcal{O}$  of  $\mathbb{R}^2$ . In order to know whether the solution can indeed be defined at a point  $(x^*, y^*, t^*) \in Q = \mathcal{O} \times \mathbb{R}^{+*}$ , one draws as above the characteristic line (1.4) and looks for the intersection with the boundary  $\Sigma$  of  $Q$ . If the line remains in  $Q$  and intersects the plane  $t = 0$  at a point  $x_0 = x^* - at^*$ ,  $y_0 = y^* - bt^*$  that lies inside  $\mathcal{O}$ , then the solution  $u$  is determined by the initial condition

$$u(x^*, y^*, t^*) = u_0(x_0, y_0) = u_0(x^* - at^*, y^* - bt^*).$$

In the other case, the line intersects  $\Gamma$  at some point  $(x_0, y_0, t_0)$  with time  $t_0$  satisfying  $0 \leq t_0 < t^*$ . On the one hand, the points are on the same characteristic line, which yields

$$(x^* - x_0, y^* - y_0)^T = (t^* - t_0)\mathbf{c}.$$

On the other hand, since  $(x^*, y^*) \in \mathcal{O}$  and  $(x_0, y_0) \in \partial\mathcal{O}$ , we have (provided  $\mathcal{O}$  is not characteristic at  $m_0 = (x_0, y_0)$ )

$$(x^* - x_0, y^* - y_0) \cdot \mathbf{n} < 0,$$

where  $\mathbf{n}$  is the outward normal to  $\mathcal{O}$  in the  $(x, y)$ -plane (see Fig. 1.3).

Therefore, boundary data have to be prescribed on the part  $\Gamma_-$  of the boundary that corresponds to incoming characteristics

$$\partial\mathcal{O}_- = \{(x, y) \in \partial\mathcal{O}, \mathbf{c} \cdot \mathbf{n}(x, y) < 0\},$$

$$u(\cdot, t) = g(\cdot, t) \text{ on } \partial\mathcal{O}_- \iff u = g \text{ on } \Gamma_- = \partial\mathcal{O}_- \times \mathbb{R}_+,$$

and not on the part  $\partial\mathcal{O}_+ = \{(x, y) \times \partial\mathcal{O}, \mathbf{c} \cdot \mathbf{n}(x, y) \geq 0\}$  where they are outgoing. Note that if  $\mathcal{O}$  is characteristic at  $m_0 = (x_0, y_0)$ , it is easily seen that  $u$  cannot be specified on the corresponding part of  $\Gamma$ .

Denoting by  $t_0 = t_0(M^*)$  the time when the characteristic intersects  $\Sigma$ ,  $t_0 = \inf\{t \geq 0 / (x(M^*, t), y(M^*, t), t) \in \bar{\mathcal{Q}}\}$ , and by  $(x_0, y_0)$  the coordinates of the intersection point, we have

$$\begin{aligned} u(x^*, y^*, t^*) &= u_0(x_0, y_0) = u_0(x^* - at^*, y^* - bt^*) \\ &\quad \text{if } t_0(x^*, y^*, t^*) = 0, \\ u(x^*, y^*, t^*) &= g(x_0, y_0, t_0) = g(x^* - a(t^* - t_0), y^* - b(t^* - x_0), t_0), \\ &\quad \text{if } t_0(x^*, y^*, t^*) > 0, \end{aligned}$$

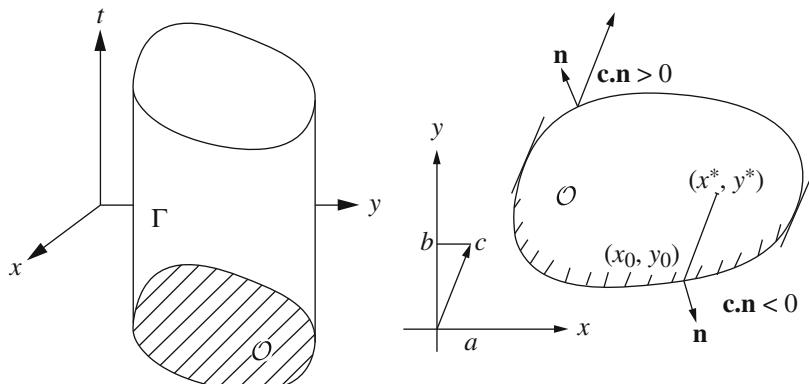
since  $t_0 > 0$  implies  $x(M^*, t_0), y(M^*, t_0) \in \partial\mathcal{O}_-$  and  $(x(M^*, t_0), y(M^*, t_0), t_0) \in \Gamma_-$ .

The resulting solution is  $C^1$  if the data, which are assumed to be smooth enough, satisfy some compatibility relations. In particular,  $u$  is continuous on both sides of the surface  $S$  (corresponding to the line  $x = at$  in the one-dimensional case; see Fig. 1.1) defined by  $S = \{M^*/(x(M^*, 0), y(M^*, 0)) \in \partial\mathcal{O}_-\}$ , where  $(x(M^*, 0), y(M^*, 0))$  is the intersection of the backward characteristic (1.4) through  $M^*$  with the plane  $t = 0$ .  $S$  is an oblique cylinder (with base  $\partial\mathcal{O}_-$ ) of direction along the vector  $\mathbf{C} = (\mathbf{c}, 1)$ . Therefore, the continuity of  $u$  across  $S$  supposes

$$u_0(x, y) = g(x, y, 0) \text{ on } \partial\mathcal{O}_-.$$

Otherwise, we get a weak solution: since the equation is linear, the Rankine–Hugoniot jump relation across  $S$  is simply (see the Chap. I, Sect. 4, formula (4.8))

$$an_x + bn_y + n_t = 0,$$



**Fig. 1.3** Boundary corresponding to incoming characteristics

where  $(n_x, n_y, n_t)^T$  is a normal vector to  $S$ . By definition of  $S$ , this is indeed satisfied.

*Remark 1.1.* The arguments can easily be extended to a scalar problem with  $C^1$  variable coefficients  $\mathbf{c}(x, y, t) = (a(x, y, t), b(x, y, t))^T$  that are Lipschitz w.r.t space variables. Only the characteristics through  $M^* = (X^*, t^*)$ , defined as the integral curves  $t \mapsto \mathbf{X}(t; \mathbf{X}^*, t^*) = (x(M^*, t), y(M^*, t))$  of

$$\begin{cases} \frac{d\mathbf{X}}{dt} = \mathbf{c}(\mathbf{X}(t), t), \\ \mathbf{X}(t^*; \mathbf{X}^*, t^*) = \mathbf{X}^*, \end{cases}$$

are no longer straight lines. One considers as above the “entrance” time  $t_0 = t_0(\mathbf{X}^*, t^*)$  corresponding to the point where the characteristic enters the domain and then follows the same lines as above, depending on whether  $t_0 = 0$  or  $t_0 > 0$ . The boundary data are still prescribed on the incoming part:  $\Gamma_- = \partial\mathcal{O}_- \times (0, T)$ ,  $\partial\mathcal{O}_- = \{(x, y, t) \in \partial\mathcal{O} \times (0, T), \mathbf{c}(x, y, t) \cdot \mathbf{n}(x, y) < 0\}$ ,

$$u = g \text{ on } \Gamma_-.$$

We leave the technical details to the reader.  $\square$

## 1.2 One-Dimensional Linear Systems. Linearization

Consider now a linear hyperbolic system in *diagonal* form,

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \boldsymbol{\Lambda} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, & 0 < x < 1, t > 0, \\ \mathbf{u}(x, 0) = \mathbf{u}_0(x), \end{cases}$$

where  $\mathbf{u} \in \mathbb{R}^p$  and  $\boldsymbol{\Lambda}$  is a diagonal matrix with eigenvalues  $a_i, 1 \leq i \leq p$ . At first, we assume that the  $a_i$  are nonvanishing:  $a_i \neq 0$ ,  $1 \leq i \leq p - q$ ,  $a_i > 0, p - q + 1 \leq i \leq p$ . One writes

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^I + \boldsymbol{\Lambda}^{II} = \boldsymbol{\Lambda}^+ + \boldsymbol{\Lambda}^-,$$

where, using the notations (1.9) in Chap. IV, Sect. 1.1, the matrix  $\boldsymbol{\Lambda}^I \equiv \text{diag}(a_i^+) = \boldsymbol{\Lambda}^+$  (resp.  $\boldsymbol{\Lambda}^{II} \equiv \text{diag}(a_i^-) = \boldsymbol{\Lambda}^-$ ) has  $q$  positive (resp.  $p - q$  negative) eigenvalues, and the corresponding partition for  $\mathbf{u}$  is  $\mathbf{u} = (\mathbf{u}^I, \mathbf{u}^{II})^T \in \mathbb{R}^q \times \mathbb{R}^{p-q}$  (incoming/outgoing dependent variables). The results of Sect. 1.2 lead to prescribing the following boundary conditions:

$$\mathbf{u}^I(0, t) = \mathbf{g}^I(t), \quad \mathbf{u}^{II}(1, t) = \mathbf{g}^{II}(t),$$

which means that one solves  $p$  uncoupled scalar equations. If one eigenvalue vanishes, the corresponding component of  $\mathbf{u}$  is constant, determined by the initial data on  $t = 0$ . Thus, it should be considered as an outgoing variable, i.e., part of  $\mathbf{u}^{II}$  on  $x = 0$  and of  $\mathbf{u}^I$  on  $x = 1$ .

A slight generalization consists in coupling these boundary conditions by introducing the already known components corresponding to the outgoing characteristics impinging on the boundary (which transport the information from the data given on the line  $t = 0$ ). Thus, one can set

$$\begin{cases} \mathbf{u}^I(0, t) = \mathbf{S}^I \mathbf{u}^{II}(0, t) + \mathbf{g}^I(t), \\ \mathbf{u}^{II}(1, t) = \mathbf{S}^{II} \mathbf{u}^I(1, t) + \mathbf{g}^{II}(t), \end{cases} \quad (1.5)$$

where  $\mathbf{S}^I$  (resp.  $\mathbf{S}^{II}$ ) is  $q \times (p - q)$  (resp.  $(p - q) \times q$ ) matrix, and the solution is still uniquely determined; (1.5) means that  $\mathbf{u}^I$  is an affine function of  $\mathbf{u}^{II}$  on  $x = 0$  (“reflection of the outgoing waves”) and conversely  $\mathbf{u}^{II}$  is an affine function of  $\mathbf{u}^I$  at  $x = 1$ . In fact, this type of boundary condition leads in the case of linear systems with variable coefficients to the theory of well-posed systems in the sense of Kreiss (see Kreiss [71]).

Now, for a *general linear* hyperbolic system with constant coefficients,

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0}, & 0 < x < 1, t > 0, \\ \mathbf{u}(x, 0) = \mathbf{u}_0(x), \end{cases} \quad (1.6)$$

one diagonalizes the matrix  $\mathbf{A}$  (see Chap. IV, Sect. 1.1)

$$\mathbf{A} = \mathbf{T} \Lambda \mathbf{T}^{-1}.$$

We recall that  $\mathbf{w} = \mathbf{T}^{-1} \mathbf{u}$  are the characteristic variables:

$$\mathbf{u} = \sum_i \mathbf{w}_i \mathbf{r}_i, \quad \mathbf{w}_i = \mathbf{l}_i^T \mathbf{u}, \quad (1.7)$$

where the  $\mathbf{r}_i$  (resp.  $\mathbf{l}_i$ ) are the eigenvectors of  $\mathbf{A}$  (resp.  $\mathbf{A}^T$ ). For ease of notation, we set  $p' =$  number of nonpositive eigenvalues of  $\mathbf{A}$  ( $a_i \leq 0, 1 \leq i \leq p'$ ) and  $q = p - p' =$  number of positive eigenvalues of  $\mathbf{A}$  ( $a_i > 0, p' + 1 \leq i \leq p$ ). Let the subscript  $I$  (resp.  $II$ ) correspond to positive eigenvalues  $a_i > 0$  (resp. negative  $a_i \leq 0$ ), and set

$$\mathbf{w}^I = (w_{p'+1}, \dots, w_p), \quad \mathbf{w}^{II} = (w_1, \dots, w_{p'})^T.$$

Therefore,  $\mathbf{w} = \mathbf{T}^{-1} \mathbf{u}$  is a solution of a decoupled system

$$\frac{\partial \mathbf{w}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{w}}{\partial x} = \mathbf{0},$$

which is well-posed if the boundary conditions for  $\mathbf{w} = (\mathbf{w}^I, \mathbf{w}^{II})^T \in \mathbb{R}^{p-p'} \times \mathbb{R}^{p'}$  can take the above form, i.e., are specified at  $x = 0$  for  $p' + 1 \leq i \leq p$ ,

$$\mathbf{w}^I(0, t) = \mathbf{g}^I(t), \quad (1.8a)$$

where  $\mathbf{g}^I(t)$  is a given  $(p - q')$ -component vector function.

If we consider the strip  $0 < x < 1$ , we set  $p'' =$  the number of negative eigenvalues of  $\mathbf{A}$  ( $a_i < 0$ , for  $1 \leq i \leq p''$ ), and at  $x = 1$

$$\mathbf{w}^I = (w_{p''+1}, \dots, w_p), \quad \mathbf{w}^{II} = (w_1, \dots, w_{p''})^T,$$

the boundary data take the form

$$\mathbf{w}^{II}(1, t) = \mathbf{g}^{II}(t),$$

where  $\mathbf{g}^{II}(t)$  is a given  $(p''\text{-component vector})$  function; we can take more generally

$$\begin{cases} \mathbf{w}^I(0, t) = \mathbf{S}^I \mathbf{w}^{II}(0, t) + \mathbf{g}^I(t), \\ \mathbf{w}^{II}(1, t) = \mathbf{S}^{II} \mathbf{w}^I(1, t) + \mathbf{g}^{II}(t). \end{cases} \quad (1.8b)$$

Now, if we are given boundary data at  $x = 0$  in the form

$$\mathbf{E}\mathbf{u}(0, t) = \mathbf{g}(t), \quad (1.9)$$

where  $\mathbf{E}$  is a  $N \times p$  matrix and  $\mathbf{g}$  is an  $N$ -component given function, we can ask whether the corresponding I.B.V.P. is well-posed. In terms of characteristic variables, the relation

$$\mathbf{ET}\mathbf{w} = \mathbf{g}, \quad \text{at } x = 0$$

can be decomposed in blocks corresponding to the partition  $\mathbf{w} = (\mathbf{w}^I, \mathbf{w}^{II})^T \in \mathbb{R}^q \times \mathbb{R}^{p-q}$  ( $q$  is the number of positive eigenvalues):

$$\mathbf{ET}\mathbf{w} = (\mathbf{ET})^I \mathbf{w}^I + (\mathbf{ET})^{II} \mathbf{w}^{II} = \mathbf{g} \text{ at } x = 0,$$

where  $(\mathbf{ET})^I$  (resp.  $(\mathbf{ET})^{II}$ ) is a  $N \times q$  (resp.  $N \times (p - q)$ ) matrix. Since  $\mathbf{w}^{II}(0, t)$  is given by the initial data, it reads

$$(\mathbf{ET})^I \mathbf{w}^I = \mathbf{g} - (\mathbf{ET})^{II} \mathbf{w}^{II}.$$

Hence, the problem is well-posed iff one can compute  $\mathbf{w}^I$ , which supposes first that  $N = q$ , so that  $(\mathbf{ET})^I$  is a square  $q \times q$  matrix. Then, it is easily seen that

$$(\mathbf{ET})^I = \mathbf{E}(\mathbf{T})^I,$$

where the columns of  $\mathbf{T}^I$  are the  $q$  eigenvectors of  $\mathbf{A}$  corresponding to positive eigenvalues. The condition is therefore given in the following lemma.

*Lemma 1.1*

Consider the boundary condition (1.9) for the system (1.6). The resulting problem is well-posed if  $\mathbf{E}$  is a  $q \times p$  matrix such that  $\mathbf{E}(\mathbf{T})^I$  is invertible, where  $\mathbf{T}^I$  denotes the  $p \times q$  matrix with columns the  $q$  eigenvectors of  $\mathbf{A}$  corresponding to positive eigenvalues.

We shall give below (Sect. 3.1) an example of such a situation in gas dynamics.

For a *nonlinear system*, one can linearize about a constant state (“freezing” theory) and apply the above procedure to the linearized system. We shall give another approach later.

In short, the number of boundary conditions should be equal to the number of incoming characteristics, i.e., pointing into the region, an argument that we have already used when introducing the Lax entropy condition (Chap. II, Sect. 5.2).

### 1.3 Multidimensional Linear Systems

Consider the “half-space” model I.B.V.P.

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} + \sum_{j=1}^{d-1} \mathbf{B}_j \frac{\partial \mathbf{u}}{\partial y_j} = \mathbf{0}, \quad x > 0, \quad y \in \mathbb{R}^{d-1}, \quad t > 0.$$

To simplify the presentation of the theory, which is already rather complicated, we shall mainly restrict ourselves to the two-dimensional constant coefficient case

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial y} + \mathbf{B} \frac{\partial \mathbf{u}}{\partial y} = \mathbf{0}, \quad x > 0, y \in \mathbb{R}, t > 0, \quad (1.10a)$$

$$\mathbf{u}(x, y, 0) = \mathbf{u}_0(x, y), \quad (1.10b)$$

$$\mathbf{E}\mathbf{u}(0, y, t) = \mathbf{g}(y, t), \quad (1.10c)$$

where  $\mathbf{E}$  is a  $q \times p$  matrix and  $\mathbf{g}$  a given  $q$ -component function. The system is hyperbolic if for any  $(\xi, \eta) \in \mathbb{R}^2$ ,  $\mathbf{A}\xi + \mathbf{B}\eta$  has real eigenvalues and is diagonalizable. We assume that the boundary  $x = 0$  of the space domain is noncharacteristic, i.e.,  $\mathbf{A}$  is invertible (see Chap. V, Sect. 1.2). Hence, with a possible change of variable, we can assume that  $\mathbf{A}$  is diagonal  $\mathbf{A} = \Lambda = (\Lambda^I, \Lambda^{II})$ . Following the considerations of the scalar case, a necessary condition for the I.B.V.P. to be well posed is that  $q$  boundary conditions are prescribed, where  $q = \text{number of positive eigenvalues of } \Lambda$ , i.e.,  $\mathbf{u}^I(0, y, t) = \mathbf{g}^I(y, t)$  or  $\mathbf{u}^I(0, y, t) = \mathbf{S}^I \mathbf{u}^{II}(0, y, t) + \mathbf{g}^I(y, t)$ . But this is not sufficient, and there are examples of solutions of the corresponding I.B.V.P. with arbitrary growth in time (see Gustafsson and Kreiss [572], Higdon [615], Yee [1204]), and more restrictions may be necessary.

In fact, the theory in the multidimensional case is not straightforward. We just want to sketch the main lines of the arguments because of their link to topics already developed such as Fourier modes, group velocity, and characteristic surfaces. This is only an introduction to the topic, and most details are skipped; for a precise study, we refer to [615].

The system is assumed to be either strictly hyperbolic or symmetric hyperbolic; for simplicity, we again assume that the boundary  $x = 0$  of the space domain is noncharacteristic, i.e.,  $\mathbf{A}$  is invertible.

### 1.3.1 Uniform Kreiss Condition (U.K.C.)

The theory developed by Kreiss relies on “normal mode analysis.” Normal modes are elementary waves  $\mathbf{u}(x, y, t) = \varphi(x)e^{iny-st}$ ,  $s \in \mathbb{C}$ , which are introduced as follows. Applying a Laplace–Fourier transform, i.e., a Fourier transform in the  $y$ -variable and a Laplace transform in the time variable  $t$  ( $u \mapsto L(u)(s) = \int_0^\infty e^{st}u(t)dt$ ,  $\operatorname{Re}(s) < 0$ ) to the system (1.10a), gives

$$\begin{cases} -s\tilde{\mathbf{u}} + \mathbf{A}\frac{\partial\tilde{\mathbf{u}}}{\partial x} + i\eta\mathbf{B}\tilde{\mathbf{u}} = \mathbf{0}, \\ \tilde{\mathbf{u}} = \tilde{\mathbf{u}}(x, \eta, s) = \int_{\mathbb{R}} \int_0^\infty e^{-i\eta y} e^{st} \mathbf{u}(x, y, t) dy dt, \end{cases} \quad (1.11)$$

which can also be written as

$$\frac{\partial\tilde{\mathbf{u}}}{\partial x} = \mathbf{A}^{-1}(s\mathbf{I} - i\eta\mathbf{B})\tilde{\mathbf{u}} = \mathbf{D}(\eta, s)\tilde{\mathbf{u}},$$

where we have set

$$\mathbf{D}(\eta, s) = \mathbf{A}^{-1}(s\mathbf{I} - i\eta\mathbf{B}). \quad (1.12)$$

By the inverse transform, we see that the solution  $\mathbf{u}$  is a “superposition” of elementary modes  $\hat{\mathbf{u}}e^{\ell x}e^{iny-st}$ , where  $\ell$  is an eigenvalue of  $\mathbf{D}(\eta, s)$ . We can immediately see that  $\operatorname{Re}(\ell) \neq 0$ .

*Lemma 1.2*

Assume  $\operatorname{Re}(s) < 0$  and  $\eta \in \mathbb{R}$ . The matrix  $\mathbf{D}(\eta, s)$  defined by (1.12) has no purely imaginary eigenvalue.

*Proof.* We write

$$\mathbf{D}(\eta, s)\varphi = i\xi\varphi \iff i(\xi\mathbf{A} + \eta\mathbf{B})\varphi = s\varphi.$$

Now, for  $\mathbf{k} = (\xi, \eta)^T \in \mathbb{R}^2$ , the matrix  $\xi\mathbf{A} + \eta\mathbf{B}$  has real eigenvalues and a complete set of eigenvectors because the system is hyperbolic. Therefore

$$\xi \in \mathbb{R} \implies s \text{ purely imaginary } (s = i\omega, \omega \in \mathbb{R}, \text{ and } \operatorname{Re}(s) = 0),$$

which gives the result.  $\square$

Thus, consider a normal mode  $\mathbf{u}(x, y, t) = \varphi(x)e^{i\eta y - st}$ , with  $s \in \mathbb{C}$ , which, once substituted in (1.10), is a solution of an ordinary differential equation for the amplitude function  $\varphi$ ,

$$-s\varphi + \mathbf{A}\varphi' + i\eta\mathbf{B}\varphi = \mathbf{0},$$

or

$$\varphi' = \mathbf{A}^{-1}(s\mathbf{I} - i\eta\mathbf{B})\varphi = \mathbf{D}(\eta, s)\varphi.$$

The main idea in order to derive necessary conditions on the boundary data so that the problem is well-posed is to exclude the cases that can lead to an ill-posed problem.

First, looking for particular normal modes (those giving rise to solutions of the form  $\mathbf{u}_\alpha(x, y, t) = \varphi_\alpha(x)e^{i\alpha\eta y - \alpha st}$  that cannot satisfy an energy estimate) yields that the problem is ill-posed if, for some  $\eta$ , the o.d.e. has an “eigenvalue”  $s$  with  $\operatorname{Re}(s) < 0$ .

Then, for those  $s$  with  $\operatorname{Re}(s) < 0$  that yield an ill-posed problem, using the fact that the set of solutions of the o.d.e. is spanned by  $p$  independent solutions  $\varphi_j(x) = \hat{\varphi}_j e^{\ell_j x}$ ,  $\ell_j$  an eigenvalue and  $\hat{\varphi}_j$  a corresponding eigenvector of  $\mathbf{D}(\eta, s)$  (if  $\ell_j$  is simple; otherwise  $\varphi_j(x)$  is multiplied by a polynomial in  $x$ ), one shows that  $q$  among them have finite norm, say  $\varphi_1, \dots, \varphi_q$ . These functions  $\varphi_i$  correspond to the  $q$  eigenvalues with negative real part of the matrix  $\mathbf{D}(\eta, s)$  (see Lemma 1.2).

We want, moreover, to prevent the possibility of such solutions satisfying the boundary conditions. Denoting by  $\mathbf{u}_j = \varphi_j(x)e^{i\eta y - st} = \hat{\varphi}_j e^{\ell_j x + i\eta y - st}$  the corresponding solutions of (1.10a), this eventually results in the necessary condition for well-posedness, which reads

$$\mathbf{E}[\mathbf{u}_1(0), \dots, \mathbf{u}_q(0)] \text{ nonsingular.}$$

(otherwise for  $\mathbf{g} = 0$  in (1.10c), we would have a solution of the I.B.V.P. that does not satisfy an energy estimate). This condition can be written equivalently as

$$\mathbf{E}[\varphi_1(x=0), \dots, \varphi_q(x=0)] \text{ nonsingular}$$

Setting

$$\mathbf{N}(\eta, s) = \mathbf{E}[\varphi_1(0), \dots, \varphi_q(0)], \quad (1.13)$$

$\mathbf{N}(\eta, s)$  is a square  $q \times q$  matrix, and this in turn is equivalent to

$$\det \mathbf{N}(\eta, s) \neq 0, \quad \forall \eta \in \mathbb{R}, \forall s, \operatorname{Re}(s) < 0.$$

This is a necessary condition for well-posedness. The sufficient “uniform Kreiss condition” (U.K.C.) is written in a very similar way,

$$\det \overline{\mathbf{N}}(\eta, s) \geq \delta > 0, \quad \forall \eta \in \mathbb{R}, \forall s, \operatorname{Re}(s) < 0$$

(where  $\bar{\mathbf{N}}$  is obtained as was  $\mathbf{N}$ , only after some normalization of the eigenfunctions), but the proof of sufficiency is not easy.

### 1.3.2 The Characteristic Manifold

We would like to interpret the above U.K.C. condition in a more intuitive way involving “incoming” and “outgoing” notions. Following [615], we introduce the following definition.

*Definition 1.1*

The characteristic manifold for system (1.10a) is the set of points  $(\xi, \eta, \omega) \in \mathbb{R}^3$ , such that  $\det(-\omega\mathbf{I} + \xi\mathbf{A} + \eta\mathbf{B}) = 0$ .

In view of Definition 1.2, Chap. V, the characteristic manifold is the set of  $(\xi, \eta, \omega)$  such that the plane  $\{(x, y, t), \omega t = \xi x + \eta y\}$  is characteristic. Therefore,  $\omega$  is an eigenvalue of the “principal symbol”  $\xi\mathbf{A} + \eta\mathbf{B}$ , which is real since the system is hyperbolic. Note that setting  $\mathbf{n} = (-\omega, \mathbf{k}), \mathbf{k} = (\xi, \eta)^T$ , we have

$$i\mathbf{A}(\mathbf{D}(\eta, i\omega) - i\xi\mathbf{I}) = -\omega\mathbf{I} + \eta\mathbf{B} + \xi\mathbf{A} = \mathbf{M}(\mathbf{n}),$$

where the matrix  $\mathbf{M}$  is defined by (1.6), Chap. V. Let us give an example.

*Example 1.1.* Consider the isentropic Euler equations (see Chap. V, Sect. 1.1, Example 1.1; see also Sect. 2.4), linearized about a state  $(\rho, u, v)$ ; denoting by  $(\rho', u', v')$  the perturbations, they are a solution of the linear system

$$\begin{aligned} \frac{\partial \rho'}{\partial t} + \rho \frac{\partial u'}{\partial x} + u \frac{\partial \rho'}{\partial x} + \rho \frac{\partial v'}{\partial y} + v \frac{\partial \rho'}{\partial y} &= R_1, \\ \frac{\partial u'}{\partial t} + u \frac{\partial u'}{\partial x} + \left(\frac{c^2}{\rho}\right) \frac{\partial \rho'}{\partial x} + v \frac{\partial u'}{\partial y} &= R_2, \\ \frac{\partial v'}{\partial t} + u \frac{\partial v'}{\partial x} + \left(\frac{c^2}{\rho}\right) \frac{\partial \rho'}{\partial y} + v \frac{\partial v'}{\partial y} &= R_3, \end{aligned}$$

where the terms  $R_i$  do not contain derivatives of  $(\rho, u, v)$ . This system can be symmetrized, defining the new dependent variables by  $(u', v', (\frac{c}{\rho})\rho')$ . The characteristic manifold is the union of the plane

$$\omega = \mathbf{u} \cdot \mathbf{k} \quad \text{if } \mathbf{k} = (\xi, \eta)^T$$

and the cones  $C$  defined by

$$\omega = \mathbf{u} \cdot \mathbf{k} \pm c|\mathbf{k}|$$

(here  $\mathbf{u} = (u, v)^T$  is the constant velocity of the state at which the system is linearized and  $c$  the speed of sound).  $\square$

*Example 1.2. Maxwell's equations in empty space.* Consider the system

$$\frac{\partial}{\partial t}(\varepsilon_0 \mathbf{E}) - \frac{1}{\mu_0} \operatorname{curl} \mathbf{B} = -J \text{ (Ampère's law),} \quad (1.14a)$$

$$\frac{\partial \mathbf{B}}{\partial t} + \operatorname{curl} \mathbf{E} = 0 \text{ (Faraday's law),} \quad (1.14b)$$

$$\operatorname{div}(\varepsilon_0 \mathbf{E}) = \rho \text{ (Poisson's or Gauss' law),} \quad (1.14c)$$

$$\operatorname{div} \mathbf{B} = 0, \quad (1.14d)$$

where  $\mathbf{B}$  denotes the magnetic field,  $\mathbf{E}$  the electric field,  $\rho$  the charge density, and  $J$  the current density. In a vacuum, we assume constant permittivity  $\varepsilon_0$  and permeability  $\mu_0$ , and  $c^2 = \frac{1}{\varepsilon_0 \mu_0}$  is the speed of light. If we take the divergence of Eq. (1.14a), we get

$$\frac{\partial}{\partial t} \operatorname{div}(\varepsilon_0 \mathbf{E}) = -\operatorname{div} J,$$

so that with Eq. (1.14c)

$$\frac{\partial \rho}{\partial t} + \operatorname{div} J = 0,$$

which is the equation of conservation of the total charge. If we take the divergence of (1.14b), we get

$$\frac{\partial}{\partial t} (\operatorname{div} \mathbf{B}) = 0,$$

so that if we assume that

$$\operatorname{div} \mathbf{B}_0 = 0$$

and

$$\operatorname{div}(\varepsilon_0 \mathbf{E}_0) = \rho_0$$

at some initial time, (1.14c) and (1.14d) may be omitted. Then, the system simplifies to

$$\begin{aligned} \frac{\partial \mathbf{E}}{\partial t} - c^2 \operatorname{curl} \mathbf{B} &= -\frac{1}{\varepsilon_0} J, \\ \frac{\partial \mathbf{B}}{\partial t} + \operatorname{curl} \mathbf{E} &= 0. \end{aligned}$$

(For more details concerning Maxwell's equation, see Dautray and Lions [390]). Setting

$$\mathbf{u} = (\mathbf{E}, c^2 \mathbf{B})^T,$$

introducing the matrices

$$\mathbf{D}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & c \\ 0 & -c & 0 \end{pmatrix}, \quad \mathbf{D}_2 = \begin{pmatrix} 0 & 0 & -c \\ 0 & 0 & 0 \\ c & 0 & 0 \end{pmatrix}, \quad \mathbf{D}_3 = \begin{pmatrix} 0 & c & 0 \\ -c & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$\mathbf{A}_j = \begin{pmatrix} 0 & \mathbf{D}_j \\ \mathbf{D}_j^T & 0 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} -J/\varepsilon_0 \\ 0 \end{pmatrix},$$

the system is written as

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^3 \mathbf{A}_j \frac{\partial \mathbf{u}}{\partial x_j} = \mathbf{f}.$$

The matrices  $\mathbf{A}_j$  are symmetric, and for any  $\mathbf{k} = (\xi, \eta) \in \mathbb{R}^3$ , the matrix  $\xi\mathbf{A}_1 + \eta_2\mathbf{A}_2 + \eta_3\mathbf{A}_3$  has three real double eigenvalues,  $\omega_{1,2} = -c|\mathbf{k}|$ ,  $\omega_{3,4} = 0$ , and  $\omega_{5,6} = c|\mathbf{k}|$ . We see in this example that the assumption  $\det \mathbf{A}_1 \neq 0$  is not always valid! The characteristic manifold is made up of the horizontal plane  $\omega = 0$  and the right circular cones  $C : \omega = \pm c|\mathbf{k}|$  (propagation of electromagnetic waves).  $\square$

### 1.3.3 Group Velocity and Incoming/Outgoing Modes

The link between the characteristic manifold and the above arguments introducing the U.K.C. appears in the following simple fact that we have already observed.

*Lemma 1.3*

*The point  $(\xi, \eta, \omega) \in \mathbb{R}^3$  lies on the characteristic manifold for system (1.10a) iff  $i\xi$  is a purely imaginary eigenvalue of the matrix  $\mathbf{D}(\eta, i\omega)$  defined by (1.12).*

*Proof.* Given any point  $(\xi, \eta, \omega) \in \mathbb{R}^3$ , we write as previously

$$-\omega\mathbf{I} + \xi\mathbf{A} + \eta\mathbf{B} = -\mathbf{A}(-\xi\mathbf{I} + \mathbf{A}^{-1}(\omega\mathbf{I} - \eta\mathbf{B})) = i\mathbf{A}(\mathbf{D}(\eta, i\omega) - i\xi\mathbf{I}).$$

Thus  $(\xi, \eta, \omega)$  lies on the characteristic surface iff  $i\xi$  is a purely imaginary eigenvalue of the matrix  $\mathbf{D}(\eta, i\omega) = \mathbf{A}^{-1}(i\omega\mathbf{I} - i\xi\mathbf{B})$  introduced above. Note also that

$$\mathbf{D}(\eta, i\omega)\varphi = i\xi\varphi \iff (\xi\mathbf{A} + \eta\mathbf{B})\varphi = \omega\varphi,$$

which shows that the eigenvectors of  $\mathbf{D}(\eta, i\omega)$  are those of  $\xi\mathbf{A} + \eta\mathbf{B}$ .  $\square$

The reason for introducing the characteristic manifold results from the following elementary lemma.

*Lemma 1.4*

A plane wave of the form

$$\mathbf{u}(x, y, t) = \hat{\mathbf{u}} e^{i(\xi x + \eta y - \omega t)}$$

is a solution of (1.10a) iff the point  $(\xi, \eta, \omega) \in \mathbb{R}^3$  lies on the characteristic manifold.

Thus, by this lemma, the ‘‘characteristic manifold’’ also describes the set of all wave numbers  $\mathbf{k} = (\xi, \eta)^T$  and frequencies  $\omega$  of plane wave solutions  $\hat{\mathbf{u}} e^{i(\xi x + \eta y - \omega t)}$ . The propagation velocity of such an individual plane wave is the phase velocity vector  $\frac{\omega \mathbf{k}}{|\mathbf{k}|} = (\frac{\omega}{|\mathbf{k}|})(\xi, \eta)^T$ ; thus the wave is incoming (in the domain  $x > 0$ ) if  $\text{sgn}(\omega \xi) \leq 0$ . However, the direction in which a group of these waves (such as the superposition used in the inverse Fourier transform) propagates is directly linked to the group velocity

$$\gamma = \text{grad } \omega = \left( \frac{\partial \omega}{\partial \xi}, \frac{\partial \omega}{\partial \eta} \right)^T$$

(see Chap. V, Sect. 3.1.2). To illustrate this with a simple example, take the scalar case (1.3). The direction of propagation lies along bicharacteristics (1.4) with direction  $\mathbf{c} = (a, b)^T$ . The dispersion relation is  $\omega = \mathbf{c} \cdot \mathbf{k}$ , and the group velocity coincides with that of the bicharacteristics since  $\gamma = \text{grad } \omega = \mathbf{c}$ , which points into or out of the domain  $x > 0$  according to the sign of  $a$  (see Fig. 1.2) (but the phase velocity vector of an individual plane wave can point in any desired direction). The characteristic manifold is the plane  $\omega = a\xi + b\eta = \mathbf{c} \cdot \mathbf{k}$ .

We now focus on Example 1.1, of linearized isentropic Euler equations, which is representative of more general systems. On the one hand, the plane  $\omega = \mathbf{u} \cdot \mathbf{k} = u\xi + v\eta$  (if  $\xi = \frac{\omega - v\eta}{u}$ ,  $i\xi$  is an eigenvalue of  $\mathbf{D}(\eta, i\omega)$ ) corresponds to a constant group velocity  $\mathbf{u}$  and thus to a translational motion. On the other hand, the cones correspond to group velocity  $\gamma = \mathbf{u} \pm c \frac{\mathbf{k}}{|\mathbf{k}|}$  (propagation of sound or acoustic waves). This can indeed be generalized in a rather simple way by using bicharacteristics, but we shall not go into details. Consider a cone  $C(\omega = \mathbf{u} \cdot \mathbf{k} \pm c|\mathbf{k}|)$ , and denote by  $\Xi$  the projection of this cone  $C$  onto the  $(\eta, \omega)$ -plane (see Fig. 1.4). When  $(\eta, \omega)$  lies in the interior of  $\Xi$ , there are two corresponding points  $(\xi_j, \eta, \omega)$ ,  $j = 1, 2$ , on the cone, where  $\xi_j = \xi(\eta, \omega)$  is computed from

$$\omega = u\xi + v\eta \pm c(\xi^2 + \eta^2)^{1/2};$$

in that case,  $\mathbf{D}(\eta, i\omega)$  has two other imaginary eigenvalues  $i\xi_j$ . When  $(\eta, \omega)$  approaches the boundary of  $\Xi$ , the corresponding points on the cone coalesce. Outside of  $\Xi$ , the eigenvalues are  $\ell = i(\xi \pm i\rho) = \pm\rho + i\xi$  (this is because  $-\omega\mathbf{I} + \xi\mathbf{A} + \eta\mathbf{B}$  is real). When  $\frac{\partial \omega}{\partial \eta} = 0$  (at points  $Q, Q'$  on Fig. 1.4), the group velocity is tangent to the plane  $y = 0$ , corresponding to group velocity vectors  $\text{grad } \omega$  pointing into or out of the domain (depending on whether  $\frac{\partial \omega}{\partial \xi}$

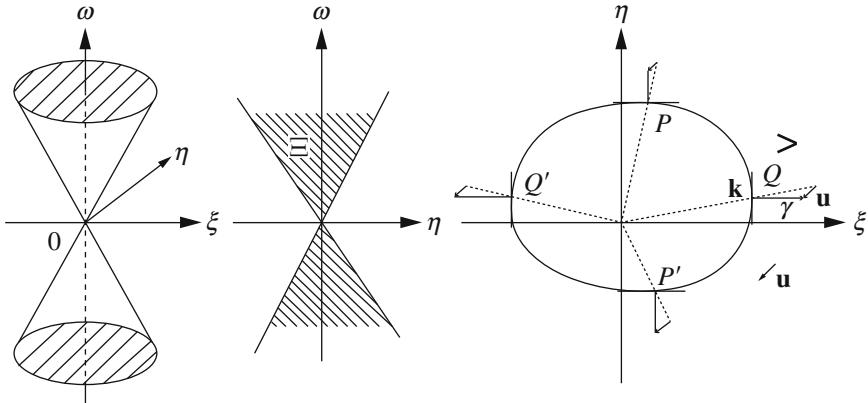


Fig. 1.4 Characteristic cone and its projections

is  $< 0$  or  $> 0$ ). On the other hand, when  $\frac{\partial \omega}{\partial \eta} = 0$  (at points  $P, P'$ ), the group velocity is tangent to the boundary  $x = 0$ . To understand the location of these points in Fig. 1.4, note that the cross section of  $C$  has locally an equation of the form  $\eta = \eta(\xi)$ , and the points in the  $(\xi, \eta)$ -plane where  $\frac{\partial \omega}{\partial \xi} = 0$  coincide with those where  $\eta'(\xi) = 0$ . Note also that in Fig. 1.4,  $\gamma = \mathbf{u} \pm c \frac{\mathbf{k}}{|\mathbf{k}|}$  is the sum of  $\mathbf{u}$  and  $c$  times a unit vector  $\frac{\mathbf{k}}{|\mathbf{k}|}$ ; the direction of  $\mathbf{u}$  implies that in the present case,  $x = 0$  is an outflow boundary.

In short, the above considerations show that for  $(\eta, \omega)$  in  $\Xi$ , each determination of  $\xi = \xi(\eta, \omega)$  such that  $(\xi, \eta, \omega)$  belongs to the cone  $C$  (i.e., such that  $i\xi$  is an eigenvalue of  $\mathbf{M}(\eta, i\omega)$ ) can be associated with a motion into ( $\frac{\partial \omega}{\partial \xi} < 0$ ) or out of ( $\frac{\partial \omega}{\partial \xi} > 0$ ) the spatial domain  $x > 0$ .

We have information on the behavior of the eigenvalues of  $\mathbf{D}(\eta, s)$  as  $\text{Res} = 0$ . Now, we can think that in normal modes  $\hat{\mathbf{u}} e^{\ell x + i\eta y - st}$ ,  $\ell$  (which is an eigenvalue of  $\mathbf{D}(\eta, s)$ ) and  $s$  (with  $\text{Re}(s) < 0$ ) are “perturbations” of  $i\xi$  and  $i\omega$ , respectively:

$i\xi$  is an eigenvalue of  $\mathbf{D}(\eta, i\omega)$ .

$\ell = \text{Re}\ell + i\xi$  is an eigenvalue of  $\mathbf{D}(\eta, s)$ , where  $s = \text{Re}(s) + i\omega$ .

We have  $\ell = \ell(\eta, s)$  and, at least formally,

$$\frac{\partial \ell}{\partial s} = \frac{\partial \xi}{\partial \omega}.$$

One can then see, under some regularity assumption on the dependence of  $\ell$  on  $s$ , that

$$\frac{\partial s}{\partial \ell} = \frac{\partial \omega}{\partial \xi}.$$

Therefore it appears that when  $s$  is “perturbed” so that  $\text{Re}(s) < 0$ , the sign of  $\text{Re}(\ell)$  will indicate whether the mode corresponds to group velocity vectors

pointing into ( $\frac{\partial \omega}{\partial \xi} > 0$ ; thus  $\operatorname{Re}(\ell) < 0$  since  $\operatorname{Re}(s) < 0$ ) or out of ( $\operatorname{Re}(\ell) > 0, \frac{\partial \omega}{\partial \xi} < 0$ ) the domain.

Informally speaking, identifying “incoming” and “outgoing” portions of the solution  $\tilde{\mathbf{u}}$ , obtained by means of Laplace and Fourier transform as a “superposition” of elementary modes  $\hat{\mathbf{u}} e^{\ell x} e^{iny - st}$ , is linked to the sign of  $\operatorname{Re}(\ell)$ ,  $\ell$  an eigenvalue of  $\mathbf{D}(\eta, s)$ .

Now, there exists a matrix  $\mathbf{Q} = \mathbf{Q}(\eta, s)$  such that  $\mathbf{Q}^{-1} \mathbf{D} \mathbf{Q}(\eta, s)$  is made up of two Jordan blocks  $\mathbf{D}_1(\eta, s)$  and  $\mathbf{D}_2(\eta, s)$ ,

$$\mathbf{Q} \mathbf{D} \mathbf{Q}^{-1}(\eta, s) = \begin{pmatrix} \mathbf{D}_1 & \\ & \mathbf{D}_2 \end{pmatrix}, \quad (1.15a)$$

and such that the  $q$  eigenvalues  $\ell$  of  $\mathbf{D}_1$  (resp.  $(p - q)$  eigenvalues of  $\mathbf{D}_2$ ) have  $\operatorname{Re}(\ell) < 0$  (resp.  $> 0$ ). Let us set in (1.11)

$$\tilde{\mathbf{v}}(x, \eta, s) = \mathbf{Q} \tilde{\mathbf{u}} = (\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2)^T. \quad (1.15b)$$

Then, the differential equation (1.11) splits into

$$\begin{aligned} \frac{\partial \tilde{\mathbf{v}}_1}{\partial x} &= \mathbf{D}_1(\eta, s) \tilde{\mathbf{v}}_1, \\ \frac{\partial \tilde{\mathbf{v}}_2}{\partial x} &= \mathbf{D}_2(\eta, s) \tilde{\mathbf{v}}_2. \end{aligned}$$

The solutions  $\tilde{\mathbf{v}}_j$ , of the form  $\tilde{\mathbf{v}}_j e^{\ell x}$ , are associated with “incoming” (resp. “outgoing”) waves since  $\ell$  is an eigenvalue of  $\mathbf{D}_1$  (resp.  $\mathbf{D}_2$ ). Using the fact that  $\mathbf{Q} = \mathbf{Q}(\eta, s)$  and that the Laplace–Fourier transform only acts on the function of the  $(y, t)$  variables, it happens that the boundary condition (1.10),  $\mathbf{E}\mathbf{u}(0, y, t) = \mathbf{g}(y, t)$ , becomes

$$\mathbf{E} \mathbf{Q}^{-1} \tilde{\mathbf{v}}(0, \eta, s) = \tilde{\mathbf{g}}(\eta, s).$$

We can also decompose the columns of  $\mathbf{Q}^{-1}$  in two parts,  $\mathbf{Q}^{-1} = (\mathbf{Q}_1, \mathbf{Q}_2)$ , so that

$$\mathbf{Q}^{-1} \tilde{\mathbf{v}} = \mathbf{Q}_1 \tilde{\mathbf{v}}_1 + \mathbf{Q}_2 \tilde{\mathbf{v}}_2,$$

and the boundary conditions are transformed into

$$\mathbf{N}(\eta, s) \tilde{\mathbf{v}}_1 = -\mathbf{E} \mathbf{Q}_2 \tilde{\mathbf{v}}_2 + \tilde{\mathbf{g}}$$

by definition (1.13) of  $\mathbf{N}(\eta, s)$  (up to a linear change of dependent variables  $\tilde{\mathbf{v}}_i$ , which does not alter the class of “incoming” or “outgoing”). This hints that the uniform Kreiss condition can be interpreted as a uniform solvability condition, which enables one to solve for incoming dependent variables in terms of outgoing variables and boundary data. Again, this is just an outline. In particular, tangential group velocities corresponding to waves moving tangent to the boundary, which we have completely skipped, are not so eas-

ily handled. For precise results and the nonconstant coefficient extension, we refer to Higdon [615] and the references therein; see also Dutt [453].

## 2 The Nonlinear Approach

### 2.1 Nonlinear Equations

In this section, we introduce the reader to some results for a nonlinear system or equation, obtained without linearization. We shall not give any proof, and we refer to the cited papers for details.

Consider first the simple Burgers' equation in the strip  $0 < x < 1$ , with boundary conditions given on  $x = 0$  and  $x = 1$ . The problem is obviously more complex than in the linear case.

We take

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial u^2}{\partial x} = 0, \\ u(x, 0) = u_0(x), \\ u(0, t) = g(t), \\ u(1, t) = h(t). \end{cases}$$

According to the data, there may exist or not a weak solution satisfying the boundary data, and these data may be effectively necessary for the computation of the solution or not. Indeed, we leave as an exercise for the reader the computation of a solution corresponding to  $u_0(x) = 0$ ,  $g(t) = 1$ ,  $h(t) = -1$  or  $g(t) = -1$ ,  $h(t) = 1$  and to  $u_0(x) = 1$ ,  $g(t) = -1$ ,  $h(t) = 1$ .

Kruzhkov's existence and uniqueness result has been extended by Bardos, LeRoux, and Nedelec [86] to the scalar multidimensional I.B.V.P. in a domain  $\mathcal{O}$  of  $\mathbb{R}^d$ ,

$$\begin{cases} \frac{\partial u}{\partial t} + \sum_{i=1}^d \frac{\partial}{\partial x_i} f_i(u) = 0, & x \in \mathcal{O}, t \in (0, T), \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), \text{ a.e. in } \mathcal{O}, \\ u(\mathbf{x}, t) = 0, \text{ on } \partial\mathcal{O} \times (0, T), \end{cases} \quad (2.1)$$

via the vanishing viscosity method (see the Chap. I, Sect. 5.2, and G.R., Chapter 2 [539]), by studying more closely the boundary integrals in the entropy inequality. We just mention their result, corresponding for simplicity to boundary conditions equal to zero. Note that the result supposes that one can define the trace of the entropy fluxes on the boundary  $t = 0$  and on  $\partial\mathcal{O} \times (0, T)$ , which holds, for instance, for a function in  $BV(\mathcal{O} \times (0, T))$  (the

proof can be found in [520]); here, we assume that  $\mathcal{O}$  is bounded and that  $u_0$  belongs to  $C^2(\bar{\mathcal{O}})$ .

*Theorem 2.1*

*There exists a unique entropy solution  $u$  of (2.1), in the sense that  $u$  is the unique function in  $BV(\mathcal{O} \times (0, T))$  satisfying for any test function  $\varphi \in C_0^2(\bar{\mathcal{O}} \times [0, T])$ ,  $\varphi \geq 0$*

$$\begin{aligned} & \int_0^T \int_{\mathcal{O}} \left\{ |u - k| \frac{\partial \varphi}{\partial t} + \operatorname{sgn}(u - k) \sum_{i=1}^d (f_i(u) - f_i(k)) \frac{\partial \varphi}{\partial x_i} \right\} d\mathbf{x} dt \\ & + \int_0^T \int_{\partial\mathcal{O}} \operatorname{sgn}(k) \left( \sum_{i=1}^d (f_i(0) - f_i(k)) \nu_i \right) \varphi(s, t) ds dt \geq 0, \end{aligned}$$

where  $\nu$  is the unit outward normal to  $\partial\mathcal{O}$  and

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}) \text{ a.e. in } \mathcal{O}.$$

The solution is obtained as the limit of solutions of the following viscous equations (with  $(\varepsilon > 0)$ ):

$$\begin{cases} \frac{\partial u_\varepsilon}{\partial t} + \sum_{i=1}^d \frac{\partial}{\partial x_i} f_i(u_\varepsilon) - \varepsilon \Delta u_\varepsilon = 0, & \mathbf{x} \in \mathcal{O}, t \in (0, T), \\ u_\varepsilon(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \mathcal{O}, \\ u_\varepsilon(\mathbf{x}, t) = 0, & \text{on } \partial\mathcal{O} \times (0, T). \end{cases}$$

The general case of nonzero boundary data  $g$  belonging to  $C^2(\partial\mathcal{O} \times (0, T))$  follows by translating the solution by  $\tilde{g}$ , where  $\tilde{g}$  is an extension of  $g$  to  $\mathcal{O} \times (0, T)$ . The definition of a weak entropy solution in the theorem can be interpreted in the following terms. If the boundary data  $g$  are given, the value  $u(x, t)$ , the trace of  $u$  on the boundary at time  $t$ , must satisfy

$$(\operatorname{sgn}(u(x, t) - g(x, t)) \sum_{i=1}^d (f_i(u) - f_i(k)) \nu_i) \geq 0 \text{ a.e. on } \partial\mathcal{O} \times (0, T),$$

$\forall k \in I(u(x, t), g(x, t))$ , where  $I(a, b)$  denotes the interval  $[\min(a, b), \max(a, b)]$ . For a one-dimensional equation, this condition can be characterized in a rather simple way:

$$\frac{f(u) - f(k)}{u - k} \leq 0, \quad \forall k \text{ between } u = u(0, t) \text{ and } g = g(t), \quad (2.2)$$

which says that the slope of the chord joining  $(u, f(u))$  to  $(k, f(k))$  is negative.

*Example 2.1.* Take again the above example of Burgers' equation with  $u_0(x) = 1$ ,  $g(t) = -1$ ,  $h(t) = 1$ . On the one hand, if we take the boundary data  $g(t) = -1$  into account, obviously a difficulty appears at the boundary  $x = 0$ . On the other hand, if we do not consider the boundary data at  $x = 0$ , there is an infinity of weak continuous solutions. Indeed, consider, for instance, the functions

$$u(x, t) = \begin{cases} \frac{(x + \alpha)}{t}, & x \leq t - \alpha, \\ 1, & x \geq t - \alpha, \end{cases}$$

depending on a parameter  $\alpha \in [0, 1]$ . For  $t \leq \alpha$ ,  $u(x, t) = 1$  for all  $x \geq 0$ ; for  $t \geq \alpha$ , the solution consists of a rarefaction and a constant state (see Fig. 2.1). Let us study these solutions regarding the requirements (2.2) given by Theorem 2.1:

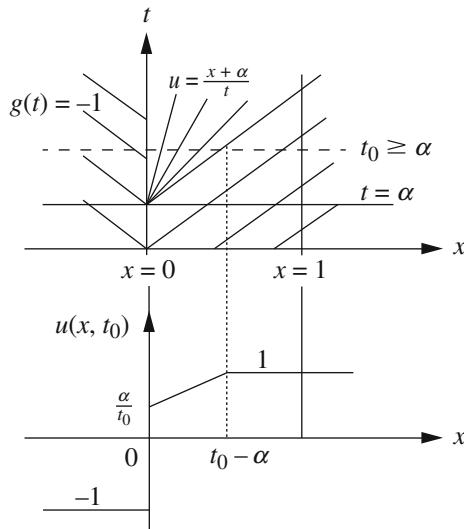
If  $\alpha > 0$ , for  $t \leq \alpha$ ,  $I(u(0, t), g(t)) = [-1, 1]$ ; for  $t \geq \alpha$ ,  $I(u(0, t), g(t)) = [-1, \frac{\alpha}{t}]$ , and of course the boundary entropy condition (2.2) is not satisfied: for instance, if  $t \leq \alpha$ ,  $k \in [-1, 1]$  (see Fig. 2.1),

$$\frac{f(u(0, t) - f(k)}{1 - k} > 0.$$

If  $\alpha = 0$ , the solution is such that the discontinuity at  $x = 0$  indeed satisfies (2.2);  $\forall k \in I(u(0, t), g(t)) = [-1, 0]$ ,

$$\frac{f(u(0, t) - f(k)}{u - k} = \frac{f(0) - f(k)}{-k} \leq 0.$$

It is thus the unique solution given by Theorem 2.1.



**Fig. 2.1** Example of I.B.V.P. for Burgers' equation

If we consider the discontinuity connecting the state  $u_L = -1 = g(t)$  to  $u_R = u(0, t)$ , the Rankine–Hugoniot condition gives

$$\sigma = \frac{f(u(0, t)) - f(g(t))}{u(0, t) - g(t)} = \frac{u(0, t) - 1}{2}.$$

Remembering Oleinik's entropy condition (see G.R., Chapter 2, Lemma 6.1), that condition is obviously violated for any  $\alpha$ , because  $u(0, t) > g(t)$ . However, for the solution given by the theorem,  $\sigma < 0$  and the discontinuity leaves the domain immediately.  $\square$

## 2.2 Nonlinear Systems

The vanishing viscosity method has been used for one-dimensional systems

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = \mathbf{0}, \quad x > 0, t > 0, \quad (2.3a)$$

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x), \quad x > 0, \quad (2.3b)$$

in particular by Benabdallah and Serre [106], Dubois and LeFloch [437], and Gisclon and Serre [518]. The first authors define a set  $\mathcal{E}(\mathbf{g})$  of admissible boundary values for which a boundary entropy inequality holds. The boundary condition is therefore written in the form  $\mathbf{u}(0+, t) \in \mathcal{E}(\mathbf{g}(t))$ , which stands in place of the strong nonhomogeneous Dirichlet condition  $\mathbf{u} = \mathbf{g}$ . Dubois and LeFloch also propose a second way of selecting admissible boundary conditions involving the resolution of Riemann problems. This leads to the introduction of another set of admissible boundary conditions  $\mathbf{u}(0+, t) \in \mathcal{V}(\mathbf{g})$ . The two kinds of boundary conditions are linked in general by

$$\mathbf{u} \in \mathcal{V}(\mathbf{g}) \implies \mathbf{u} \in \mathcal{E}(\mathbf{g}),$$

and they coincide in some cases (scalar, linear systems, systems of the Temple class), but not always (see Benabdallah and Serre [106] for a counterexample). Gisclon and Serre [518] obtain a still different way of writing the boundary condition  $\mathbf{u}(0, t) \in C(\mathbf{g}(t))$ , with  $C(\mathbf{g}) \subset \mathcal{E}(\mathbf{g})$ . It is interesting to note that their “residual” boundary condition involves in general the viscosity matrix  $\mathbf{B}$  of the perturbed system

$$\frac{\partial \mathbf{u}_\varepsilon}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}_\varepsilon) = \varepsilon \frac{\partial}{\partial x} \left( \mathbf{B}(\mathbf{u}_\varepsilon) \frac{\partial}{\partial x} \mathbf{u}_\varepsilon \right).$$

Before entering into the description of these sets, we stress the point that, since very few results concerning the pure Cauchy problem are available (see the notes at the end of Chap. II), one cannot expect any general existence

theorem concerning the I.B.V.P. (see, however, Li Ta-Tsien [794] for results in the case of a “reducible”  $2 \times 2$  system and Li Ta-Tsien and Yu Wen-Ci [795] for piecewise smooth solutions). The membership in  $\mathcal{E}(\mathbf{g})$  or  $\mathcal{V}(\mathbf{g})$  appears as a necessary condition to prove the existence of the corresponding I.B.V.P. Only in some particular examples (either the scalar case, or linear systems, constant data, or  $p$ -system; see Benabdallah [105] and Dubroca and Gallice [441] and [442]) is this condition sufficient to prove existence and uniqueness. Other references can be found in [437]. Gisclon and Serre [518] prove that their condition leads to a well-posed I.B.V.P., with convergence of the viscous solutions to the hyperbolic one, and also [106].

*Definition 2.1*

*For each state  $\mathbf{g}$  in  $\Omega$ , the set  $\mathcal{E}(\mathbf{g})$  of admissible values at the boundary is defined as the set  $\mathcal{E}(\mathbf{g})$  of states  $\mathbf{u}$  in  $\Omega$  such that*

$$F(\mathbf{u}) - F(\mathbf{g}) - U' \mathbf{g} \cdot \{\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{g})\} \leq 0 \quad (2.4)$$

*for each entropy pair  $(U, F)$ .*

The “boundary condition” for the problem is then

$$\mathbf{u}(0+, t) \in \mathcal{E}(\mathbf{g}(t)), t > 0.$$

Let us now introduce another set, which is easier to deal with and directly applies to numerical approximation by finite volume methods since it involves Riemann problems. Given a state  $\mathbf{g} \in \Omega$ , we consider the following set.

*Definition 2.2*

*The set  $\mathcal{V}(\mathbf{g})$  of admissible values at the boundary is the set  $\mathcal{V}(\mathbf{g})$  of states  $\mathbf{w}$  in  $\Omega$  such that*

$$\mathcal{V}(\mathbf{g}) = \{\mathbf{w} / \mathbf{w} = \mathbf{w}_R(0+; \mathbf{g}, \mathbf{u}_R), \mathbf{u}_R \in \Omega\}, \quad (2.5)$$

*where  $\mathbf{w}_R$  denotes the solution of the Riemann problem.*

We assume for simplicity that  $\Omega$  is such that the Riemann problem has a solution (otherwise, we must add some restriction on  $\mathbf{u}_R$ ). The corresponding boundary condition for the problem is then

$$\mathbf{u}(0+, t) \in \mathcal{V}(\mathbf{g}(t)), \quad t > 0.$$

Consider first the *scalar* convex case. If we assume that  $\lim_{u \rightarrow \pm\infty} f(u) = +\infty$  as  $u \rightarrow \pm\infty$ , we can easily characterize  $\mathcal{E}(g)$ .

*Proposition 2.1*

*Denote by  $\bar{u}$  the state where  $f$  is minimum. For a given  $g \neq \bar{u}$ , the set of “admissible” values  $\mathcal{E}(g)$  is*

$$\begin{aligned}\mathcal{E}(g) &= (-\infty, \bar{u}] \quad \text{if } g \leq \bar{u}, \\ \mathcal{E}(g) &= (-\infty, \bar{g}] \cup \{g\} \quad \text{if } g \geq \bar{u},\end{aligned}$$

where  $\bar{g}$  denotes the solution  $\bar{g} \neq g$  of  $f(\bar{g}) = f(g)$ .

We also mention the fact that in the scalar (nonconvex) case, the two sets coincide.

*Proposition 2.2*

In the case of a scalar conservation law, for any  $g$  in  $\Omega$ , the sets  $\mathcal{V}(g)$  and  $\mathcal{E}(g)$  coincide.

*Example 2.1* (Revisited). Let us illustrate the above results with Burgers' equation for which  $\bar{u} = 0$ ,  $\bar{g} = -g$ . On the one hand,  $g(t) = -1 \leq 0$ , and the set of states  $u$  such that the chord has a negative slope is indeed  $\mathcal{E}(g) = (-\infty, 0)$ . On the other hand, if we look at all the limits  $w_R(0+; -1, u_R), u_R \in \Omega$ , of the Riemann problems, we find:

- (a) If  $u_R \geq -1$ , the solution of the Riemann problem is a rarefaction, and when  $u_R$  varies,  $w_R(0+; -1, u_R)$  lies in  $[-1, 0]$ .
- (b) If  $u_R < -1$ , the solution of the Riemann problem is a shock, and when  $u_R$  varies,  $w_R(0+; -1, u_R)$  lies in  $(-\infty, -1]$  so that  $\mathcal{V}(g) = (-\infty, -1] \cup [-1, 0] = (-\infty, 0]$  as expected.

Consider, for instance, the constant data  $g \geq 0$ . The proposition gives  $\mathcal{V}(g) = (-\infty, -g] \cup \{g\}$ . Let us interpret this result:

- (i) If  $u(0, t) < -g$ , the trace of the solution  $u$  on the boundary  $x = 0$  is “far” from the data; however,  $u \in \mathcal{V}(g)$ , and this outgoing wave ( $u(0, t) \leq 0$ ) is admissible: this is a truly nonlinear effect and could not have been obtained by linearization. The result is such that the solution of the Riemann problem (the discontinuity connecting the state  $u_L = g$  to  $u_R = u$  leaves the domain since it propagates with speed  $\sigma = \frac{1}{2}(u + g) < 0$ ).
- (ii) If  $u$  is “near” the data  $g$ , then the proposition yields that  $u = g$ : that could have been obtained by linearization.

The characterization is also quite simple in the case of a strictly hyperbolic *linear system*; the sets  $\mathcal{V}(\mathbf{g})$  and  $\mathcal{E}(\mathbf{g})$  coincide, and we recover the formulation of Sect. 1.3.1. Denoting as previously by  $p' = p - q$  the number of nonpositive eigenvalues of  $\mathbf{A}$  ( $a_i \leq 0, 1 \leq i \leq p'$ ), we can state the following proposition.

*Proposition 2.3*

Assume that the system (2.3) is linear. For any  $\mathbf{g}$  in  $\mathbb{R}^n$ ,

$$\mathcal{V}(\mathbf{g}) = \mathcal{E}(\mathbf{g}) = \left\{ \mathbf{u}, \exists (\alpha_i) \in \mathbb{R}^{p'}, \mathbf{u} = \mathbf{g} + \sum_{i=1}^{p'} \alpha_i \mathbf{r}_i \right\}. \quad (2.6)$$

Using Chap. II, Sect. 1, it is easy to check that  $\mathcal{V}(\mathbf{g})$  is indeed the set  $\{\mathbf{u} = \mathbf{g} + \sum_i \alpha_i \mathbf{r}_i\}$ . For the other identity, we need the following lemma.

*Lemma 2.1*

*For a strictly hyperbolic linear system, the pairs of entropy–entropy flux functions are given by*

$$U(\mathbf{u}) = \sum_{i=1}^p \varphi_i(w_i), \quad F(\mathbf{u}) = \sum_{i=1}^p a_i \varphi_i(w_i)$$

for any arbitrary convex functions  $\varphi_i$ , where  $\mathbf{w} = (w_i)$  are the characteristic variables (1.7).

This characterization of entropy–entropy flux pairs for a linear system is an easy consequence of Theorem 5.1 in the Chap. I, Sect. 5 (which states that  $U''\mathbf{A}$  is symmetric). For a detailed proof, we refer to [437], Lemma 1.2.

Let us see that (2.6) gives another equivalent way of formulating (1.8) (i.e.,  $\mathbf{w}^I(0, t) = \mathbf{g}^I(t)$ , where we recall that the subscript  $I$  (resp.  $II$ ) corresponds to positive (resp. nonpositive) eigenvalues  $a_i$ ,  $\mathbf{w}^I = (w_{p'+1}, \dots, w_p)^T$ ,  $\mathbf{w}^{II} = (w_1, \dots, w_{p'})^T$ ).

Indeed (2.6) says that  $\mathbf{u}$  and  $\mathbf{g}$  have the same components relative to the  $q = p - p'$  positive eigenvalues (incoming characteristics).

If  $\mathbf{g} \in \mathbb{R}^p$  is given and  $\mathbf{u} \in \mathcal{V}(\mathbf{g})$ , let  $\mathbf{g}^I$  be the projection of  $\mathbf{g}$  onto the space spanned by the  $\mathbf{r}_i$ ,  $i = p' + 1, \dots, p$ , i.e.,

$$\mathbf{g} = \sum_{i=1}^p g_i \mathbf{r}_i, \quad \mathbf{g}^I = \sum_{i=p'+1}^p g_i \mathbf{r}_i.$$

Then (1.8) holds. Conversely, if  $\mathbf{w}^I(0, t) = \mathbf{g}^I(t)$ , then obviously  $\mathbf{u} \in \mathcal{V}(\mathbf{g})$ .

For a general *nonlinear system*, a first result regarding the well-posedness of the I.B.V.P. concerns the case where the initial data  $\mathbf{u}_0(x) = \mathbf{u}_0$  and the boundary data  $\mathbf{g}(t) = \mathbf{g}_0$  are constant; the I.B.V.P. is then defined by

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = 0, \quad x > 0, \quad t > 0, \tag{2.7a}$$

$$\mathbf{u}(x, 0) = \mathbf{u}_0, \quad x > 0, \tag{2.7b}$$

$$\mathbf{u}(0+, t) \in \mathcal{V}(\mathbf{g}_0), \quad t > 0. \tag{2.7c}$$

If the I.B.V.P. (2.7) is well-posed, this means that  $\mathbf{u}(x, t)$  has an extension as an entropy solution of a Riemann problem on  $\mathbb{R} \times (0, \infty)$  in the “usual class” (discussed in Chap. II, Sect. 6). This extension is precisely  $\mathbf{w}_R(\frac{x}{t}; \mathbf{g}_0, \mathbf{u}_0)$ , and  $\mathbf{u}(0, t)$  is the trace of this solution of the Riemann problem

$$\mathbf{u}(0, t) = \mathbf{w}_R(0+; \mathbf{g}_0, \mathbf{u}_0).$$

*Theorem 2.2*

Assuming that  $\mathbf{u}_0$  and  $\mathbf{g}_0$  are constant, the problem (2.7) admits a unique entropy solution in the usual class of constant states separated by elementary waves.

*Example 2.2.*  $\mathcal{V}(\mathbf{g}_0)$  can be characterized in the particular case of the isentropic one-dimensional Euler system

$$\begin{aligned}\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) &= 0 \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) &= 0,\end{aligned}$$

(see Example 1.1 above or Chap. V, Sect. 1, Example 1.1 with  $v = 0$ ) where  $p = k\rho^\gamma$ , for which the rarefaction and shock curves can be explicitly computed, by studying closely the Riemann problem according to the position of  $\mathbf{g}_0$  (sub- or supersonic, in- or outlet).

For instance, for supersonic inflow  $\mathbf{g}_0 = (\bar{\rho}_0, \bar{q}_0 = \bar{\rho}_0 \bar{u}_0)$ , with  $0 < \lambda_-(\mathbf{g}_0) = \bar{u}_0 - \bar{c}_0 < \lambda_+(\mathbf{g}_0) = \bar{u}_0 + \bar{c}_0$ , one finds that  $\mathbf{g}_0$  is the only supersonic inflow state ( $u - c > 0$ ) in  $\mathcal{V}(\mathbf{g}_0)$ . Then, the boundary condition is  $\mathbf{u}(0, t) = \mathbf{g}_0$ , corresponding to the case where the linearization is valid. However, as we have already observed in the scalar case, a strong nonlinear behavior can be observed:  $\mathcal{V}(\mathbf{g}_0)$  is reduced to  $\{\mathbf{g}_0\}$  only near  $\mathbf{g}_0$ . Since the intersection of  $\mathcal{V}(\mathbf{g}_0)$  with the subsonic states  $\{\mathbf{u}, -c \leq u \leq c\}$  is a part of the 1-shock curve  $\mathcal{S}_1(\mathbf{g}_0)$ ,  $\mathbf{u}(0, t)$  can be such that  $\mathbf{g}_0$  is connected to  $\mathbf{u}(0, t)$  by a 1-shock. Otherwise, since the intersection of  $\mathcal{V}(\mathbf{g}_0)$  with the subsonic states  $\{\mathbf{u}, \lambda_-(\mathbf{u}) < \lambda_+(\mathbf{u}) < 0\}$  is locally of dimension 2,  $\mathbf{u}(0, t)$  may even correspond to a supersonic outflow ( $\mathbf{u}(0, t) = (\rho, \rho u)$  with  $u + c < 0$ ). Again, we refer to [437] for details.  $\square$

We also mention the recent result of Gisclon and Serre [518] concerning a strictly hyperbolic system with nonzero eigenvalues. Since in the general case the definition of  $C(g(t))$  needs further development, we refer to Gisclon [516] for details.

### 3 Gas Dynamics

Concerning physical problems and in particular gas dynamics, one has usually to distinguish between two types of boundary conditions:

- *Actual boundary conditions:* the boundary is that of the spatial (bounded) domain, which can be a fluid boundary, a solid boundary, or a free surface (a case we shall not consider).
- *Artificial boundary conditions:* when the spatial domain is unbounded (for instance, fluid flow in an exterior domain such as flow past an air-

foil or in an interior infinite domain such as a channel), one limits the area of computation and introduces artificial boundaries. Then arises the problem of specifying boundary data on this artificial boundary.

If we are given an external state  $\mathbf{U}_\infty$  (uniform flow condition) at infinity, then, in view of the above analysis, one can think of two approaches. The usual one is linearization, which we shall detail later; one linearizes at this external state and “forces” some of the variables, say  $w$ , which is assigned the value  $w_\infty$  (for instance, supersonic inflow,  $\mathbf{U} = \mathbf{U}_\infty$ ; supersonic outflow, no condition). This applies particularly to a stationary state computation, and the initial condition is then taken as  $\mathbf{U}(\mathbf{x}, 0) = \mathbf{U}_\infty$ .

One also can take into account nonlinear effects and compute the state at the boundary by solving Riemann problems.

One may derive absorbing (or radiation or nonreflecting) boundary conditions at the artificial boundary. The idea is that outgoing waves should be absorbed and not artificially reflected back into the flow from the computational boundary. We shall discuss some of these notions in Sect. 4.

Let us now detail the usual inflow and outflow boundary conditions obtained by linearization (see Oliger and Sundstrom [915]).

### 3.1 Fluid Boundary (Linearized Approach)

The Euler system in one dimension linearized about a smooth state  $\mathbf{U}_0$  is obtained by substituting  $\mathbf{U} = \mathbf{U}_0 + \mathbf{U}'$  into the system (see Example 2.4, Chap. II) and neglecting terms of second order in  $\mathbf{U}'$ . We get, dropping the “prime,”

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}(\mathbf{U}_0) \frac{\partial \mathbf{U}}{\partial x} = \mathbf{R}(\mathbf{U}, \mathbf{U}_0),$$

where the Jacobian matrix  $\mathbf{A}$  is given in (4.20), Chap. IV, and the source term  $\mathbf{R}(\mathbf{U}, \mathbf{U}_0)$ , which does not contain derivatives of  $\mathbf{U}$ , does not modify the boundary conditions, which are determined by the matrix  $\mathbf{A}(\mathbf{U}_0)$ .

The above considerations in Sect. 1.2 concern, a priori, the linearized characteristic variables (1.7), defined by  $\mathbf{W} = \mathbf{T}^{-1}(\mathbf{U}_0)\mathbf{U}$ , which satisfy the linear diagonal system

$$\frac{\partial \mathbf{W}}{\partial t} + \mathbf{A}(\mathbf{U}_0) \frac{\partial \mathbf{W}}{\partial x} = \mathbf{S}(\mathbf{W}, \mathbf{W}_0),$$

where  $\mathbf{A}(\mathbf{U}_0) = \text{diag}(u_0 - c_0, u_0, u_0 + c_0)$  and  $c^2 = \frac{\partial p}{\partial \rho}(\rho, s)$ .

One decomposes  $\mathbf{W} = (\mathbf{w}^I, \mathbf{w}^{II})^T \in \mathbb{R}^q \times \mathbb{R}^{3-q}$ , in incoming/outgoing variables, and then one has to specify  $\mathbf{w}^I$  at  $x = 0$ .

In practice, it is more convenient to prescribe “physical,” i.e., measurable, variables such as pressure, velocity, or temperature. Thus, one can be interested in specifying as many conservative  $\mathbf{u}^I$  or nonconservative variables  $\mathbf{v}^I$ . One has then to check that the corresponding boundary conditions on  $\mathbf{W}$

lead to a well-posed problem. For instance, in the case  $\mathbf{V}' = (\rho, u, p)^T$ ,  $\mathbf{V}'$  is a solution of (with the notations of Chap. II, Remark 2.2)

$$\frac{\partial \mathbf{V}'}{\partial t} + \mathbf{B}'(\mathbf{V}'_0) \frac{\partial \mathbf{V}'}{\partial x} = \mathbf{R}'(\mathbf{V}', \mathbf{V}'_0),$$

where  $\mathbf{T}'^{-1} \mathbf{B}' \mathbf{T}' = \Lambda$ ; setting

$$\mathbf{W}(\mathbf{V}') = \mathbf{T}'^{-1}(\mathbf{V}'_0) \mathbf{V}',$$

the relation between  $\mathbf{V}'$  and  $\mathbf{W}(\mathbf{V}')$  is

$$\begin{aligned} \mathbf{W}(\mathbf{V}') &= \left( -\frac{\rho_0 u}{2c_0} + \frac{p}{2c_0^2}, \rho - \frac{p}{c_0^2}, \frac{\rho_0 u}{2c_0} + \frac{p}{2c_0^2} \right)^T \\ &= \frac{1}{2c_0^2} (-\rho_0 c_0 u + p, 2(\rho c_0^2 - p), \rho_0 c_0 u + p)^T. \end{aligned} \quad (3.1)$$

*Remark 3.1.* If we take instead  $\mathbf{V} = (\rho, u, s)^T$ ,  $\mathbf{V}$  is a solution of (with the notations of Chap. II, Remark 2.1)

$$\frac{\partial \mathbf{V}}{\partial t} + \mathbf{B}(\mathbf{V}_0) \frac{\partial \mathbf{V}}{\partial x} = \mathbf{r}(\mathbf{V}, \mathbf{V}_0),$$

with  $\mathbf{T}^{-1} \mathbf{B} \mathbf{T} = \Lambda$ ; setting

$$\mathbf{W}(\mathbf{V}) = \mathbf{T}^{-1}(\mathbf{V}_0) \mathbf{V},$$

the relation between  $\mathbf{V}$  and  $\mathbf{W}(\mathbf{V})$  is

$$\mathbf{W}(\mathbf{V}) = \left( \frac{\rho}{2\rho_0} - \frac{u}{2c_0} + \frac{p_s^0 s}{2\rho_0 c_0^2}, -\frac{s}{c_0^2}, \frac{\rho}{2\rho_0} + \frac{u}{2c_0} + \frac{p_s^0 s}{2\rho_0 c_0^2} \right)^T.$$

If we introduce a linearized pressure  $p$ ,

$$p = p_s^0 s + p_\rho^0 \rho = p_s^0 s + c_0^2 \rho,$$

we can write

$$\mathbf{W}(\mathbf{V}) = \frac{1}{2\rho_0 c_0^2} (p - \rho_0 c_0 u, -2\rho_0 s, p + \rho_0 c_0 u)^T.$$

We notice that the first and last components of  $\mathbf{W}(\mathbf{V}')$  and  $\mathbf{W}(\mathbf{V})$  coincide (up to a constant, which corresponds to the fact that we have not normalized the eigenvectors) but not the second one, though the equation  $\frac{\partial w_2}{\partial t} + u_0 \frac{\partial w_2}{\partial x} = 0$  is of course satisfied by both  $w_2(\mathbf{V}')$  and  $w_2(\mathbf{V})$ .  $\square$

Then, if we have a partition  $\mathbf{V} = (\mathbf{v}^I, \mathbf{v}^{II}) \in \mathbb{R}^N \times \mathbb{R}^{p-N}$ , the relation  $\mathbf{W}(\mathbf{V}) = \mathbf{T}^{-1}(\mathbf{V}_0) \mathbf{V}$  can be written in matrix form,

$$\begin{pmatrix} \mathbf{w}^I \\ \mathbf{w}^{II} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_I & \mathbf{a}_{II} \\ \mathbf{b}_I & \mathbf{b}_{II} \end{pmatrix} \begin{pmatrix} \mathbf{v}^I \\ \mathbf{v}^{II} \end{pmatrix},$$

with matrices that are rectangular ( $\mathbf{a}_{II}, \mathbf{b}_I$ ) or square ( $\mathbf{a}_I, \mathbf{b}_{II}$ ) and possibly empty, and this last relation yields

$$\begin{aligned} \mathbf{w}^I &= \mathbf{a}_I \mathbf{v}^I + \mathbf{a}_{II} \mathbf{v}^{II}, \\ \mathbf{w}^{II} &= \mathbf{b}_I \mathbf{v}^I + \mathbf{b}_{II} \mathbf{v}^{II}. \end{aligned}$$

Assume that the boundary data are “ $\mathbf{v}^I$  prescribed.” The condition to get admissible boundary conditions is first that the partition of  $\mathbf{W}$  corresponds to incoming/outgoing variables. Hence, we must have  $N = q$ . Next, we want for  $\mathbf{w}^I$  an expression of the form

$$\mathbf{w}^I(0, t) = \mathbf{S}^I \mathbf{w}^{II}(0, t) + \mathbf{g}^I(t);$$

this condition supposes that either  $\mathbf{b}_{II}$  is invertible, in which case

$$\mathbf{w}^I = \mathbf{a}_{II} \mathbf{b}_{II}^{-1} \mathbf{w}^{II} + (\mathbf{a}_I - \mathbf{a}_{II} \mathbf{b}_{II}^{-1} \mathbf{b}_I) \mathbf{v}^I,$$

or  $\mathbf{b}_{II}$  is empty and  $\mathbf{w}^I = \mathbf{a}_I \mathbf{v}^I$ . This should be checked in all the particular examples.

For a one-dimensional problem, assuming that the boundary is on the left of the domain (for instance, the domain is  $x > 0$ ), we have four possible cases. We exclude the case  $u_0 = 0$ , which corresponds to a rigid wall boundary and will be considered later.

- (1) Supersonic inflow,  $u_0 > c_0, q = 3$ ; all three eigenvalues are positive: three boundary conditions, which means that the whole state must be prescribed.
- (2) Subsonic inflow,  $q = 2$  positive eigenvalues,  $c_0 > u_0 > 0 > -c_0$ : two conditions. A precise study (Oliger and Sundström 1978) gives that, in conservative variables, one can impose any pair among  $(\rho, \rho u, \rho e)$  or even any two linear combinations of  $(\rho, \rho u, \rho e)$ . In primitive variables, one can prescribe  $(\rho, u)$  or  $(\rho, p)$  but not the pair  $(u, p)$ . Indeed, setting  $\mathbf{v}^I = (u, p)$ ,  $\mathbf{v}^{II} = \rho$ , (3.1) gives  $\mathbf{b}^{II} = 0$  and yields an ill-posed problem.
- (3) Subsonic outflow,  $q = 1$  positive eigenvalue,  $0 > u_0 > -c_0$ : one condition, which can be the density, the pressure, or the velocity (in fact any combination of  $(\rho, \rho u, \rho e)$  or  $(\rho, u, p)$ ).
- (4) Supersonic outflow  $u_0 < -c_0, q = 0$ : no conditions.

For a two-dimensional problem, in the domain, say  $x > 0, y \in \mathbb{R}, t > 0$ , taking into account the fact that two eigenvalues collapse, the corresponding inflow–outflow problems lead to 4, 3, 1, and 0 prescribed boundary conditions.

### 3.2 Solid or Rigid Wall Boundary

The usual “slip boundary condition” is prescribed:

$$\mathbf{u} \cdot \mathbf{n} = \mathbf{0}, \quad (3.2)$$

which means that the flow does not cross the boundary but may move tangentially.

*Remark 3.2.* In the case of Maxwell system (Example 1.2), the corresponding boundary condition would be

$$\mathbf{E} \times \mathbf{n} = \mathbf{0}; \quad (3.3)$$

the notation  $\mathbf{E} \times \mathbf{n}$  denotes the vector  $(E_y n_z - E_z n_y, E_z n_x - E_x n_z, E_x n_y - E_y n_x)^T$ . Condition (3.3) means that the boundary is a perfect conductor. Then Faraday’s law gives

$$\mathbf{B} \cdot \mathbf{n} = \mathbf{B}_0 \cdot \mathbf{n}$$

(see Dautray and Lions [390], Chapter 1, Part A, §4, Section 2.4.3). □

## 4 Absorbing Boundary Conditions

In general, a boundary condition is called “exact” if the boundary is “transparent,” i.e., the (approximate) solution, in the finite domain with artificial boundary obtained with this boundary condition, coincides with the exact solution in the unbounded domain. The boundary condition is called “absorbing” if it yields the decreasing with time of some energy function. One also speaks of “radiation” or “nonreflecting” boundary conditions if they allow the wave motion to pass through the boundary of the domain without generating reflections back into the interior, or at least with a reduced amount of spurious reflection, but allow true physical reflections (see Higdon [614] and Hagstrom and Hariharan [576]; we refer to [521], for a review of the problem).

If one uses normal mode analysis, one can derive exact or “perfectly absorbing” boundary conditions at normal incidence (or in dimension one) by imposing boundary conditions that annihilate the outgoing waves (or prevent the generation of incoming waves). Otherwise, one minimizes the amplitude of waves reflected from the artificial boundary (but the reflection coefficients depend on the incidence). Let us illustrate these ideas with a very simple example.

*Example 4.1.* Consider the wave equation

$$\frac{\partial^2 \rho}{\partial t^2} - c^2 \frac{\partial^2 \rho}{\partial x^2} = 0, \quad c > 0.$$

which is often taken as the simplest example for illustrating absorbing boundary conditions. The solution of this equation is a function of the form  $f(x - ct) + g(x + ct)$ , i.e., two waves traveling to the right (resp. to the left) with constant speed  $c$  (resp.  $-c$ ). If we put an artificial boundary at  $x = 0$  (domain  $x > 0$ ), we want to let the wave  $g(x + ct)$  that travels to the left leave the domain, and we do not want to let a wave enter the domain; thus, we do exclude waves of the form  $f(x - ct)$ . The perfectly absorbing condition is the Sommerfeld condition (see Givoli and Cohen [522]):

$$\frac{\partial \rho}{\partial t} - c \frac{\partial \rho}{\partial x} = 0 \text{ on } x = 0, \quad t > 0. \quad (4.1)$$

For an artificial boundary at  $x = 1$  (domain  $x < 1$ ), it would be

$$\frac{\partial \rho}{\partial t} + c \frac{\partial \rho}{\partial x} = 0 \text{ on } x = 1, \quad t > 0.$$

The motivation to study this equation here can be found in Chap. V, Remark 2.5. The wave equation was obtained from the linear system

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho_0 u) &= 0, \\ \frac{\partial}{\partial t}(\rho_0 u) + c_0^2 \frac{\partial \rho}{\partial x} &= 0, \end{aligned}$$

which we write

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} = 0;$$

it is hyperbolic, with eigenvalues  $-c_0, c_0$  and corresponding eigenvectors  $\mathbf{r}_1 = (-1, c_0)^T$  and  $\mathbf{r}_2 = (1, c_0)^T$ , which give the columns of  $\mathbf{T}$ . The characteristic variables are  $\mathbf{w} = \mathbf{T}^{-1}\mathbf{u}$ , i.e.,  $w_1 = -c_0\rho + \rho_0 u$  and  $w_2 = c_0\rho + \rho_0 u$ , which thus satisfy a decoupled system

$$\begin{aligned} \frac{\partial w_1}{\partial t} - c_0 \frac{\partial w_1}{\partial x} &= 0, \\ \frac{\partial w_2}{\partial t} + c_0 \frac{\partial w_2}{\partial x} &= 0. \end{aligned}$$

We can annihilate the rightward traveling waves by setting

$$w_2 = c_0\rho + \rho_0 u = 0. \quad (4.2)$$

Condition (4.2) implies (4.1) (with  $c \equiv c_0$ ) since then

$$\frac{\partial \rho}{\partial t} - c_0 \frac{\partial \rho}{\partial x} = \frac{\partial \rho}{\partial t} - c_0 \left( \frac{-\rho_0}{c_0} \right) \frac{\partial u}{\partial x} = 0.$$

Conversely, if (4.1) is satisfied,

$$\frac{\partial w_2}{\partial t} = c_0 \frac{\partial \rho}{\partial t} + \rho_0 \frac{\partial u}{\partial t} = c_0^2 \frac{\partial \rho}{\partial x} + \rho_0 \frac{\partial u}{\partial t} = 0;$$

similarly,  $\frac{\partial w_2}{\partial x} = 0$ , and  $w_2 = 0$  up to a constant.

Let us now use normal mode analysis as an introduction to the two-dimensional case that follows. We note that a mode  $e^{ik(x+c_0t)}$ , with phase velocity  $-c_0$ , travels out of the domain  $x > 0$ . We want to prevent waves of the form  $e^{ik(x-c_0t)}$  from entering. By Fourier transform, if we define

$$\hat{\mathbf{u}}(x, \omega) = \int e^{+i\omega t} u(x, t) dt$$

(the + sign in  $e^{+i\omega t}$  is taken here only for convenience), the linear system becomes

$$\frac{\partial \hat{\mathbf{u}}}{\partial x} = i\omega \mathbf{A}^{-1} \hat{\mathbf{u}}(x, \omega),$$

and diagonalizing  $\mathbf{A}^{-1} = \mathbf{T} \mathbf{A}^{-1} \mathbf{T}^{-1}$ , setting  $\hat{\mathbf{v}} = T^{-1} \hat{\mathbf{u}}$ , it decouples into

$$\begin{aligned} \frac{\partial \hat{v}_1}{\partial x} &= i\left(\frac{\omega}{c_0}\right) \hat{v}_1(x, \omega), \\ \frac{\partial \hat{v}_2}{\partial x} &= -i\left(\frac{\omega}{c_0}\right) \hat{v}_2(x, \omega). \end{aligned}$$

If we set

$$\hat{v}_1 = 0 \text{ at } x = 0, \quad (4.3)$$

(here the subscript 1 corresponds to the positive eigenvalue  $+c$ ), there will be no incoming wave, since then

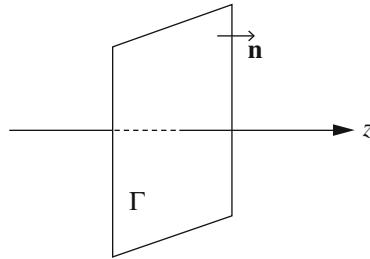
$$\begin{aligned} \mathbf{u} &= \int e^{-i\omega t} \hat{\mathbf{u}}(x, \omega) d\omega = \int e^{-i\omega t} \mathbf{T} \hat{\mathbf{v}}(x, \omega) d\omega \\ &= \int e^{-i\omega t - i\omega x/c_0} \hat{v}_2(0, \omega) \mathbf{r}_2 d\omega \end{aligned}$$

is only a function of  $x + c_0 t$ . Thus, these approaches give the same result. They provide a perfectly absorbing condition in one dimension or at normal incidence in several dimensions since a wave  $e^{ik(x+ct)}$  can be regarded as a wave in several dimensions traveling at normal incidence to the boundary  $x = 0$ .  $\square$

*Example 1.2* (Revisited). The same approach applies to the (homogeneous) Maxwell system in one dimension. Assuming slab symmetry, the variables depend only on  $z$  and  $t$ ,

$$\mathbf{E} = \mathbf{E}(z, t), \quad \mathbf{B} = \mathbf{B}(z, t)$$

(Fig. 4.1).  $E_z$  and  $B_z$  are independent of  $(z, t)$  (no propagation), and  $(E_x, E_y)$ ,  $(B_x, B_y)$  satisfy (1.14a),



**Fig. 4.1** Boundary for Maxwell system with slab symmetry

$$\frac{\partial E_x}{\partial t} + c^2 \frac{\partial B_y}{\partial z} = 0,$$

$$\frac{\partial E_y}{\partial t} + c^2 \frac{\partial B_x}{\partial z} = 0,$$

together with (1.14b)

$$\frac{\partial B_y}{\partial t} + \frac{\partial E_x}{\partial z} = 0,$$

$$\frac{\partial B_x}{\partial t} + \frac{\partial E_y}{\partial z} = 0.$$

The system can also be written in the characteristic form of two waves propagating with characteristic velocity  $\pm c$ :

$$\frac{\partial}{\partial t} \left\{ \begin{pmatrix} E_x \\ E_y \end{pmatrix} + c \begin{pmatrix} B_y \\ -B_x \end{pmatrix} \right\} + c \frac{\partial}{\partial z} \left\{ \begin{pmatrix} E_x \\ E_y \end{pmatrix} + c \begin{pmatrix} B_y \\ -B_x \end{pmatrix} \right\} = 0,$$

$$\frac{\partial}{\partial t} \left\{ \begin{pmatrix} E_x \\ E_y \end{pmatrix} - c \begin{pmatrix} B_y \\ -B_x \end{pmatrix} \right\} - c \frac{\partial}{\partial z} \left\{ \begin{pmatrix} E_x \\ E_y \end{pmatrix} - c \begin{pmatrix} B_y \\ -B_x \end{pmatrix} \right\} = 0.$$

Therefore, the incoming (resp. outgoing) wave corresponds to

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} - c \begin{pmatrix} B_y \\ -B_x \end{pmatrix} \quad (\text{resp. } \begin{pmatrix} E_x \\ E_y \end{pmatrix} + c \begin{pmatrix} B_y \\ -B_x \end{pmatrix}) = 0.$$

The perfectly absorbing boundary condition corresponds to no incoming wave (no reflected wave at the artificial boundary),

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} - c \begin{pmatrix} B_y \\ -B_x \end{pmatrix} = \mathbf{0} \iff (\mathbf{E} - c\mathbf{B} \times \mathbf{n}) \times \mathbf{n} = \mathbf{0}$$

on the artificial boundary, which are called the Silver–Müller boundary conditions. When applied more generally, they yield perfect absorption for plane

waves at normal incidence. We can also admit a given incoming wave

$$(\mathbf{E} - c\mathbf{B} \times \mathbf{n}) \times \mathbf{n} = \mathbf{e} \times \mathbf{n} \text{ on the artificial boundary,}$$

or equivalently

$$\left( \mathbf{B} + \left( \frac{1}{c} \right) \mathbf{E} \times \mathbf{n} \right) \times \mathbf{n} = \mathbf{b} \times \mathbf{n}.$$

See Cioni et al. [310] for the numerical approximation.

Consider now a linear, strictly hyperbolic system with constant coefficients, in dimension  $d = 2$ , in the whole space

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A} \frac{\partial \mathbf{u}}{\partial x} + \mathbf{B} \frac{\partial \mathbf{u}}{\partial y} = 0;$$

for example, consider the case of linearized isentropic Euler equations (Example 1.1). We set the artificial boundary at  $x = 0$  and assume that it is noncharacteristic; the interior of the computational domain corresponds to  $x > 0$ , and we thus recover problem (1.10a). Applying a Fourier transform w.r.t. to  $y, t$  yields

$$\frac{\partial \hat{\mathbf{u}}}{\partial x} = \mathbf{A}^{-1} (i\omega \mathbf{I} - i\eta \mathbf{B}) \hat{\mathbf{u}} = \mathbf{D}(\eta, i\omega) \hat{\mathbf{u}},$$

where  $\hat{\mathbf{u}} = \hat{\mathbf{u}}(x, \eta, \omega) = \iint e^{-i\eta y} e^{+i\omega t} u(x, y, t) dy dt$ , and following arguments similar to those of Sect. 1.3.3, we can derive “perfectly absorbing” boundary conditions by annihilating the wave entering the domain. Recall that  $\mathbf{D}(0, i) = i\mathbf{A}^{-1}$  has distinct (purely imaginary) eigenvalues and for  $|\frac{\eta}{\omega}| + |\omega|$  small enough  $\mathbf{D}(\eta, i\omega)$  has also distinct purely imaginary eigenvalues (we have seen that it is the case if  $(\eta, \omega)$  lies in  $\Xi$ ; see Fig. 1.4).  $\mathbf{D}(\eta, i\omega)$  is diagonalizable in this neighborhood, which geometrically corresponds to near-normal incidence; indeed, in the present situation (half-space problem), the pair  $(\eta, \omega) = (0, 1)$  corresponds to normal incidence. Therefore, the eigenvalues of  $\mathbf{D}(\eta, i\omega)$  are of the form  $i\xi_k$ , where  $\xi = \xi(\eta, \omega)$  is real. Following (1.15), let us set

$$\hat{\mathbf{v}}(x, \eta, \omega) = \mathbf{Q}(\eta, i\omega), \quad \hat{\mathbf{u}} = (\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2)^T,$$

where  $\mathbf{Q}^{-1}(\eta, i\omega)$  is the matrix with columns the eigenvectors of  $\mathbf{D}(\eta, i\omega)$  ( $\mathbf{Q}$  is the matrix with rows the eigenvectors  $\mathbf{l}_i^T, \mathbf{l}_i$ , the eigenvectors of  $\mathbf{D}^T$ , and  $\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}(\eta, i\omega)$  is diagonal), and the eigenvalues of  $\mathbf{D}_1$  (resp.  $\mathbf{D}_2$ ) are such that  $\operatorname{sgn} \xi(\frac{\eta}{\omega}, 1) \geq 0$  (resp.  $\leq 0$ ). The system satisfied by  $\hat{\mathbf{u}}$  is decoupled,

$$\frac{\partial \hat{\mathbf{v}}_i}{\partial x} = \mathbf{D}_i(\eta, i\omega) \hat{\mathbf{v}}_i,$$

and thus

$$\hat{v}_{1k}(x, \eta, \omega) = \exp(i\xi(\eta, \omega)) \hat{v}_{1k}(0, \eta, \omega),$$

and  $\hat{\mathbf{v}}_1$  corresponds to the incoming waves  $\xi_k(\frac{\eta}{\omega}, 1) \geq 0$ . Indeed the phase velocity of a wave  $e^{+i\eta y - i\omega t + i\xi x}$ , which is involved in the inverse Fourier transform  $\mathbf{u} = \int \int e^{+i\eta y - i\omega t} \hat{\mathbf{u}}(x, \eta, \omega) d\eta d\omega$ , is the vector  $\frac{(\omega\xi, \omega\eta)^T}{|\mathbf{k}|^2}$ , and it enters the domain if  $\operatorname{sgn} \omega\xi \geq 0$  and thus is an incoming wave if  $\operatorname{sgn} (\omega\xi(\eta, \omega)) = \operatorname{sgn} \xi(\frac{\eta}{\omega}, 1) \geq 0$ . Then, the condition

$$\hat{\mathbf{v}}_1 = 0 \text{ at } x = 0 \quad (4.4)$$

is the “perfectly absorbing” boundary condition. It can be equivalently written as

$$\pi_q \mathbf{Q} \hat{\mathbf{u}} = \mathbf{0},$$

where  $\pi_q$  denotes the projection

$$\pi_q(v_1, \dots, v_p) = (v_1, \dots, v_q)$$

on the coordinates related to the “incoming” eigenvalues of  $\mathbf{D}(\eta, i\omega)$ , i.e., as we have just seen, related to eigenvalues with positive imaginary part  $\xi_k(\frac{\eta}{\omega}, 1) \geq 0$ .

However, since the Fourier transform is used and  $\mathbf{Q}$  is not constant (it depends on  $\eta, \omega$ ), this provides global conditions for  $\mathbf{u}$ . Therefore, local conditions (boundary values of the incoming variables  $\mathbf{u}^I$  are given by a function of  $\mathbf{u}^{II}$  at the same place and time) should be derived by some approximation w.r.t.  $\frac{\eta}{\omega}$ . Let  $\mathbf{P}^{-1}(\eta, \omega)$  denote the matrix with columns the eigenvectors of  $\mathbf{D}(\frac{\eta}{\omega}, i) = (\frac{1}{\omega})\mathbf{D}(\eta, i\omega)$  ( $\mathbf{Q}^{-1}(\eta, \omega)$  denotes the matrix with columns the eigenvectors of  $\mathbf{D}(\eta, i\omega)$ ). We assume that we can write near-normal incidence (for  $|\frac{\eta}{\omega}| + |\omega|$  small enough)

$$\mathbf{P}(\eta, \omega) = \mathbf{V}(0, 1) + \frac{\eta}{\omega} \frac{\partial}{\partial \eta} \mathbf{P}(0, 1) + O\left(\frac{\eta}{\omega}\right)^2,$$

where  $\frac{\partial}{\partial \eta} \mathbf{P}$  denotes the derivative w.r.t. the first variable. The condition (4.4) yields

$$\pi_q \left\{ \mathbf{P}(0, 1) + \frac{\eta}{\omega} \frac{\partial}{\partial \eta} \mathbf{P}(0, 1) + O\left(\frac{\eta}{\omega}\right)^2 \right\} \hat{\mathbf{u}} = \mathbf{0}.$$

We get then first-order approximating boundary conditions

$$\pi_q^0 \mathbf{P}(0, 1) \hat{\mathbf{u}} = \mathbf{0} \quad \text{at } x = 0,$$

which by inverse transform gives

$$\pi_q^0 \mathbf{P}(0, 1) \mathbf{u} = \mathbf{P}_I(0, 1) \mathbf{u} = \mathbf{0}.$$

Here  $\mathbf{P}_I$  denotes the matrix with lines  $\mathbf{l}_i^T$  corresponding to the  $q$  eigenvectors of  $\mathbf{A}^{-1}$ , i.e., of  $\mathbf{A}$  associated with positive eigenvalues.

*Remark 4.1.* We have previously denoted  $\mathbf{P}(0, 1)$  by  $\mathbf{T}^{-1}$  (for instance, in (1.8), Chap. IV). Some elementary algebra gives

$$\mathbf{u} = \sum v_i \mathbf{r}_i = \mathbf{P}^{-1} \mathbf{v} = \mathbf{T} \mathbf{v} \text{ (where } v_i = \mathbf{l}_i^T \mathbf{u}) \implies (v_1, \dots, v_q) = \mathbf{P}_I \mathbf{u},$$

or

$$\mathbf{l}_i^T \mathbf{u} = \mathbf{0} \text{ at } x = 0.$$

$\mathbf{P}_I \mathbf{u}$  is the projection on the space spanned by the eigenvectors corresponding to “incoming” eigenvalues of  $\mathbf{A}^{-1}$ , i.e., to positive eigenvalues of  $\mathbf{A}$ . This is usually denoted as above with a projection operator, say  $\pi_q^0$ , which coincides with  $\pi_q$  if we assume that the eigenvalues do not vanish in the neighborhood.  $\square$

This means exactly that the characteristic variables corresponding to incoming characteristics are set to zero. We can also admit a given incoming wave, “ $\mathbf{P}_I(0, 1)\mathbf{u}$  given at  $x = 0$ ,” and we recover the boundary condition (1.8).

A second-order approximation is

$$\pi_q^0 \left\{ \omega \mathbf{P}(0, 1) + \eta \frac{\partial}{\partial \eta} \mathbf{P}(0, 1) \right\} \hat{\mathbf{u}} = \mathbf{0},$$

to which we apply again the inverse transform; now the variables  $\eta, \omega$  in the symbol correspond by inverse transform to partial differential operators, and we get

$$\pi_q^0 \left( \mathbf{P}(0, 1) \frac{\partial u}{\partial t} + \left( \frac{\partial}{\partial \eta} \mathbf{P}(0, 1) \right) \frac{\partial u}{\partial y} \right) = 0 \text{ at } x = 0.$$

Again, this is just an outline, and for more details concerning the derivation of these conditions, we refer to Engquist and Majda [460] and Kröner [712]. The first authors apply the theory to the particular example of the linearized shallow water equation, whereas the last author studies the linearized Euler system in dimension  $d = 2$  ( $p = 4$ ) and derives precisely the corresponding first-order and second-order absorbing boundary conditions by computing explicitly  $\mathbf{V}(\eta, \omega)$ . For instance, a subsonic outflow requires one condition, which often reduces to fixing the pressure at the boundary. However, it can reflect pressure disturbance back into the computational domain, and a first-order approximation of the nonreflecting boundary condition is  $p - \bar{p}cu = 0$  ( $p - pc$  is the incoming characteristic variable; see Remark 1.2). See also Higdon [615], Gustafsson and Fern [571], Jiang and Wong [656], Hagstrom and Hariharan [576], Rudy and Strikwerda [995], and Bayliss and Turkel [94].

*Remark 4.2.* In fact, Engquist and Majda [460] do not work with the Fourier transform but with the corresponding differential operators. The matrix  $\mathbf{D}(\eta, -i\omega) = -\mathbf{A}^{-1}(i\omega \mathbf{I} + i\eta \mathbf{B})$  (which they denote by  $\mathbf{M}(\eta, \omega)$ ) corresponds to the symbol of the differential operator  $\mathbf{A}^{-1}\left(\frac{\partial}{\partial t}\right) + \mathbf{A}^{-1}\mathbf{B}\left(\frac{\partial}{\partial y}\right)$  that results from rewriting (1.10a) as

$$\frac{\partial \mathbf{u}}{\partial x} + \mathbf{A}^{-1} \left( \frac{\partial \mathbf{u}}{\partial t} \right) + \mathbf{A}^{-1} \mathbf{B} \left( \frac{\partial \mathbf{u}}{\partial y} \right) = \mathbf{0},$$

and then the matrices  $\mathbf{D}_i$  in (1.15) are the symbols of differential operators, more precisely pseudo-differential operators, since the matrices depend on  $x, y$ . We refer to [460], Section 2, for more details.  $\square$

The approach followed by Hedstrom [597] and Thompson [1121] for a nonlinear hyperbolic system relies on characteristics. In one dimension (see Chap. II, Sect. 5), we have written the system in characteristic form,

$$\mathbf{l}_i^T(\mathbf{u}) \left\{ \frac{\partial \mathbf{u}}{\partial t} + \lambda_i(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x} \right\} = \mathbf{0}.$$

Hedstrom's nonreflecting boundary condition can be written as

$$\mathbf{l}_i^T(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial t} = 0, \quad \text{for all } i \text{ such that } \lambda_i(\mathbf{u}) > 0,$$

i.e., the amplitude of an incoming wave is constant with time at the boundary. If the system is linear, it means exactly that the characteristic variables corresponding to incoming waves are constant. In the nonlinear case, he shows that if there are only simple waves going out, this condition gives no wave coming into the domain from the boundary  $x = 0$ . Otherwise, the strength of the reflected shock is of order 3. In the previous example of subsonic outflow, it gives  $\frac{\partial p}{\partial t} - \rho c \frac{\partial u}{\partial t} = 0$ . Such a boundary condition is found, for instance, in Cambier, Escande and Veuillot [226].

Thompson has extended this condition to the multidimensional case

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) + \frac{\partial}{\partial y} \mathbf{g}(\mathbf{u}) = \mathbf{0}. \quad (4.5)$$

Considering as previously a boundary  $x = 0$  of the domain  $x > 0$  and defining

$$\mathbf{L}_i(\mathbf{u}) = \mathbf{l}_i^T(\mathbf{u}) \lambda_i(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial x},$$

where  $\lambda_i(\mathbf{u}) = \lambda_i(\mathbf{u}, \mathbf{e}_1)$  are the eigenvalues of  $\mathbf{A} = \mathbf{f}'$  and  $\mathbf{l}_i^T(\mathbf{u})$  the eigenvectors of  $\mathbf{A}^T$ , the characteristic equations of the system (4.4) projected on the normal  $\mathbf{e}_1$  to the boundary (see (1.5), Chap. V, Sect. 1) are

$$\mathbf{l}_i^T(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial t} + \mathbf{L}_i(\mathbf{u}) = -\mathbf{l}_i^T(\mathbf{u}) \mathbf{B}(\mathbf{u}) \frac{\partial \mathbf{u}}{\partial y}. \quad (4.6)$$

The terms on the right-hand side contain derivatives in the direction transverse to the boundary and may be evaluated from values in the interior, as well as the terms corresponding to outgoing waves. There remains to specify "nonreflecting" boundary conditions that determine the values of  $\mathbf{L}_i(\mathbf{u})$  for incoming waves ( $\lambda_i(\mathbf{u}) > 0$ ) by requiring as above that the amplitude of

an incoming wave remain constant in time. For subsonic outflow, the condition can be written as  $\frac{\partial p}{\partial t} - \rho c \frac{\partial u}{\partial t} = 0$ , where  $u$  now represents the normal component of the velocity; we refer to Thompson [1121] for details; see also [93, 956, 1077, 1161].

## 5 Numerical Treatment

Concerning the numerical treatment, we have to distinguish between the given “analytical” boundary conditions that follow from the considerations of the preceding sections and the “numerical” conditions required for computation: if a variable cannot be prescribed, its value may be needed in the computation, and eventually one must describe the way to compute it. For example, we have just seen that for subsonic inflow, two conditions, for instance,  $(\rho, u)$ , are specified, but one must still specify the value of  $p$ . Now, there are many ways of deriving a boundary treatment, so we just give some examples, and the reader is asked to refer to the papers cited for details.

### 5.1 Finite Difference Schemes

For a finite difference scheme, in the boundary condition (1.8),  $\mathbf{w}_0^{I,n} \sim \mathbf{w}_1(0, t_n)$  is prescribed (given by  $\mathbf{g}^I(t_n)$ ), whereas the nonspecified value  $\mathbf{w}_0^{II,n} \sim \mathbf{w}^{II}(0, t_n)$  must be computed from the values or from the differential equation in the interior of the domain: the value  $\mathbf{w}^{II}(0, t_n)$  is therefore approximated either by extrapolation or by a finite difference discretization. Note that “ $\mathbf{w}^I(0, t_n)$  is prescribed” means that the information for  $\mathbf{w}^I(0, t)$  cannot be extracted solely from the differential equations in the interior of the domain as for  $\mathbf{w}^{II}(0, t)$ . If  $\mathbf{g}^I(t)$  is not exactly known, as can be the case in some practical situations where no data are available, it must still be determined by additional information, or else instability occurs (see Gustafsson [570] and Tadmor [1087]).

*Example 5.1.* Consider a 3-point (linear) conservative scheme for approximating the simple advection equation (1.1) on the strip  $0 < x < 1$ . The boundary condition (see Sect. 1.1.1) is  $u(0, t) = g(t)$  if  $a > 0$ , and there is no boundary condition at  $x = 1$ . In the formula

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

where  $0 \leq j \leq N$ , and  $\Delta x = \frac{1}{N+1}$ ,  $u_0^n = g(n\Delta t)$  is given by the boundary condition, whereas the value  $u_{N+1}^n$  is needed and should be computed from the interior values.

The simplest zeroth-order (locally first-order) extrapolation is

$$u_{N+1}^n = u_N^n \text{ (horizontal),}$$

and a first-order (or linear) extrapolation is

$$u_{N+1}^n = 2u_N^n - u_{N-1}^n.$$

Other possible oblique extrapolations are

$$\begin{aligned} u_{N+1}^n &= u_N^{n-1} \text{ (zeroth order) or} \\ u_{N+1}^n &= 2u_N^{n-1} - u_{N-1}^{n-2} \text{ (first order).} \end{aligned}$$

Otherwise, we can use the upwind scheme (first order)

$$u_{N+1}^{n+1} = u_{N+1}^n - \lambda a(u_{N+1}^n - u_N^n),$$

and one can also use implicit schemes (Beam-Warming [95] and with Yee [96]). The order of accuracy is studied as always by means of Taylor expansion.

If one uses a 5-point scheme, one still sets  $u_0^n = g(t_n)$ , and  $u_{-1}^n$  will be computed from the Taylor expansion w.r.t  $x$ , where the space derivatives are replaced by time derivatives of  $g$ , thanks to Eq. (1.1),

$$\frac{\partial u}{\partial x}(0, t_n) = -\left(\frac{1}{a}\right)g'(t_n),$$

and so on, which gives, for instance,

$$u_{-1}^n = g(t_n) + \Delta x \frac{g'(t_n)}{a}.$$

One then needs  $u_{N+1}^n$  and  $u_{N+2}^n$ . If one takes for  $u_{N+1}^n$  and  $u_{N+2}^n$  the same formula, i.e., the same coefficients in the extrapolation or difference scheme, for example,

$$\begin{aligned} u_{N+1}^n &= 2u_N^n - u_{N-1}^n \text{ and} \\ u_{N+2}^n &= 2u_{N+1}^n - u_N^n, \end{aligned}$$

the boundary conditions are called “translatory.” □

*Example 5.2.* For a linear system, a usual approach introduces “compatibility relations.” The boundary condition (1.8b)

$$\mathbf{w}^I(0, t) = \mathbf{S}^I \mathbf{w}^{II}(0, t) + \mathbf{g}^I(t)$$

provides  $q$  relations at the boundary. The  $p - q$  other relations for  $\mathbf{w}^{II}$  are obtained from the discretization of the  $p - q$  differential equations in the characteristic outgoing variables

$$\frac{\partial \mathbf{w}^{II}}{\partial t} + \mathbf{A}_{II} \frac{\partial \mathbf{w}^{II}}{\partial x} = \mathbf{0}.$$

As above in the scalar case, one uses an upwind scheme for the spatial derivative and then computes  $\mathbf{w}^{II}(0, t_{n+1})$  from  $\mathbf{w}(0, t_n)$ , where  $\mathbf{w}^{II}(0, t_n)$  is given by the scheme and  $\mathbf{w}^I(0, t_n)$  by  $\mathbf{S}^I \mathbf{w}^{II}(0, t_n) + \mathbf{g}^I(t_n)$ .  $\square$

*Example 5.3.* In two dimensions, for instance, the linearized Euler system, one also uses the characteristic equations of the system projected on the outward normal  $\boldsymbol{\nu}$  to the boundary (see Chap. V, Sect. 1, (1.8))

$$\mathbf{l}_k^T(\mathbf{u}, \boldsymbol{\nu}) \left\{ \frac{\partial \mathbf{u}}{\partial t} + \lambda_k(\mathbf{u}, \boldsymbol{\nu}) \frac{\partial \mathbf{u}}{\partial \boldsymbol{\nu}} \right\} = S_k,$$

where the right-hand side  $S_k$  involves derivatives in the tangential direction  $\boldsymbol{\nu}^\perp$  only. For the Euler system that is invariant by rotation, it is equivalent to consider a boundary of the type  $x = 0$ . One keeps only the equations corresponding to outgoing characteristics for the one-dimensional projected system (i.e., to positive eigenvalues). The equations are discretized, using the same upwind scheme as in the interior and without taking into account the boundary conditions, using extrapolations to estimate the derivatives or modifying the scheme so as to involve only interior points; the term  $S_k$  may be considered as known from values inside the domain. This gives a value noted, say  $\mathbf{u}^*$ ; then  $\mathbf{u}^{n+1}$  satisfies the “compatibility relations,” which are

$$\mathbf{l}_k^T(\mathbf{u}^*, \boldsymbol{\nu}) \{ \mathbf{u}^* - \mathbf{u}^{n+1} \} = 0, \text{ if } \lambda_k(\mathbf{u}^*, \boldsymbol{\nu}) \geq 0;$$

the other relations needed in order to compute  $\mathbf{u}^{n+1}$  are given by the prescribed boundary data if  $\lambda_k(\mathbf{u}^*, \boldsymbol{\nu}) \leq 0$ . The expression is often linearized by taking  $\mathbf{l}_k^T(\mathbf{u}^n, \boldsymbol{\nu})$  instead of  $\mathbf{l}_k^T(\mathbf{u}^*, \boldsymbol{\nu})$ . This approach is well suited for subdomain computations (Veuillot and Cambier [1168]).

Note that if we take for  $\mathbf{u}$  the primitive variables  $(\rho, u_\nu, u_\xi, p)$ , the expressions for if  $\mathbf{l}_i^T(\mathbf{u}, \boldsymbol{\nu})$  are very simple (see Cambier et al. [227]).

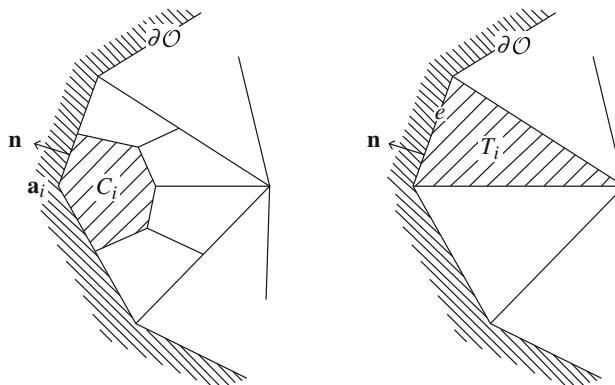
The same idea is used by Hagstrom and Hariharan [576] for the numerical treatment of their “reflecting” boundary conditions.  $\square$

For what concerns stability, the relevance of these different boundary schemes results from the G.K.S. theory (for Gustafsson, Kreiss, Sundstrom); see Gustafsson et al. [574], Gustafsson [570], and Goldberg and Tadmor for convenient stability criteria [546–548]; Trefethen [1133, 1134], Beam, Warming, and Yee [96], Sod [1069], Daru and Lerat [389]. However, this stability analysis does not extend to the nonlinear case, and linear extrapolation may be unsatisfactory [673]).

In fact, let us see that the finite volume method offers a very simple treatment of the boundary via fluxes, which does not require the explicit computation of the nonprescribed variables.

## 5.2 Finite Volume Approach

Assume that the boundary of the spatial domain  $\mathcal{O}$  is made up of boundaries of the finite volumes (one uses a “body-fitted” grid and, for another approach, the so-called Cartesian grid method, see, for instance, Pember et al. [936]). Following the arguments of Chap. V, Sect. 4.1.1, we have to integrate the system over a boundary element, say  $\Omega_i$ , and thus to compute a flux through the boundary of  $\Omega_i$ . We have already taken care of the parts of  $\partial\Omega_i$  adjacent to another cell, and there remains to approximate the flux through an edge  $e$  that is on the boundary of  $\mathcal{O}$  (Fig. 5.1). We denote this flux by  $\Phi(\mathbf{V}_i, \mathbf{V}_\infty)$  and describe its computation according to the different cases. The notation  $\mathbf{V}_\infty$  is not significant: it refers to a state in the far field, which in fact will not always be the case.



**Fig. 5.1** Boundary cell in the cell vertex and cell center approaches

### 5.2.1 Solid Wall Boundary

Taking into account the slip boundary condition  $\mathbf{u} \cdot \mathbf{n} = 0$ , we see that the only contribution to the exact flux comes from the pressure

$$\mathbf{F} \cdot \mathbf{n} = (0, p_b \mathbf{n}, 0)^T.$$

This pressure  $p_b$  at the body can be estimated differently according to the chosen method. (i) In the case of a node-based scheme (cell vertex), set  $p_b = p_i$ . (ii) In the case of a center-based scheme, given a state  $\mathbf{V}_i$  in a cell next to the boundary, one introduces a mirror state  $\tilde{\mathbf{V}}_i$ , with the same  $\rho_i, p_i, \mathbf{u}_{\tau i} = \mathbf{u}_i \cdot \mathbf{n}^\perp$ , only opposite normal velocity  $\tilde{u}_{ni} = -\mathbf{u}_i \cdot \mathbf{n}$ , and then one solves the corresponding Riemann problem in an exact or approximate way (it can be viewed as an external state in a fictitious cell opposite to the edge  $e$ )

$$\Phi(\mathbf{V}_i, \mathbf{V}_\infty) \equiv \Phi(\mathbf{V}_i, \tilde{\mathbf{V}}_i). \quad (5.1)$$

The exact solution at  $\xi = 0$  of the one-dimensional Riemann problem corresponding to  $\mathbf{U}_L = \mathbf{U} = (\rho, -u, p)$  and  $\mathbf{U}_R = \tilde{\mathbf{U}} = (\rho, u, p)$ , with, for instance,  $u > 0$  (and  $u$  not too large), is a 1-rarefaction connecting  $\mathbf{U}$  to  $\mathbf{U}^* = (\rho_I, 0, p^*)$ , followed by a 2-contact discontinuity on  $x = 0$  connecting to  $(\rho_{II}, 0, p^*)$ , and a 3-rarefaction connecting to  $\tilde{\mathbf{U}}$  (see Chap. III, Sect. 3). The one-dimensional Godunov flux gives

$$\mathbf{g}^{\text{God}}(\mathbf{U}, \tilde{\mathbf{U}}) = (0, p^*, 0)^T.$$

The expressions with Roe's scheme are also very simple, since Roe's average state is easily computed (see Chap. IV, Sect. 4),

$$\bar{H} = H, \quad \bar{u} = 0, \quad \bar{c} = \kappa \left( H - \frac{u^2}{2} \right) = c, \quad \bar{\rho} = p.$$

Roe's flux gives

$$\mathbf{g}^{\text{Roe}}(\mathbf{U}, \tilde{\mathbf{U}}) = (0, p_b, 0)^T,$$

where

$$p_b = p + \rho u^2 + \rho c u.$$

Thus, if on the right-hand side of (5.1) one uses Godunov's flux, we get for the boundary flux  $\Phi^{\text{God}}(\mathbf{V}_i, \tilde{\mathbf{V}}_i) = (0, p^* \mathbf{n}, 0)^T$ , where  $p^*$  is the pressure “on the boundary,” i.e., the pressure of the abovementioned intermediate state in the solution of the Riemann problem.

Now, if one uses Roe's flux, an easy computation (see Chap. V, Sect. 4.3.1) shows that Roe's intermediate state is such that

$$\begin{aligned} \bar{\rho}_i &= \rho_i, & \bar{H}_i &= H_i, & \bar{u}_{\tau i} &= \bar{\mathbf{u}}_i \cdot \mathbf{n}^\perp = u_{\tau i}, & \bar{u}_{ni} &= \bar{\mathbf{u}}_i \cdot \mathbf{n} = 0, \\ \bar{c}_i &= \kappa \left( H_i - \frac{u_{\tau i}^2}{2} \right), \end{aligned}$$

and then one computes Roe's flux (using formula in Chap. IV, Sect. 4), which in the present case are very simple because in (4.29), for instance,  $\Delta\rho = \Delta p = 0$ ; this yields

$$\Phi^{\text{Roe}}(\mathbf{V}_i, \tilde{\mathbf{V}}_i) = (0, p_b \mathbf{n}, 0)^T,$$

with

$$p_b = p_i + \rho_i u_{ni}^2 + \rho_i \bar{c}_i u_{ni}.$$

We might also use other schemes.

### 5.2.2 Fluid Boundary

Be aware that  $\mathbf{n}$  is the outward unit normal to  $e$  in  $\Omega_i$ , and thus in the case of a fluid boundary, in the above one-dimensional linearized study,  $-\mathbf{n}$  gives the direction  $x > 0$ , and inflow corresponds to  $\mathbf{u} \cdot \mathbf{n} < 0$ , outflow to  $\mathbf{u} \cdot \mathbf{n} > 0$ , and the prescribed boundary conditions to negative eigenvalues. The state  $\mathbf{V}_\infty$  is a state in the far field, which represents the flow in a fictitious cell adjacent to the boundary that “satisfies the linearized boundary conditions” in the sense that for a subsonic inflow (for instance, stagnant gas taken from a reservoir expanding in a nozzle), total enthalpy and physical entropy of  $\mathbf{V}_\infty$  are specified; for a subsonic outflow (for instance, flow in a tunnel exhausting into the atmosphere), the pressure is prescribed; and for supersonic inflow (external flow), the whole state is given, whereas for supersonic outflow, no condition is required on  $\mathbf{V}_\infty$ . The reason for specifying enthalpy  $H$  for a subsonic inflow is that  $H$  is constant along streamlines for steady adiabatic flow; thus, for a flow originating from a reservoir of common total enthalpy, the total enthalpy remains constant throughout the complete flow field (see J.D. Anderson [40], Chapter 6.4). Other problems may lead one to choose other quantities (velocity, temperature, pressure).

In order to compute the boundary flux,  $\Phi(\mathbf{V}_i, \mathbf{V}_\infty)$ , a very natural approach consists of Steger and Warming’s flux splitting formula (see Chap. V, Sect. 4.3.5), slightly modified in the sense that the matrix is computed at the same state,  $\mathbf{V}_i$ , i.e.,

$$\Phi(\mathbf{V}_i, \mathbf{V}_\infty) = \mathbf{A}_n^+(\mathbf{V}_i)\mathbf{V}_i + \mathbf{A}_n^-(\mathbf{V}_\infty)\mathbf{V}_\infty$$

is replaced by

$$\Phi(\mathbf{V}_i, \mathbf{V}_\infty) = \mathbf{A}_n^+(\mathbf{V}_i)\mathbf{V}_i + \mathbf{A}_n^-(\mathbf{V}_i)\mathbf{V}_\infty,$$

which uses known information since, roughly speaking, negative eigenvalues correspond to prescribed boundary data. Note that for a supersonic inflow,  $\mathbf{A}_n^+ = 0$  and thus one “forces” in a weak sense  $\mathbf{V}$  to take the value  $\mathbf{V}_\infty$  (only the flux is prescribed, not the state, but the four dependent variables are involved), whereas for a supersonic outflow,  $\mathbf{A}_n^- = 0$ ,

$$\Phi(\mathbf{V}_i, \mathbf{V}_\infty) = \mathbf{A}_n^+(\mathbf{V}_i)\mathbf{V}_i, \quad (5.2)$$

which corresponds to an extrapolation.

This linearized approach is valid for “weak” nonlinear effects, not for “strong” nonlinear ones. Assume, for instance, that in the last case of a supersonic outflow ( $a_4(\mathbf{V}_\infty, -\mathbf{n}) = u_{-n\infty} + c_\infty = -u_{n\infty} + c_\infty \leq 0$ ), the internal state is subsonic ( $a_4(\mathbf{V}_i, -\mathbf{n}) = u_{-ni} + c_i \geq 0$ ,  $a_{2,3}(\mathbf{V}_i, -\mathbf{n}) \leq 0$ ) and thus does not correspond to a supersonic outflow. Then, one expects the extrapolation (5.2) to be unstable, and instead one introduces an intermediate “boundary” state. This state is obtained by again solving a Riemann

problem following the approach of Osher and Chakravarthy (1983) (see [919] and Sect. 2.2 above). It consists in solving a Riemann problem between an unknown left state  $\mathbf{V}_0$  and the right state  $\mathbf{V}_i$ , so that we obtain a state on the boundary  $\mathbf{V}_\infty = \lim \mathbf{w}_R(\xi; \mathbf{V}_0, \mathbf{V}_i)$  as  $\xi \rightarrow 0$ . It can also be implemented by introducing a fictitious cell on the other side of the boundary, where the state is  $\mathbf{V}_0$ .

More precisely, according to the (linearized) boundary conditions, the “state”  $\mathbf{V}_\infty$  is only characterized by the fact that it belongs to some manifold  $\mathcal{V}$  with codimension  $q = \text{number of specified conditions} = \text{number of positive eigenvalues}$ . One solves the “partial Riemann problem” between this manifold  $\mathcal{V}$  (or rather the adherence  $\bar{\mathcal{V}}$ ) and  $\mathbf{V}_i$ , i.e., one looks for a state  $\mathbf{V}_0$  in  $\bar{\mathcal{V}}$  and  $q-1$  intermediate states such that  $\mathbf{V}_0$  can be connected to  $\mathbf{V}_i$  by a succession of at most  $q$  waves of the families corresponding to positive eigenvalues; then one sets

$$\Phi(\mathbf{V}_i, \mathbf{V}_\infty) \equiv \Phi(\mathbf{V}_i, \mathbf{V}_0).$$

On the right-hand side, the flux may be that of Godunov, Osher, and Mehlmann (Chap. IV, Sect. 3.4). In the case of Godunov’s flux, it means that we solve exactly a one-dimensional Riemann problem where the right state  $\mathbf{V}_i$  is given and the left state  $\mathbf{V}_0$  is unknown but belongs to  $\bar{\mathcal{V}}$ , and the resulting flux through  $x = 0$  provides the boundary flux.

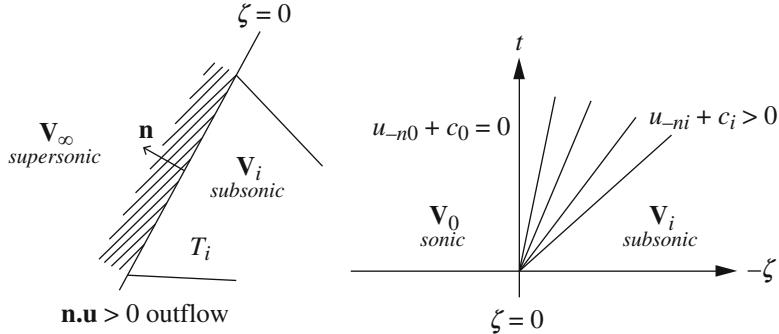
For outflow and weak nonlinearities ( $\mathbf{V}_i$  is “near”  $\mathcal{V}$ ), the waves solution of the Riemann problem are all in the same quadrant  $\frac{x}{t} > 0$ , so that there is only one state  $\mathbf{V}_0$  in the quadrant  $\frac{x}{t} \leq 0$ , and one can speak of a boundary state  $\mathbf{V}_0$ .

In the above example (supersonic outflow with subsonic internal state Fig. 5.2), one looks for one ( $q = 0$ ) supersonic (or sonic) state  $\mathbf{V}_0$  that can be connected to  $\mathbf{V}_i$  by a 4-wave. Since  $a_4(\mathbf{V}_0, -\mathbf{n}) = -u_{n0} + c_0 \leq 0 \leq a_4(\mathbf{V}_i, -\mathbf{n}) = -u_{ni} + c_i$ , we find that this wave is a 4-rarefaction (see Chap. III, Sect. 3, Lemma 3.1), and for  $\mathbf{V}_0$ , we get the unique sonic state connected to  $\mathbf{V}_i$  by a 4-rarefaction ( $\mathbf{V}_0$  is obtained by writing that the state is sonic  $-u_{n0} + c_0 = 0$  and that the 4-Riemann invariants  $\mathbf{u} \cdot \mathbf{n} - \ell, \mathbf{u} \cdot \mathbf{n}^\perp$ , and  $s$  are constant across the 4-wave). Then, if Godunov’s flux is used,

$$\Phi(\mathbf{V}_i, \mathbf{V}_\infty) \equiv \mathbb{F} \cdot \mathbf{n}(\mathbf{V}_0).$$

For a subsonic outflow  $a_{2,3}(\mathbf{V}_\infty, -\mathbf{n}) \leq 0 \leq a_4(\mathbf{V}_\infty, -\mathbf{n})$ , the manifold is  $\mathcal{V} = \{\mathbf{U} = (\rho, \mathbf{u}, p) / p = p_\infty\}$ . In the weak nonlinear case,  $a_4(\mathbf{V}_i, -\mathbf{n}) \geq 0$ , and we look for a state  $\mathbf{V}_0$  with given pressure connected to  $\mathbf{V}_i$  by a 4-rarefaction; again, the three 4-Riemann invariants are constant across the 4-simple wave, so that  $\mathbf{V}_0$  is completely determined. Otherwise,  $\mathbf{V}_i$  corresponds to a supersonic outflow  $a_4(\mathbf{V}_i, -\mathbf{n}) \leq 0$ , and the 4-wave can be a shock. The state  $\mathbf{V}_0$  is obtained by taking in the  $(u, p)$ -plane the intersection of the 4-shock curve through  $\mathbf{V}_i$  with the horizontal line  $\{p = p_\infty\}$ .

For details concerning the other cases, and the use of Osher’s scheme, see the work of F. Dubois [435] and references therein.



**Fig. 5.2** Supersonic outflow with subsonic internal state

## Notes

We have already mentioned in each section of this chapter the contributions to which we refer (in particular, Higdon (1986) [615], Yee (1981) [1204], Gustafsson et al. (1972) [574], Engquist and Majda (1977) [460], Hedstrom (1979) [597], Thompson (1987) [1121], Bardos, LeRoux and Nedelec (1979) [86] and [284], [725] for extensions of their result, [769], Dubois and LeFloch (1988) [437] [435], Gisclon [517], Gisclon and Serre (1994) [518, 519], Cambier et al. (1984, 1985) [227] [1168], to which we can add a nice (unpublished) report by Wu [765, 1192, 1193] for the application; see also [600], the book of D. Serre [1038, 1041] for theoretical results, and the papers of Sablé-Tougeron (1993) [997, 998], Asakura (1996) [61]; Jameson [648], Karni [678], [901] for the numerical treatment and [772] for intermediate boundary conditions for time-split methods. The reader will find in the above papers many other important references that we did not quote. There are also more details concerning many parts of this chapter in the book of Hirsch [617, 618]. Boundary conditions for reactive flows can be found in [93] and for magnetohydrodynamic flows in [1077], both following Thompson's approach, for shallow water equations [568].

### Note Added in the Second Edition

We did not consider other type of boundary conditions than Dirichlet type, where the state is prescribed (in a weak sense), though other types of boundary conditions, such as zero flux boundary condition, have received interest since the first edition; see [218], the work of Andreianov and coauthors [44, 46–48], and references therein; see also [911] for the kinetic and hydrodynamic limit point of view. Concerning linearization, see [223]; for the discretization, see [367, 369] and references therein (besides [366] for more theoretical results), and see [793, 1100] for the numerical treatment of boundary conditions for high-order schemes.



# VII

## Source Terms

### 1 Introduction to Source Terms

We now consider general systems of  $p$  balance laws

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{j=1}^d \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{u}) = \mathbf{s}(\mathbf{u}, \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d \quad t > 0,$$

and we will mostly focus on systems in one dimension which are written in the form

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x} \mathbf{f}(\mathbf{u}) = \mathbf{s}(\mathbf{u}, x), \quad x \in \mathbb{R} \quad t > 0, \quad (1.1)$$

where  $\mathbf{u} \in \Omega$ , subset of  $\mathbb{R}^p$ , and the source  $\mathbf{s}$  in the right hand side of (1.1) is some  $\mathcal{C}^1$  given function from  $\Omega \times \mathbb{R}$  into  $\mathbb{R}^p$ . Note that  $\mathbf{s}(\mathbf{u}, x)$  may depend on  $x$  but does not contain derivative terms; we do not consider diffusion, for instance, though diffusion terms may be considered later on in the study of relaxation systems, but we will not call them *source terms*. For the Cauchy problem, we will moreover be given the initial condition  $\mathbf{u}(., 0) = \mathbf{u}_0(.)$ .

We have already encountered such systems in the Chap. I, see in that chapter (2.8) or (2.19) and the examples beyond introducing combustion or the gravity force (2.24) (2.26), with  $\mathbf{s} = \mathbf{s}(\mathbf{u})$ , and other systems where the source may depend on the space variable with the example of the shallow water model (3.1); in the present chapter, we will mainly consider this model in the one-dimensional case (3.2) (in the Chap. I), or the example of flow in a duct of variable section (3.3), examples which we will study again below. We have also encountered source terms in Chap. IV, for the kinetic and relaxation approaches introduced to derive the corresponding kinetic and relaxation schemes.

We are interested in giving some insight into the new problems raised by the treatment of source terms and developing some of the ideas which are at the basis of their recent numerical treatment, focusing on some classical examples. We will refer frequently to the existing literature for the statement of results and a more complete description, since it is out of scope to go into detailed proofs for all the examples.

## 1.1 Some General Considerations for Systems with Source Terms

We begin by some simple remarks concerning results which are easy extensions of those obtained for conservation laws (without source term), results which we state without proof, and using the notations introduced in the previous chapters.

First, one can easily prove that, as for conservation laws, discontinuities may occur, even with a regular initial data. If the solution is piecewise  $C^1$  and is discontinuous across a surface in the  $(\mathbf{x}, t)$  space with outward normal  $\mathbf{n}$ , the Rankine-Hugoniot condition remains the same as for the conservation law (if the source term is a function, not a measure),

$$[\mathbf{u}]n_t + \sum_{j=1}^d [\mathbf{f}_j(\mathbf{u})]n_{x_j} = \mathbf{0},$$

and in one dimension

$$[\mathbf{f}(\mathbf{u})] = \sigma[\mathbf{u}], \quad (1.2)$$

where  $\sigma = \sigma(t)$  denotes the speed of propagation of the discontinuity. The characteristic fields, i.e., the eigenvalues and eigenvectors of the Jacobian matrix  $\mathbf{f}'(\mathbf{u})$ , will thus play a major role as for systems of conservation laws.

The entropy condition associated to a convex entropy  $U$  with entropy flux  $F$  must be written, taking the source term into account,

$$\frac{\partial}{\partial t} U(\mathbf{u}) + \sum_{j=1}^d \frac{\partial}{\partial x_j} F(\mathbf{u}) \leq U'(\mathbf{u})\mathbf{s}(\mathbf{u}, \mathbf{x});$$

see the Chap. I, Remark 5.1, with equality for smooth solutions. In one dimension, the entropy jump inequality in case of a discontinuous solution, i.e. (5.15), remains unchanged, which means

$$[F(\mathbf{u})] \leq \sigma[U(\mathbf{u})]. \quad (1.3)$$

In the scalar case, the results of existence and uniqueness of Theorem 5.4 (of the Chap. I) extend to a multidimensional scalar balance law with flux and source depending on the space and even on the time variables as proved initially by Kruzhkov (for recent stability results and total variation estimates, see [343, 751]).

There are however some important new features: in the scalar one-dimensional case, the solution is no longer constant along characteristics which are no longer straight lines when the flux is nonlinear. Note also that the solution of the Riemann problem is no longer self-similar, as we will see in the next section, and constants are no longer trivial solutions (see [619]).

Then one may be interested in the asymptotic behavior of the solution when the source term depends on some parameter which becomes very small or very large (depending on the chosen examples). For example, for relaxation systems which we have already considered and will again consider later on, the source is often written in the form of a BGK like kernel  $\mathbf{s}(\mathbf{u}) = (\mathbf{u}_{eq} - \mathbf{u})/\tau$  (at least for simplicity at this point), where  $\tau$  is the *relaxation time*, and when  $\tau \rightarrow 0$ , the solution  $\mathbf{u}$  (which depends on  $\tau$ ) tends to  $\mathbf{u}_{eq}$ :  $\mathbf{u} \rightarrow \mathbf{u}_{eq}$  at least formally. These states  $\mathbf{u}_{eq}$  are then called *equilibrium* states (or sometimes Maxwellian, in connection with the kinetic approach, as seen in Chap. IV, Sect. 8), and the solution  $\mathbf{u}$  is expected to *relax* toward an *equilibrium* (which supposes some stability process).

More generally, we are often interested in source terms  $\mathbf{s}(\mathbf{u}, \alpha)$  depending on some parameter  $\alpha$  which may become very large and with the asymptotic behavior of the solution when  $\alpha \rightarrow \infty$ . If  $\mathbf{s}(\mathbf{u}, \alpha) = \mathbf{s}(\mathbf{u})\alpha$ , then the roots of  $\mathbf{s}(\mathbf{u}) = \mathbf{0}$  are again equilibrium states. However, the above derivation is only formal, and the fact that the solution does approach such a state as  $\alpha \rightarrow \infty$  depends on stability properties, a subject which will be evoked for relaxation source terms, and may involve some scaling of the time variable (as for the Euler system with friction, see Example 1.6 below).

In some important cases, stationary (or steady state) solutions play an important role. These solutions, which thus satisfy  $\partial_t \mathbf{u} = 0$ , come naturally when considering the large time behavior of time-dependent solutions. Indeed, setting  $t = \alpha s$ , and  $\tilde{\mathbf{u}}(x, s) = \mathbf{u}(x, t)$ , we get that  $\tilde{\mathbf{u}}$  satisfies  $\alpha^{-1} \partial_s \tilde{\mathbf{u}} + \partial_x \mathbf{f}(\tilde{\mathbf{u}}) = \mathbf{s}(\tilde{\mathbf{u}}, x)$ , and when  $\alpha \rightarrow \infty$  we get (at least formally) that  $\tilde{\mathbf{u}}$  tends to a stationary solution.

For such solutions, the source and the flux equilibrate, one also speaks of *equilibria* (for instance, a lake at rest for the shallow water system). In practice, using the same term should not bring confusion in a specific context.

Note that it will be important to design schemes which are able to reproduce these last features. We will speak of *asymptotic preserving* (resp. *well-balanced*) property for schemes which are consistent with the asymptotic behavior (resp. preserve steady equilibria). For the AP property moreover, the stability condition should not depend on the parameter  $\alpha$ .

## 1.2 Simple Examples of Source Terms in the Scalar Case

Let us illustrate the direct influence of source terms by exhibiting explicit solutions for some simple examples.

*Example 1.1. Linear advection-reaction equation.* Let us consider a scalar linear advection-reaction equation

$$\partial_t u + a \partial_x u = -\alpha u \quad (1.4)$$

with given initial data  $u(x, 0) = u_0(x)$  in  $\mathbf{L}^\infty(\mathbb{R})$ . In the right-hand side,  $\alpha$  is a given constant parameter, which is a “damping” coefficient if  $\alpha > 0$  (in the more complex model of Example 1.6 below, it is linked to friction). One can think of the evolution of the mass fraction for a simplified model of reactive flow (see Example 2.6 in the Chap. I, assuming constant density, velocity and temperature, and Arrhenius reaction rate).

We can easily compute the solution using characteristics. Along a line  $\frac{dx}{dt} = a$ , the solution  $\tilde{u}(t) = u(x(t), t)$  satisfies  $\frac{d\tilde{u}}{dt} = -\alpha\tilde{u}$ . Thus, the solution is no longer constant on a characteristic line, and

$$u(x, t) = e^{-\alpha t} u_0(x - at).$$

When  $\alpha > 0$ , if  $u_0$  is bounded,  $u(x, t)$  tends to the trivial equilibrium state  $u = 0$  as  $t \rightarrow \infty$ .

A steady state or stationary solution of (1.4) satisfies  $a\partial_x u = -\alpha u$ ; thus,  $u(x) = \kappa e^{-\alpha x/a}$ , for some constant  $\kappa$ . If the initial data is given by  $u_0(x) \equiv \kappa e^{-\alpha x/a}$  then  $u(x, t) = u_0(x)$ ,  $\forall t > 0$ , this is indeed a steady state solution. However, if  $\kappa \neq 0$ , this situation supposes the problem is posed on a half line, say for  $x > 0$  if  $a > 0$  (or on a bounded interval), otherwise  $u$  is not bounded, and then a boundary condition is needed. For instance, if the boundary condition  $u(0, t)$  is a constant, given by some  $\kappa$ , then the solution of (1.4) will indeed reach a stationary state, since for  $t > \frac{x}{a}$ ,  $u(x, t) = \kappa e^{-\alpha x/a}$ , whatever  $u_0$ .

Note that when the coefficient  $\alpha$  increases, the asymptotic behavior is also easily characterized and  $u \rightarrow 0$ .  $\square$

Let us now consider more generally a nonlinear balance law

$$\partial_t u + \partial_x f(u) = s(u), \quad (1.5)$$

with smooth source  $s$  (in the sense that  $s$  is a bounded function, not a measure). The Rankine-Hugoniot condition satisfied by a discontinuous solution is the same as for the homogeneous equation, see (1.2), which writes

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

for a discontinuity propagating with speed  $\sigma$ , and the entropy condition for a strictly convex flux  $f$  is again  $u_-(t) \geq u_+(t)$ . We illustrate the solution of the Riemann problem on the simple example of Burgers' equation.

*Example 1.2. Burgers' equation with damping.* Consider Burgers' equation with damping

$$\partial_t u + \partial_x \frac{u^2}{2} = -\alpha u. \quad (1.6)$$

In a domain where  $u$  is  $\mathcal{C}^1$ , along a characteristic curve  $\frac{dx}{dt}x(t) = u(x(t), t)$ , the solution  $\tilde{u}(t) = u(x(t), t)$  satisfies  $\frac{d\tilde{u}}{dt} = -\alpha\tilde{u}$ . If  $\alpha \neq 0$ , the solution is no longer constant along the characteristics, and these are no longer straight

lines. If the domain reaches the line  $t = 0$ , and  $x(0) = x_0$ , then  $\tilde{u}(t) = u_0(x_0)e^{-\alpha t}$ , and from  $\frac{d}{dt}x(t) = u_0(x_0)e^{-\alpha t}$ , we get

$$x(t) = \frac{u_0(x_0)}{\alpha}(1 - e^{-\alpha t}) + x_0.$$

Let us assume that  $\alpha > 0$  and start from a smooth decreasing initial data  $u_0(x)$ . We can see that characteristics may intersect in finite time by taking two characteristics, starting from  $x_0, x_1$  with  $x_0 < x_1$  if  $|u'_0|$  on the interval  $(x_0, x_1)$  is large enough wrt.  $\alpha$ .

Consider now the Riemann problem with data  $(u_L, u_R)$ . Introducing the characteristics, we get that when  $u_L > u_R$  the entropy solution is discontinuous and given, for  $t > 0$ , by

$$u(x, t) = \begin{cases} u_L e^{-\alpha t}, & x < \sigma(t) \\ u_R e^{-\alpha t}, & x > \sigma(t), \end{cases}$$

where, for Burgers' equation, the discontinuity propagates with speed  $\sigma(t) = \frac{1}{2}(u_+(t) + u_-(t))$ . Thus  $\sigma(t) = \frac{1}{2}e^{-\alpha t}(u_L + u_R)$ , and the trajectory of the shock (discontinuity curve) is  $\Sigma = \{(x, t); x(t) = \frac{u_L + u_R}{2\alpha}(1 - e^{-\alpha t})\}$ . When  $u_L < u_R$ , the solution for  $t > 0$  is a “rarefaction,” i.e., a piecewise  $C^1$  continuous function

$$u(x, t) = \begin{cases} u_L e^{-\alpha t}, & x < x_L(t), \\ \frac{\alpha x}{1 - e^{-\alpha t}} e^{-\alpha t}, & x_L(t) \leq x \leq x_R(t) \\ u_R e^{-\alpha t}, & x > x_R(t) \end{cases}$$

where  $x_{L,R}(t) = u_{L,R} \frac{1 - e^{-\alpha t}}{\alpha}$  are characteristic curves. Indeed, first, we know that  $e^{\alpha t}\tilde{u}$  is constant on a characteristic. Then let us assume that the characteristic curves in the domain  $\{(x, t), x_L(t) \leq x \leq x_R(t), t > 0\}$  can be parametrized by  $\xi \in [u_L, u_R]$ , so that  $x = x(t; \xi) = \xi \frac{1 - e^{-\alpha t}}{\alpha}$ ; this idea comes from the fact that a characteristic curve satisfies for any  $x_0 \neq 0, t > 0$ ,  $x(t) - x_0 = u_0(x_0) \frac{1 - e^{-\alpha t}}{\alpha}$ . Then we can write  $\xi = \frac{\alpha x}{1 - e^{-\alpha t}}$ , and the right-hand side is constant on a characteristic so that we may look for a solution  $u$  which, for  $t > 0$ , has the form  $u(x, t) = e^{-\alpha t}w(\xi)$ , with  $\xi = \frac{\alpha x}{1 - e^{-\alpha t}}$ . We derive the ODE for  $w$  from the balance law satisfied by  $u$ ; we get that  $w$  satisfies  $w'(\xi)(-\xi + w(\xi)) = 0$ . This yields in turn the above formula for  $u$ .

We check that when  $\alpha \rightarrow 0$ , we recover the usual structure of the Riemann solution (see [777] section 17.15 for further developments).

Again, when the “friction” coefficient  $\alpha$  increases, the asymptotic behavior is damped. Note that the smooth stationary solutions of (1.6) are linear in  $x$  (which supposes that the domain is bounded and thus boundary conditions are needed).  $\square$

*Example 1.3. Advection with geometric source term.* We now consider

$$\partial_t u + a \partial_x u = s(x), \quad a > 0 \tag{1.7}$$

with a source term depending on  $x$  (which is usually called a geometric source term). Assume  $a \neq 0$ , the characteristics defined by  $\frac{dx}{dt}(t) = a$  are straight lines:  $x - x_0 = at$ , the solution restricted to a characteristic  $\tilde{u}(t) = u(x(t), t)$  satisfies  $\frac{d\tilde{u}}{dt}(t) = s(x(t))$ , thus we get

$$u(x, t) = u_0(x - at) + \frac{1}{a} \int_{x-at}^x s(y) dy.$$

An equilibrium satisfies  $a \frac{du}{dx} = s(x)$ , which can be easily solved by introducing a primitive  $S$  of  $s$ , then  $au - S = \text{const.}$   $\square$

These simple examples already reveal different new features which we will again encounter on more realistic models: constant states are no longer solution of the balance law; according to the nature of the source, there can exist some interesting steady states, or the behavior wrt. the source may be interesting to study.

### 1.3 Numerical Treatment of Source Terms

From the numerical point of view, there are several ways to treat the source term in an equation of the form (1.6)

$$\partial_t u + \partial_x f(u) = s(u).$$

We begin by the simplest idea, following the finite volume approach. A cell is an interval  $C_j = (x_{j-1/2}, x_{j+1/2})$ , with length  $\Delta x$  which we assume is uniform for simplicity, as well as the time step  $\Delta t$ , and we set  $\lambda = \frac{\Delta t}{\Delta x}$ . Then  $u_j^n$  will denote an approximation at time  $t_n = n\Delta t$  of the average value  $\frac{1}{\Delta x} \int_{C_j} u(x, t) dx$ .

#### 1.3.1 Direct Finite Volume Treatment of the Equation

One may use a direct treatment in a finite volume-type approach which means that one integrates the equation on a cell  $C_j$  and treats both sides in a “natural” way: integrating the left-hand side leads to writing a usual scheme with numerical flux, and for the source, using some quadrature rule, one gets for  $n \geq 0, j \in \mathbb{Z}$ ,

$$u_j^{n+1} = u_j^n - \lambda \left( g_{j+1/2}^n - g_{j-1/2}^n \right) + \Delta t s(u_j^n), \quad (1.8)$$

with a well-chosen numerical flux  $g_{j+1/2}^n = g(u_j^n, u_{j+1}^n)$  (if we take a 3-point scheme), and the initial data is given as usually by  $u_j^0 = \frac{1}{\Delta x} \int_{C_j} u_0(x) dx$ , for  $j \in \mathbb{Z}$ . To obtain (1.8), we have approximated the integral of the source term on a cell  $C_j$  by the midpoint quadrature rule; and if  $s$  depends also on  $x$ , we obtain the discretized source term  $s(u_j^n, x_j)$ . It is only an example; one may use any other quadrature rule, such as the trapezoidal rule, and one may also introduce some upwinding as we will see below.

If we assume that the interface flux  $g_{j+1/2}^n = g(u_j^n, u_{j+1}^n)$  is consistent with  $f$  (in the sense  $g(u, u) = f(u)$ ), it does not take into account the interaction between the transport part  $\partial_t u + \partial_x f(u)$  and the source. Indeed the consistency of the numerical flux  $g$  with the flux  $f$  expresses the property of a conservation law without source that constant states are solution. This is no longer true for a balance law as we have already observed, and we will see situations where the definition of the numerical flux must involve the stationary solutions.

If consistency is rather easily treated, it is not obvious to prove stability results for general equations. It can be done in the linear case (1.4), for the upwind scheme (1.8), for example, assuming a CFL condition  $|a|\Delta t \leq \Delta x$ . The scheme writes for  $a > 0$

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

Then, under CFL 1,  $0 \leq \lambda a \leq 1$ , we obtain

$$\sum_j |u_j^{n+1}| \leq (1 + |\alpha| \Delta t) \sum_j |u_j^n|,$$

so that we can control the growth in the  $\mathbf{L}^1$  norm, and, if  $n\Delta t \leq T$ ,

$$\sum_j |u_j^n| \leq e^{|\alpha|T} \sum_j |u_j^0|.$$

Besides if  $\alpha > 0$  and moreover  $\alpha \Delta t \leq 1 - \lambda a$  (which means the friction coefficient is not too large if one does not want to reduce the time step), both the  $\mathbf{L}^1$  and the  $\mathbf{L}^\infty$  norms decrease.

### 1.3.2 Splitting or Fractional Step Methods

We have already encountered the operator splitting method in Chap. IV, for the kinetic (resp. relaxation) schemes in Sect. 7.3.1 (resp. Sect. 8.1.2) (however, recall that the relaxation step was instantaneous) and also in Chap. V, Sect. 3.2, with dimensional splitting, for a problem in two space dimensions, which supposes a Cartesian mesh.

There may be several possibilities in the decomposition of a general balance law. Presently, the idea of using a splitting method comes naturally since

one may be interested in taking advantage of all the known results concerning schemes for “homogenous” conservation laws (i.e., without source term). Thus, one splits the solution operator in two successive operators: one for the advection part, and in a first step, and one solves the homogeneous system

$$\partial_t u + \partial_x f(u) = 0, \quad (1.9)$$

with some adequate finite volume scheme

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

(to simplify, we have taken again a 3-point scheme), and in a second step, one takes the source term into account, and one solves the ODE

$$\partial_t u = s(u), \quad (1.10)$$

on a time interval of length  $\Delta t$ , taking as initial data the solution  $u_j^{n+1-}$  obtained at the end of the previous step, with some adequate ODE solver. If one chooses the implicit Euler method in this last step, it gives

$$u_j^{n+1} = u_j^{n+1-} + \Delta t s(u_j^{n+1}),$$

and thus the resulting scheme writes

$$u_j^{n+1} = u_j^n - \lambda (g(u_i^n, u_{j+1}^n) - g(u_{j-1}^n, u_j^n)) + \Delta t s(u_j^{n+1}). \quad (1.11)$$

Again, one may use any other ODE solver, and in some situations, one can even solve the ODE step exactly; (1.11) is taken as an example, and it can be easily generalized. For more complex systems in CFD, see [798]; then, for instance, for reaction terms, if the ODE solver is used on each grid cell, one needs to use a second-order scheme to maintain accuracy. This may be achieved on the simple example by  $\frac{1}{2}\Delta t(s(u_j^{n+1}) + s(u_j^{n+1-}))$  or by using a Runge-Kutta method [624].

This fractional step approach is very well suited in some cases and is indeed followed in the derivation of relaxation schemes.

It is well known that the resulting scheme is first order in time. Indeed, formally, one solves a PDE of the form  $\partial_t u = \mathcal{A}u + \mathcal{S}u$ , where  $\mathcal{A}, \mathcal{S}$  are operators in some well-chosen functional spaces, by solving successively  $\partial_t u = \mathcal{A}u$ ,  $\partial_t u = \mathcal{S}u$ , each on a time interval of length  $\Delta t$ . As already said in Chap. V, Sect. 3.2, the justification relies on the Trotter formula (see Chap. V, Remark 3.4) which may be written with the present notations  $e^{t(\mathcal{A}+\mathcal{S})} = \lim_{n \rightarrow \infty} (e^{\Delta t \mathcal{S}} e^{\Delta t \mathcal{A}})^n$ , where  $t = n\Delta t$  (called the Lie formula in the case of matrices; see [306, 337, 683]). Formally again, using a Taylor expansion for each solution operator, for instance  $e^{\Delta t \mathcal{A}} = I + \Delta t \mathcal{A} + \frac{\Delta t^2}{2} \mathcal{A}^2 + \mathcal{O}(\Delta t^3)$ , we write

$$e^{\Delta t(\mathcal{A}+\mathcal{S})} = I + \Delta t(\mathcal{A} + \mathcal{S}) + \frac{\Delta t^2}{2}(\mathcal{A}^2 + \mathcal{A}\mathcal{S} + \mathcal{S}\mathcal{A} + \mathcal{S}^2) + \mathcal{O}(\Delta t^3),$$

while

$$\begin{aligned} e^{\Delta t\mathcal{S}}e^{\Delta t\mathcal{A}} &= (I + \Delta t\mathcal{S} + \frac{\Delta t^2}{2}\mathcal{S}^2)(I + \Delta t\mathcal{A} + \frac{\Delta t^2}{2}\mathcal{A}^2) + \mathcal{O}(\Delta t^3) \\ &= I + \Delta t(\mathcal{A} + \mathcal{S}) + \frac{\Delta t^2}{2}(\mathcal{A}^2 + 2\mathcal{S}\mathcal{A} + \mathcal{S}^2) + \mathcal{O}(\Delta t^3) = e^{\Delta t(\mathcal{A}+\mathcal{S})} + \mathcal{O}(\Delta t^2), \end{aligned}$$

which gives a first-order error for the fractional step method if the operators respectively associated to the advection step, say  $\mathcal{A}$  (resp. to the ODE step  $\mathcal{S}$ ), do not commute. One can improve the order by using the classical Strang splitting which gives a second-order method in time and corresponds to

$$\begin{aligned} e^{\Delta t\mathcal{S}/2}e^{\Delta t\mathcal{A}}e^{\Delta t\mathcal{S}/2} &= I + \Delta t(\mathcal{A} + \mathcal{S}) + \frac{\Delta t^2}{2}(\mathcal{S}^2 + \mathcal{S}\mathcal{A} + \mathcal{A}\mathcal{S} + \mathcal{A}^2) + \mathcal{O}(\Delta t^3) \\ &= e^{\Delta t(\mathcal{A}+\mathcal{S})} + \mathcal{O}(\Delta t^3). \end{aligned}$$

*Example 1.4.* (*Example 1.1 revisited*). It is important to notice that for the simple linear advection reaction equation (1.4), the solution after the two steps of a splitting method solved exactly coincides with the exact solution; hence, there is no error due to the splitting part of the scheme. Indeed, one can commute the two steps, since the operators  $\mathcal{A}$  and  $\mathcal{S}$  commute.  $\mathcal{A}$  is the operator  $v \mapsto -a\partial_x v$ , and then the solution operators also commute  $e^{t\mathcal{A}} : v(.) \mapsto v(. - at)$  (resp.  $\mathcal{S}$  is the operator  $v \mapsto -\alpha v$  and  $e^{t\mathcal{S}} : v \mapsto v e^{-\alpha t}$ ).

Now, let us look at the different methods for (1.4), in the case  $a > 0$ . The upwind method in (1.8) leads to the formula

$$u_j^{n+1} = u_j^n - \lambda a(u_j^n - u_{j-1}^n) - \Delta t \alpha u_j^n,$$

while the fractional step method leads to define first

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

then, if we use the explicit Euler method, we get

$$u_j^{n+1} = u_j^{n+1-} - \Delta t \alpha u_j^{n+1-} = u_j^n - \lambda a(u_j^n - u_{j-1}^n) - \Delta t \alpha u_j^n + \alpha \lambda a \Delta t (u_j^n - u_{j-1}^n),$$

while if we use instead the implicit Euler method, we get

$$u_j^{n+1} = \frac{1}{1 + \alpha \Delta t} (u_j^n - \lambda a(u_j^n - u_{j-1}^n)).$$

It is easy to check that the three methods are first-order accurate; however, their stability properties differ. For the implicit one, under CFL  $\lambda|a| \leq 1$ ,

$$|u_j^{n+1}| \leq |u_j^{n+1-}| \leq \max(|u_j^n|, |u_{j\pm 1}^n|)$$

whatever  $\alpha > 0$ . For the explicit one, we need to assume moreover  $\alpha\Delta t \leq 2$ ; for the direct approach, a sufficient stability condition is  $\alpha\Delta t \leq \frac{1}{2}$  under CFL 1/2. Clearly, in case of a stiff source term, i.e., for large  $\alpha$ , a bound on  $\alpha\Delta t$  limits the time step and may not be acceptable for large  $\alpha$ .  $\square$

We will not study here other possible difficulties linked to a stiff source term. An interesting example is treated in [786] with a parameter-dependent source term  $s(u) = -\alpha u(u-1)(u-1/2)$  presenting three equilibria, one of which ( $u = 1/2$ ) is unstable, and the other two ( $u = 0$  and  $u = 1$ ) are stable; it serves as a model problem for understanding the behavior of numerical methods on reacting flow problems (see [467] for the convergence of the solution as  $\alpha \rightarrow \infty$ ). Numerical difficulties appear for large ratio of the advection time scale to the “relaxation” scale  $\Delta t\alpha$ , due to a lack of spatial resolution in evaluating the source terms, leading to wrong speed of propagation of discontinuities (Chapter 17 in LeVeque’s textbook [777] presents a complete discussion).

*Remark 1.1.* In the scalar case, one can prove a convergence result for the semi-discrete scheme, using the exact solution operator for the solution of the homogeneous system (1.9), say  $H(t) = e^{At}$  and the explicit Euler solver for (1.10) say  $E(t) : v \mapsto v + ts(v)$ , so that the semi-discrete scheme writes, starting from an initial data  $v_0$ ,

$$v(., t_n) = (H(\Delta t)E(\Delta t))^n v_0(.) .$$

Let  $u(x, t)$  be the entropy solution of (1.6), with initial data  $u_0$ , and  $v(x, t_n)$  the operator splitting solution; let  $T_0 > 0$  be some given fixed time. Assume the initial condition  $u_0$  has finite total variation  $TV(u_0)$ . There exists a positive constant  $K_0$  depending on  $TV(u_0), TV(v_0), f, s, T_0$  such that for  $T = N\Delta t \leq T_0$ ,

$$\|u(., T) - v(., T)\|_{\mathbf{L}^1(\mathbb{R})} \leq K_0(\|u_0 - v_0\|_{\mathbf{L}^1(\mathbb{R})} + \Delta t) .$$

The proof can be found in [733]. Then [1102] analyzes the order of convergence. We refer to [624] for a thorough treatment. Other references are [116, 1025].  $\square$

If the splitting method works well on many examples, there may however be some problems in particular situations, for instance, in the design of intermediate boundary conditions in case of an I.B.V.P.; also if one wants to preserve equilibria at a discrete level and for stiff source terms (for an analysis in the case of a stiff source term, see [934, 935] or [221]). Indeed, it is easy to see on some examples that a splitting method in general does not preserve equilibria, in particular for a geometric source term. Starting from an equilibrium state, it is not preserved because the first step does not take into account the source and thus a priori gives birth to nonstationary waves.

If the idea of upwinding is introduced, one can derive schemes which, for this aim, have a better behavior. We just illustrate the idea; it will be more systematically developed in a following section.

*Example 1.5. (Example 1.3 revisited).* Consider again Eq. (1.7)

$$\partial_t u + a \partial_x u = s(x), \quad a > 0.$$

Unlike what occurs for Eq. (1.4), the two exact solution operators associated to (1.7) do not commute so that the solution after the two steps does not coincide with the exact solution. Indeed the first step is the same; however, for the ODE second step,  $\partial_t u = s(x)$ , the operator which we noted above  $e^{tS}$  is given by  $v \mapsto v + s(x)t$ , and it yields the formula for  $e^{\Delta t S} e^{\Delta t A}$ :  $v(x, t + \Delta t) = v(x - a\Delta t, t) + \Delta t s(x)$ , while the exact solution satisfies  $u(x, t + \Delta t) = u(x - a\Delta t, t) + \frac{1}{a} \int_{x-a\Delta t}^x s(y) dy$ .

Let us discretize directly this equation satisfied by the exact solution on a regular grid

$$u(x_j, t_{n+1}) = u(x_j - \nu \Delta x, t_n) + \frac{1}{a} \int_{x_j - \nu \Delta x}^{x_j} s(y) dy$$

where we have set  $\nu = a \frac{\Delta t}{\Delta x}$ ,  $\nu$  is the Courant number, which we assume here is not an integer (for a 3-point scheme, a CFL condition will limit  $\nu \leq 1$ ). The usual upwind method interpolates  $u(x_j - \nu \Delta x, t_n)$  by  $u_j^n - \nu(u_j^n - u_{j-1}^n) = (1 - \nu)u_j^n + \nu u_{j-1}^n$ . A quadrature rule for the source term gives

$$\frac{1}{a} \int_{x_j - \nu \Delta x}^{x_j} s(y) dy \sim \Delta t (\theta s(x_{j-1}) + (1 - \theta)s(x_j))$$

$\theta \in [0, 1]$ , which is of order 2 for  $\theta = 1/2$ ; in this case, we get finally, setting  $s_j = s(x_j)$ ,

$$u_j^{n+1} = u_j^n - \nu(u_j^n - u_{j-1}^n) + \frac{1}{2} \Delta t (s_j + s_{j-1}). \quad (1.12)$$

The formula shows an upwinding of the source term which comes directly from the nature of the equation.

On this example, in (1.12), the discrete solution reaches a stationary state, say  $\bar{u}_j$  iff  $a \frac{\bar{u}_j - \bar{u}_{j-1}}{\Delta x} = \frac{1}{2}(s_j + s_{j-1})$ , which is indeed a two-point second-order finite difference discretization of an equilibrium  $a \frac{d}{dx} u = s$ .

The other way round to see the property is that the scheme gives a steady state solution  $u_j^{n+1} = u_j^n, \forall j \in \mathbb{Z}$ , iff  $a \frac{u_j^n - u_{j-1}^n}{\Delta x} = \frac{1}{2}(s_j + s_{j-1})$ . Thus, starting from an equilibrium  $u_0(x)$  satisfying  $au'_0(x) = s(x)$  and discretizing the terms at second order of accuracy by the expression  $a(u_j^0 - u_{j-1}^0) = \frac{1}{2} \Delta x (s_j + s_{j-1})$ , we have  $u_j^1 = u_j^0$ ; this equilibrium is preserved. One says the scheme is *well-balanced* wrt. these discrete equilibria.

Indeed, in some situations, physical equilibria are known (lake at rest for shallow water, atmospheric column at rest); in these examples which we will

see below, the source term involves a gravity force), and one may want to preserve them at a discrete level, because it is important for any simulation that no spurious wave is created numerically when an equilibrium state is involved.

Now, a splitting method with the upwind scheme in a first step gives (assuming  $a > 0$ )

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

and after the second step with Euler explicit scheme, we get

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

If the initial data is an equilibrium state, discretized as previously by

$$a \frac{u_j^0 - u_{j-1}^0}{\Delta x} = \frac{1}{2}(s_j + s_{j-1})$$

(discretization of second order), we do not have this discrete steady state exactly preserved, since at the first step,

$$u_j^{1-} - u_j^0 = \lambda a(u_j^0 - u_{j-1}^0) = \frac{1}{2}\Delta t(s_j + s_{j-1}),$$

and after the second step  $u_j^1 = u_j^{1-} + \Delta t s_j$ , we get

$$u_j^1 - u_j^0 = \frac{1}{2}\Delta t(3s_j + s_{j-1}).$$

Both methods are first-order accurate, but the first one (1.12) has a better accuracy on discrete equilibria; this method can be easily generalized for a more general monotone flux  $\partial_t u + \partial_x f(u) = s(x)$ ; if  $f' > 0$ , it reads

$$u_j^{n+1} = u_j^n - \lambda(f(u_j^n) - f(u_{j-1}^n)) + \frac{1}{2}\Delta t(s_j + s_{j-1}).$$

The above scheme preserves exactly discrete equilibria in the sense that: if the initial data  $(u_j^0)_{j \in \mathbb{Z}}$  satisfies  $\forall j \in \mathbb{Z}, f(u_j^0) - f(u_{j-1}^0) = \frac{1}{2}\Delta x(s_j + s_{j-1})$ , then  $u_j^1 = u_j^0$ , and consequently  $u_j^n = u_j^0, \forall n \geq 0$ . We refer to P.L. Roe (1986) in [232]) where this idea of upwinding difference schemes was introduced, see also [119, 1166], and we will come again on this method in a following section.

□

*Remark 1.2.* Note that we have discretized the equilibrium in a finite difference way, not following a finite volume approach. If  $u_0$  is a continuous equilibrium, then  $u_0(x) = \frac{1}{a}S(x)$ , where  $S$  is a primitive of  $s$ , and if we define  $v_j^0$  as the exact average  $\frac{1}{\Delta x} \int_{C_j} u_0(x) dx$ , then  $v_j^0$  does not satisfy in general exactly the relation  $v_j^0 - v_{j-1}^0 = \frac{1}{2a}\Delta x(s_j + s_{j-1})$ . Indeed,

one checks easily that  $a(v_j^0 - v_{j-1}^0) = \Delta x s(x_{j-1/2}) + \Delta x^3 s''(\xi)$ , for some  $\xi \in (x_{j-3/2}, x_{j+1/2})$ , then  $a(v_j^0 - v_{j-1}^0) = \frac{1}{2} \Delta x (s_j + s_{j-1}) + \mathcal{O}(\Delta x^3)$ , whether  $s_j = s(x_j)$  or  $s_j = \frac{1}{\Delta x} \int_{C_j} s(x) dx$ , and unless  $s$  is linear  $s(x) = \alpha x + \beta$ , the  $\mathcal{O}(\Delta x^3)$  term remains. Then it proves that the discrete equilibria relation is satisfied at the order 2 for a general (smooth enough) function  $s$ , i.e.,  $\frac{a}{\Delta x} (v_j^0 - v_{j-1}^0) = \frac{1}{2} (s_j + s_{j-1}) + \mathcal{O}(\Delta x^2)$ .

This is to emphasize that the exact/approximate *well-balanced* property (exact/approximate preservation of discrete equilibria) needs to precise which equilibria are concerned, how they are discretized, and, if they are not exactly preserved, at what order of accuracy the scheme preserves them (for a first-order scheme, it should be at least second order).  $\square$

For some source terms, a splitting strategy is quite natural (see [330], for instance). However, there is no general assessment concerning the choice of one method rather than the other; it mostly depends on the application in view, as will be illustrated in the following sections (see [677] and references therein). It may even depend on the wanted accuracy (for a given coarseness of the mesh): for a model of two-phase flow, some elements of comparison can be found in [499], and in some contexts, for reacting gas, for instance, the time step method can be modified to take into account the different time scales of the transport part and the reactive part; see [603] and references therein. For convergence and error bounds, see [1103, 1104].

## 1.4 Examples of Systems with Source Terms

Let us now write again some of the examples of systems with source term which were considered in the Chap. I, since they illustrate different kinds of source terms. The source term may be due to external forces, such as gravity, or to some phenomenological model to describe friction. In this case, a Lagrangian description is still possible. We may also take into account some other phenomena such as chemical process (in case of reacting flow): since we already detailed the Chapman-Jouguet theory, we will not come over it again.

We will first consider *geometrical terms* which means that the source depends on the space variable, which does not allow a whole Lagrangian description. This dependence may come from an averaging procedure together with some symmetry assumption (as for flow in a duct of variable section) and may correspond to some physical boundary (as for the shallow water equations or for a nozzle).

We will not consider multiphase flow for which transfer source terms are naturally involved, taking into account interfacial mass transfer if the change of phase is considered, drag effects, or interfacial heat transfer [346, 1073]. Let us also mention relaxation-type source terms, introducing some difference

terms (such as relative velocity or pressure) taking into account the fact that the two phases are not at equilibrium [486, 487, 730, 835, 882]. Relaxation source terms are often treated numerically by a splitting method, with possible modification to take into account stiff cases [114, 221]. We just mention this important domain of application but will not go into details because of the complexity and the great variety of models of two-phase flow.

As already seen on the scalar examples, the numerical treatment of the source terms may raise new aspects or difficulties which were not present in the study of the classical systems of conservation laws.

We focus on the three following examples, already presented in the Chap. I.

*Example 1.6. The system of gas dynamics with gravity and friction.* This example (see Chap. I, Example 2.5) illustrates the case where the source term is due to external forces, namely, gravity, or to friction. Assuming slab symmetry, the system writes

$$\left\{ \begin{array}{l} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0, \\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = \rho(g - \alpha\varphi(u)), \\ \frac{\partial}{\partial t}(\rho e) + \frac{\partial}{\partial x}((\rho e + p)u) = \rho(gu - \alpha\psi(u)) \end{array} \right. \quad (1.13)$$

where  $g$  stands for the gravity force and  $\alpha > 0$  is a friction constant coefficient. Whereas the gravitational force is given by a fundamental principle, the friction term needs some modeling; we will give below some classical expressions.

Passing in Lagrangian coordinates, (1.13) becomes

$$\left\{ \begin{array}{l} \frac{\partial \tau}{\partial t} - \frac{\partial u}{\partial m} = 0, \\ \frac{\partial u}{\partial t} + \frac{\partial u}{\partial m} = g - \alpha\varphi(u), \\ \frac{\partial e}{\partial t} + \frac{\partial}{\partial m}(pu) = gu - \alpha\psi(u). \end{array} \right. \quad (1.14)$$

This example is interesting to study for its equilibria. When only gravity is considered, it is a model for a column of air, and in some applications [237, 771], it is important to design schemes that are able to preserve a column at rest, which for (1.13) corresponds to

$$u = 0, \quad \partial_x p = \rho g.$$

System (1.13) is also interesting for its asymptotic behavior as the friction coefficient becomes large (see [1191] and references therein).

*Remark 1.3.* The law of balance of momentum is obtained from Newton's second law: *force = mass × acceleration*. Till now, the only force considered was that exerted by means of stresses (for an ideal fluid, the motion is associated to the pressure  $p$ , the force of stress exerted per unit area of a surface  $\Sigma$  is  $p\mathbf{n}$ ,  $\mathbf{n}$  unit normal to  $\Sigma$ ,  $p$  pressure). The calculation of the friction terms  $(\varphi, \psi)$  from first principles is impossible and necessitates the use of empirical methods. In system (1.13) is added a given body force, here the gravity force which has only one component along the vertical  $z$  axis; note that we consider a 1d flow and use classically  $x$  as space variable instead of  $z$ .  $\square$

The functions  $\varphi(u)$  and  $\psi(u)$  satisfy  $\varphi(0) = \psi(0) = 0$ ,  $\psi'(0) = 0$ ,  $\varphi$  increasing. We will mainly consider the commonly used friction terms

$$\begin{cases} \varphi(u) = |u|^\chi u, \\ \psi(u) = a|u|^{\chi+2}, \quad 0 \leq a \leq 1, \end{cases} \quad (1.15)$$

with  $\chi \geq 0$ ,  $\chi = 0$  for a linear friction or  $\chi = 1$  for a quadratic friction term, and  $a$  is some constant. In what follows, we will often restrict to  $a = 1$ , then  $\psi(u) = \varphi(u)u$ . In the context of shallow water models which follow, this corresponds to the Darcy-Weisbach friction term.  $\square$

*Example 1.7. The Saint-Venant system.* It is a shallow water model, with topography, obtained by averaging the Euler system on a vertical column of water, assuming the fluid is incompressible (see Chap. I, Example 3.1). This averaging procedure avoids the treatment of a free boundary for the surface. The resulting system involves a geometric source term when the bottom topography  $z = Z(x)$  is not flat (not horizontal). Denoting by  $h = h(x, t)$  the water height and  $u$  the horizontal velocity (in fact its average on a column of water of height  $h$ ), the Saint-Venant system reads

$$\begin{cases} \frac{\partial h}{\partial t} + \frac{\partial}{\partial x}(hu) = 0, \\ \frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}(hu^2 + \frac{g}{2}h^2) = -ghZ'(x), \end{cases} \quad (1.16)$$

where  $g$  is the gravitational constant and the function  $Z(x)$  characterizes the topography.

System (1.16) is like the barotropic system of gas dynamics with a source term, where  $\frac{g}{2}h^2$  plays the role of a pressure. Hence, the eigenvalues of the homogeneous part of (1.16) are  $\lambda_{\pm} = u \pm c$ , with  $c = \sqrt{gh}$ , and  $\lambda_- < \lambda_+$  if  $h > 0$ .

As already mentioned in Remark 1.3, if one takes friction into account, one may add a friction term in the momentum balance equation

$$\frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}(hu^2 + \frac{g}{2}h^2) = -ghZ'(x) - C_f u|u|,$$

for some friction coefficient  $C_f$ , and this expression corresponds to the Darcy-Weisbach friction term. There are other possible expressions.

This example has been much studied in the last years because of its importance for the applications, even if more complex models are now considered. It is interesting because of the presence of equilibria and has been a powerful driving force for the derivation and numerical analysis of robust and *well-balanced* schemes. Also, it is important to be able to perform computations in the presence of a dry area, which means even if the water height vanishes (in the case of gas dynamics, it would correspond to void), and this is a real challenge to derive well-balanced schemes having moreover a good behavior in this case and satisfying some stability requirements (such as an energy estimate).

For the derivation of *well-balanced* numerical schemes, recent studies include the source terms in the PDE part. In fact, (1.16) can be written equivalently replacing  $Z'(x)$  by  $\partial_x Z$  and complementing by an equation

$$\partial_t Z = 0.$$

This approach (introduced in [554, 560]; see also [552]) is now classical for a geometric source term (which depends on  $x$  but not on  $t$ ) or some systems of conservation laws with flux depending on  $x$ . We illustrate it again in the following example.  $\square$

*Example 1.8. Flow in a duct of variable section (quasi one-dimensional nozzle flow).* We consider a flow in a duct, with axis parallel to the  $x$  axis, with symmetric cross section of area  $A(x)$ , assuming that the area varies smoothly with space (see Chap. I, Example 3.2, also [718]). Starting from the 3D Euler system and averaging on a section, we get the 1D Euler system with a geometric source term when  $A$  is not constant

$$\begin{cases} \partial_t(A\varrho) + \partial_x(A\varrho u) = 0, \\ \partial_t(A\varrho u) + \partial_x(A(\varrho u^2 + p)) = pA'(x), \\ \partial_t(A\varrho e) + \partial_x(A(\varrho e + p)u) = 0. \end{cases} \quad (1.17)$$

Note that system (1.17) can also be written in the form

$$\begin{cases} \partial_t\varrho + \partial_x(\varrho u) = -\frac{A'(x)}{A(x)}\varrho u, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + p) = -\frac{A'(x)}{A(x)}\varrho u^2, \\ \partial_t(\varrho e) + \partial_x((\varrho e + p)u) = -\frac{A'(x)}{A(x)}(\varrho e + p)u, \end{cases}$$

with the usual Euler system in the left-hand side and source terms taking into account the geometry. This proves that for smooth  $A$ , system (1.17) is

strongly hyperbolic, and the characteristic fields are the same as those of Euler system, the two fields associated to  $u \pm c$ , where  $c$  is the usual sound speed (which are GNL) and  $u$  (which is LD).

Assume for simplicity that the fluid is barotropic; we get the system

$$\begin{cases} \partial_t(A\varrho) + \partial_x(A\varrho u) = 0, \\ \partial_t(A\varrho u) + \partial_x(A(\varrho u^2 + p(\varrho))) = p(\varrho)A'(x). \end{cases} \quad (1.18)$$

System (1.18) looks like the barotropic Euler system with a source, provided we set  $r \equiv A\varrho$ , and  $P(r) \equiv Ap(\frac{r}{A})$ . However, this definition of  $P$  introduces a flux depending on  $x$  through  $A$ , which does not simplify the approach. For the numerical treatment, as for system (1.16), we may include the source terms in the PDE part. Indeed (1.18) can be written equivalently replacing  $A'(x)$  by  $\partial_x A$  and complementing by an equation

$$\partial_t A = 0.$$

It yields

$$\begin{cases} \partial_t(A\varrho) + \partial_x(A\varrho u) = 0, \\ \partial_t(A\varrho u) + \partial_x(A(\varrho u^2 + p(\varrho))) - p(\varrho)\partial_x A = 0, \\ \partial_t A = 0, \end{cases} \quad (1.19)$$

and we have introduced a nonconservative term, which will need a special treatment. Note that the pressure term, say  $P = Ap$ , can indeed be written as  $P = P(r, A) = Ap(\frac{r}{A})$ , where again  $r = A\varrho$  is the first component of the conservative variable  $(A\varrho, A\varrho u)^T$  of (1.18) and  $A$  is the new added variable. We can also check that  $\partial_r P = p'(\varrho)$ , the eigenvalues of the Jacobian (0 apart), do coincide with those of the Euler system (as expected, since it results from a change of variables; see Chap. II, Sect. 2.1.1).  $\square$

This example is important since it is representative of some models for a fluid flow in a porous medium, where now  $A$  will play the role of porosity [312, 606]. Then, it is also involved in some two-phase flow models, as the Baer-Nunziato model, where it plays the role of a volume fraction or phase fraction, and schemes for (1.19) may thus serve as a building block in the resolution of more complex problems; see [351, 352, 361] and also [1118].

## 2 Systems with Geometric Source Terms

Systems (1.16) and (1.17) or (1.18) are typical examples of systems with a geometric source term which in the general form write

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, K) = K'(x) \mathbf{s}(\mathbf{u}) \quad (2.1)$$

with  $\mathbf{u} \in \Omega \subset \mathbb{R}^p$ ,  $\mathbf{f}$  (resp.  $\mathbf{s}$ ) smooth functions from  $\tilde{\Omega} = \Omega \times \mathbb{R} \rightarrow \mathbb{R}^p$  (resp.  $\Omega \rightarrow \mathbb{R}^p$ ) and  $k \equiv K'$  a smooth real valued function. In fact, for flow in a nozzle in variable  $(r, ru)$ , for system (1.18), we have a function  $\mathbf{s}(\mathbf{u}, K)$ , since  $p(\varrho) = p(A\varrho/A)$ , but we will often restrict to  $\mathbf{s}(\mathbf{u})$  for simplicity. The source term can be transformed into a partial differential one; this leads to write a larger nonconservative system in variable  $\mathbf{v} = (\mathbf{u}, K)$  (the analog of (1.19))

$$\begin{cases} \partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, K) - \mathbf{s}(\mathbf{u}) \partial_x K = 0, \\ \partial_t K = 0. \end{cases} \quad (2.2)$$

Because we are interested in solutions of (2.1), the notion of entropy is relative to  $\mathbf{u}$ , but the entropy pair is also a function of  $K$ , and an entropy condition for (2.2) writes

$$\partial_t U(\mathbf{v}) + \partial_x F(\mathbf{v}) - S(\mathbf{v}) \partial_x K \leq 0,$$

with  $\partial_{\mathbf{u}} F = \partial_{\mathbf{u}} U \partial_{\mathbf{u}} f$  and  $S(\mathbf{v}) = \partial_{\mathbf{u}} U(\mathbf{s} - \partial_K \mathbf{f}) + \partial_K F$ . The function  $U(\mathbf{v})$  is required to be convex in  $\mathbf{u}$ , but because of the linear degeneracy of the field associated to  $K$ , the convexity wrt.  $K$  is not necessary; at least for smooth  $K$ , one may speak of partial entropy. If moreover the pair  $(U, F)$  satisfies  $\partial_K F = \partial_{\mathbf{u}} U(\partial_K \mathbf{f} - \mathbf{s})$ , the entropy relation writes for smooth solutions

$$\partial_t U(\mathbf{v}) + \partial_x F(\mathbf{v}) = 0;$$

it is an entropy conservation law. It will be the case for the applications we are interested in; there exists such an entropy pair  $(U, F)$  (see, for instance, (2.9), (2.18)).

If  $k = K'$  is a smooth function, the product  $k(x)\mathbf{s}(\mathbf{u}) = \mathbf{s}(\mathbf{u})\partial_x K$  is well defined, even if  $\mathbf{u}$  is discontinuous. However, for numerical purposes, we are interested in solving the Riemann problem for (2.2), and we may assume that it has self-similar solutions. However, it yields that  $K$  is discretized by a piecewise constant function; then, a problem occurs with the definition of the nonconservative product  $\mathbf{s}(\mathbf{u})\partial_x K$  when  $K$  is discontinuous, which means on stationary waves. Let us first consider the problem of nonconservative products from a general point of view.

## 2.1 Nonconservative Systems

For a general nonconservative system of  $p$  equations

$$\partial_t \mathbf{v} + \mathbf{A}(\mathbf{v}) \partial_x \mathbf{v} = 0, \quad (2.3)$$

where  $\mathbf{A}$  is a smooth  $p \times p$  matrix valued function defined on some set  $\Omega \subset \mathbb{R}^p$ , one can compute the characteristic fields which are naturally the eigenvalues and corresponding eigenvectors of the  $p \times p$  matrix  $\mathbf{A}(\mathbf{v})$ . We assume

the system is strictly hyperbolic, which means that the matrix  $\mathbf{A}(\mathbf{v})$  is diagonalizable on  $\mathbb{R}$ . Then the notions of GNL (genuinely nonlinear) field or LD (linearly degenerate) field and Riemann invariants (see Chap. II, Definitions 2.1 and 3.2) which were defined for a conservative system extend easily since they only involve the characteristic fields.

### 2.1.1 Nonconservative Systems Coming from Systems with a Geometric Source Term

For system (2.2), we have a result linking the characteristic fields of (2.2) and those of (2.1). As previously, we set  $\mathbf{v} = (\mathbf{u}, K)$ , and the dependence on  $K$  is sometimes skipped in the notations:  $\mathbf{f}'(\mathbf{v})$  stands for the  $p \times p$  matrix  $D_{\mathbf{u}}\mathbf{f}(\mathbf{u}, K)$ ; we assume that for any  $K$ ,  $\mathbf{f}'(\mathbf{v}) \equiv D_{\mathbf{u}}\mathbf{f}(\mathbf{u}, K)$  has  $p$  real distinct eigenvalues and thus a basis of eigenvectors.

*Lemma 2.1*

Let  $\lambda_j(\mathbf{v}), \mathbf{r}_j(\mathbf{v}), j = 1, \dots, p$  be the characteristic fields of  $\mathbf{f}'(\mathbf{v})$  in (2.1), and assume the fields are either GNL or LD.

If  $\forall j \in \{1, \dots, p\}, \forall \mathbf{v}, \lambda_j(\mathbf{v}) \neq 0$ , the system (2.2) is strictly hyperbolic, and the characteristic fields of system (2.2) are:

–  $\lambda_j(\mathbf{v}), \tilde{\mathbf{r}}_j(\mathbf{v})$  with  $\tilde{\mathbf{r}}_j(\mathbf{v}) = (\mathbf{r}_j(\mathbf{v}), 0)$ ,  $1 \leq j \leq p$ , and these fields have the same nature, GNL or LD, as the corresponding fields of (2.1),

– and  $\lambda_0 = 0, \tilde{\mathbf{r}}_0(\mathbf{v})$ , which is LD, where the eigenvector  $\tilde{\mathbf{r}}_0(\mathbf{v})$  can be chosen in the form  $(\mathbf{f}'^{-1}(\mathbf{v})\mathbf{s}(\mathbf{u}), 1)$ .

If  $w : \tilde{\Omega} \rightarrow \mathbb{R}$  is a  $j$ -Riemann invariant for (2.1), i.e.,  $w$  satisfies  $\forall \mathbf{v} \in \tilde{\Omega}, D_{\mathbf{u}}w(\mathbf{v}) \cdot \mathbf{r}_j(\mathbf{v}) = 0$ , then  $w$  is  $j$ -Riemann invariant for (2.2); moreover  $K$  is also a  $j$ -Riemann invariant  $\forall j \in \{1, \dots, p\}$ .

If  $\exists j_0 \in \{1, \dots, p\}, \exists \bar{\mathbf{v}}, \lambda_{j_0}(\bar{\mathbf{v}}) = 0$ , the system (2.2) has a basis of eigenvectors at  $\bar{\mathbf{v}}$  iff  $\partial_K \mathbf{f}(\bar{\mathbf{v}}) - \mathbf{s}(\bar{\mathbf{u}}) \in \text{Im}(\mathbf{f}'(\bar{\mathbf{v}}))$ .

Again, to simplify the notations, we use a shorthand notation for vectors in  $\mathbb{R}^{p+1}$  whose  $p$  first components are a known vector in  $\mathbb{R}^p$ .

*Proof.* The matrix  $\mathbf{A}(\mathbf{v})$ , with  $\mathbf{v} = (\mathbf{u}, K)$ , is given by

$$\mathbf{A}(\mathbf{v}) = \begin{pmatrix} \mathbf{f}'(\mathbf{v}) & \partial_K \mathbf{f} - \mathbf{s}(\mathbf{u}) \\ \mathbf{0} & 0 \end{pmatrix};$$

hence, the first point concerning the eigenvalues  $\lambda_j, j = 1, \dots, p$  is straightforward. In particular, if  $w : \tilde{\Omega} \rightarrow \mathbb{R}$  satisfies  $\forall \mathbf{v}, D_{\mathbf{u}}w(\mathbf{v}) \cdot \mathbf{r}_j(\mathbf{v}) = 0$ , then  $D_{\mathbf{v}}w(\mathbf{v}) \cdot \tilde{\mathbf{r}}_j(\mathbf{v}) = 0$ . The only difficulty concerns the field associated to the null eigenvalue.

Since  $\tilde{\mathbf{r}}_0(\mathbf{v})$  cannot lie in the space generated by the eigenvectors  $\tilde{\mathbf{r}}_j(\mathbf{v}), j = 1, \dots, p$ , we take it in the form  $\tilde{\mathbf{r}}_0(\mathbf{v}) = (\mathbf{r}_0(\mathbf{v}), 1)$ , then  $\mathbf{r}_0$  satisfies  $\mathbf{f}'(\mathbf{v})\mathbf{r}_0(\mathbf{v}) + \partial_K \mathbf{f} - \mathbf{s}(\mathbf{u}) = \mathbf{0}$ , and if the matrix  $\mathbf{f}'(\mathbf{v})$  is invertible, the result holds.

If  $\exists j_0, \exists \bar{\mathbf{v}}, \lambda_{j_0}(\bar{\mathbf{v}}) = 0$ , the matrix  $\mathbf{f}'(\mathbf{v})$  is not invertible at state  $\bar{\mathbf{v}}$ , but if  $\partial_K \mathbf{f} - \mathbf{s}(\bar{\mathbf{u}}) \in \text{Im}(\mathbf{f}'(\bar{\mathbf{v}}))$ , then  $\exists \mathbf{r}_0(\bar{\mathbf{v}}), \mathbf{f}'(\bar{\mathbf{v}})\mathbf{r}_0(\bar{\mathbf{v}}) + \partial_K \mathbf{f} = \mathbf{s}(\bar{\mathbf{u}})$  and  $\tilde{\mathbf{r}}_0(\bar{\mathbf{v}}) = (\mathbf{r}_0(\bar{\mathbf{v}}), 1)$  is independent from  $\tilde{\mathbf{r}}_{j_0}(\bar{\mathbf{v}})$ . Otherwise, one cannot find a second eigenvector associated to the eigenvalue  $\lambda_{k_0}(\bar{\mathbf{v}}) = \lambda_0 = 0$ . In this last case, one says that the system is *resonant*.

The condition for resonance to appear can be equivalently written

$$\mathbf{l}_{j_0}(\bar{\mathbf{v}})^T (\partial_K \mathbf{f}(\bar{\mathbf{v}}) - \mathbf{s}(\bar{\mathbf{u}})) \neq 0, \quad (2.4)$$

where  $\mathbf{l}_{j_0}$  is the “left” eigenvector of  $\mathbf{f}'$  associated to  $\lambda_{j_0}$ . Indeed, when  $\lambda_{j_0}(\bar{\mathbf{v}})$  vanishes,  $\text{Im}(\mathbf{f}'(\bar{\mathbf{v}}))$  is spanned by the vectors  $\mathbf{r}_j(\bar{\mathbf{v}})$ ,  $j \neq j_0$ , and  $\mathbf{l}_{j_0}(\bar{\mathbf{v}})^T \mathbf{r}_j(\bar{\mathbf{v}}) = 0$  for  $j \neq j_0$ .  $\square$

*Example 2.1.* *Example 1.8 revisited.* For the barotropic system (1.19), the eigenvalues are  $\lambda_{\pm} = u \pm c$ , where  $c = \sqrt{p'(\varrho)}$ . When  $u = \pm c$ ,  $\partial_K \mathbf{f} - \mathbf{s}(\mathbf{u}) = (0, -\varrho p')^T$  does not belong to the image of  $\mathbf{f}'(\mathbf{u})$  which is generated by vector  $(1, 2u)^T$  so that resonance appears, linked to the fact that a GNL field superimposes a LD one. The system (1.19) is strictly hyperbolic in the three subdomains  $u < -c$  where  $\lambda_- < \lambda_+ < \lambda_0$ ,  $|u| < c$  where  $\lambda_- < \lambda_0 < \lambda_+$  and  $u > c$  where  $\lambda_0 < \lambda_- < \lambda_+$ , in the  $(\varrho, u)$ -plane; these domains are limited by the sonic curves  $u = \pm \sqrt{p'(\varrho)}$ . We will come again below on this example and compute the eigenvectors directly.

For the full system with energy (1.17), the eigenvalues are  $\lambda_1 = u - c$ ,  $\lambda_2 = u$ ,  $\lambda_3 = u + c$  ( $c$  sound speed). Again resonance appears at sonic states  $u = \pm c$ . However, when  $\lambda_2 = 0$ , i.e., at a state  $\bar{\mathbf{u}}$  with vanishing velocity  $u = 0$ , a simple computation shows that there is a basis of eigenvectors, either by checking that in that case  $\partial_K \mathbf{f} - \mathbf{s}(\bar{\mathbf{u}}) \in \text{Im}(\mathbf{f}'(\bar{\mathbf{u}}))$  (see the above lemma) or by exhibiting directly two eigenvectors  $\tilde{\mathbf{r}}_2$  and  $\tilde{\mathbf{r}}_0$  which are independent when  $u = 0$ . Note that the fact that  $\lambda_2$  is LD is important.  $\square$

We are interested in computing the solution of the Riemann problem for (2.2), and we may assume that it has the same structure as for a conservative system, a self-similar solution with constant states separated by waves associated to each characteristic field. Lemma 2.1. shows that in this solution,  $K$  may be discontinuous across a  $\lambda_0$  (stationary) wave only. Then a difficulty appears in the definition of the *nonconservative product*  $\mathbf{s}(\mathbf{u})\partial_x K$  when  $\mathbf{u}$  is discontinuous too.

There are some general results concerning the definition of a nonconservative product which we briefly recall.

### 2.1.2 Definition of the Nonconservative Product Based on a Family of Paths

One can define a nonconservative product, based on a family of paths, following the work of Dal Maso-LeFloch-Murat [387]. We have already evoked

the theory in Chap. IV, Sect. 4.4; we just sketch here the great lines of some results and refer to the above cited paper for the complete theory.

Given a nonconservative system (2.3)

$$\partial_t \mathbf{u} + \mathbf{A}(\mathbf{u}) \partial_x \mathbf{u} = 0,$$

for sufficiently smooth  $\mathbf{A}$ , one can define in the general case a nonconservative product, noted  $[\mathbf{A}(\mathbf{u}) \partial_x \mathbf{u}]_\Phi$ , associated to the choice of a family of paths  $\Phi$  in the set of states, for  $\mathbf{u}$  of bounded variation. Restricting first to a step function  $\mathbf{u}(x) = \mathbf{u}_L + H(x)(\mathbf{u}_R - \mathbf{u}_L)$ , where  $\mathbf{u}_{L,R}$  are constant states, the nonconservative product can be defined as a multiple of the Dirac measure  $\delta$  concentrated on 0, with some weight  $C_\Phi$  which is associated to the choice of a family of paths (see [387], page 494 and page 531). Given a family of paths defined on  $[0, 1] \times \mathbb{R}^p \times \mathbb{R}^p$ :  $(s, \mathbf{u}, \mathbf{v}) \mapsto \Phi(s; \mathbf{u}, \mathbf{v})$ , such that  $\Phi(0; \mathbf{u}, \mathbf{v}) = \mathbf{u}$ ,  $\Phi(1; \mathbf{u}, \mathbf{v}) = \mathbf{v}$ ,  $\Phi(s, \mathbf{u}, \mathbf{u}) = \mathbf{u}$ , the definition of the nonconservative product noted  $[\mathbf{A}(\mathbf{u}) \partial_x \mathbf{u}]_\Phi$  is such that the weight  $C_\Phi$  is given by  $C_\Phi = \int_0^1 A(\Phi(s; \mathbf{u}_L, \mathbf{u}_R)) \partial_s \Phi(s; \mathbf{u}_L, \mathbf{u}_R) ds$ .

One can similarly define the nonconservative product for a propagating discontinuity: it yields the following jump relation to define a discontinuity between states  $\mathbf{u}_L$  and  $\mathbf{u}_R$  in  $\mathbb{R}^p$  moving at speed  $\sigma$

$$\sigma(\mathbf{u}_R - \mathbf{u}_L) = \int_0^1 \mathbf{A}(\Phi(s; \mathbf{u}_L, \mathbf{u}_R)) \partial_s \Phi(s; \mathbf{u}_L, \mathbf{u}_R) ds.$$

When  $\Phi$  is the family of straight lines  $\Phi(s; \mathbf{u}, \mathbf{v}) = \mathbf{u} + s(\mathbf{v} - \mathbf{u})$ , the nonconservative product coincides with Volpert's product. When  $A(\mathbf{u}) = \mathbf{f}'(\mathbf{u})$  is a Jacobian matrix, the nonconservative product does not depend on the choice of path and  $\mathbf{A}(\mathbf{u}) \partial_x \mathbf{u} = \partial_x \mathbf{f}(\mathbf{u})$ , (hopefully) consistent with the usual definition.

We refer to [759] for an existence theory to the Riemann problem for (2.3) assuming the system is strictly hyperbolic. We will consider the special case of a LD field in the next section.

*Remark 2.1.* In general, how does one choose the path  $\Phi$  in case of a GNL field? Generally speaking, discontinuous solutions of a physical system depend upon regularization mechanisms and higher-order regularization effects, such as viscosity or capillarity, which must be taken into account in the modeling, leading to systems with higher-order differential terms say  $R_\varepsilon$ , depending on a small parameter  $\varepsilon$  intended to tend to 0; this adds regularity to the system of PDE which then writes  $\partial_t \mathbf{u}_\varepsilon + A(\mathbf{u}_\varepsilon) \partial_x \mathbf{u}_\varepsilon = R_\varepsilon$ , for instance,

$$\partial_t \mathbf{u}_\varepsilon + A(\mathbf{u}_\varepsilon) \partial_x \mathbf{u}_\varepsilon = \partial_x(D(\mathbf{u}_\varepsilon) \partial_x \mathbf{u}_\varepsilon)$$

for some “viscosity” or “diffusion” matrix”  $D$  (see Sect. 5.3.2, Chap. II), which may be constant and even in some cases  $D = I$  (in the vanishing viscosity approach of Bianchini and Bressan [147]), and then  $\mathbf{u}$  is defined as  $\mathbf{u} = \lim_\varepsilon \mathbf{u}_\varepsilon$  (assuming one can prove that the limit exists for some topology). This supposes however that the regularization satisfies some stability property (see

[759], for instance). Shock waves can be determined by traveling wave solutions to the regularized system, that is, Rankine-Hugoniot relations for shock waves are determined from (and in general depend on) the given regularization (see [969, 1001]). More recently [29] introduces the notion of *approximate* shock curve, not depending on a viscosity matrix, and results in [147] are extended to viscosity matrices commuting with  $A$ .

Among such regularizations, one may also mention the Dafermos regularization [380, 805].

Numerical schemes, apart from the Glimm scheme, bring some numerical diffusion too, so that they compute some solution, which in general depends on the chosen scheme. The question is then to determine which solution is computed (see, for instance, [427] for an example where the approximate solution given by different numerical schemes converges toward different “weak solutions”).

Note that one may construct a numerical scheme associated to a family of paths; that does not mean it will converge to the right solution; some ideas may be found in [12, 244], and also, relying on [29], see [263, 264] and references therein; we also mention [241].

Then, for some particular nonconservative systems endowed with “enough” entropy pairs, a possible approach formulates the initial-value problem by supplementing the equations with a *kinetic relation* prescribing the rate of entropy dissipation across shock waves, leading to added conservation laws for smooth solutions [135]. Such systems appear in the modeling of multiphase flows, and a criterion of hyperbolicity can be found in [488]; see also [18].  $\square$

### 2.1.3 A Result Concerning Linearly Degenerate Characteristic Fields

We first encountered a nonconservative equation for a multicomponent flow, see Chap. III, Sect. 3, Remark 3.5, but in that case, the product  $u\partial_x Y$  was well defined because  $Y$  could jump only across a material contact discontinuity with speed  $u$  (which is a LD field) so that  $u$  being continuous, there was no ambiguity, and the equation  $\partial_t(\varrho Y) + \partial_x(\varrho Yu) = 0$  was satisfied by both smooth and discontinuous solutions. This case occurs frequently, and we can formalize some results on supplementary conservation laws which may be satisfied by a discontinuous solution which is a contact discontinuity.

*Definition 2.1*

Assume the  $k$ -th field of the nonconservative system (2.3) is linearly degenerate. A  $k$ -contact discontinuity between  $\mathbf{u}_-$  and  $\mathbf{u}_+$  is a discontinuous function which connects the two states  $\mathbf{u}_-$  and  $\mathbf{u}_+$ , where we assume that both states  $\mathbf{u}_-$  and  $\mathbf{u}_+$  lie on the same integral curve  $\mathcal{C}_k$  of the vector field  $\mathbf{r}_k$ .

Note that this definition implies that the  $k$ -Riemann invariants are equal on  $\mathbf{u}_\pm$ , in particular  $\lambda_k(\mathbf{u}_-) = \lambda_k(\mathbf{u}_+)$ .

Let us assume moreover that the nonconservative system (2.3) is endowed with an entropy pair  $(U, F)$ , which means that

$$\partial_t U(\mathbf{u}) + \partial_x F(\mathbf{u}) = 0, \quad (2.5)$$

for smooth solutions  $\mathbf{u}$ , this supposes in turn the relation  $U' \mathbf{A} = F'$ . In fact, we do not need to assume that  $U$  is convex to prove the following result.

*Proposition 2.1*

Assume that the nonconservative system (2.3) is endowed with a pair of real valued function  $(U, F)$ , defined on  $\Omega$ , such that  $\forall \mathbf{u} \in \Omega, U'(\mathbf{u})\mathbf{A}(\mathbf{u}) = F'(\mathbf{u})$ . Consider a linearly degenerate field  $\lambda_k(\mathbf{u})$  of (2.3), and let  $\mathbf{u}(x, t)$  be a  $k$ -contact discontinuity, connecting the states  $\mathbf{u}_-$  and  $\mathbf{u}_+$  in the sense of Definition 2.1, and set  $\lambda_k = \lambda_k(\mathbf{u}_+) = \lambda_k(\mathbf{u}_-)$ . Then  $\mathbf{u}$  satisfies the identity

$$[F(\mathbf{u})] = \lambda_k[U(\mathbf{u})],$$

which implies that the contact discontinuity  $\mathbf{u}(x, t)$  is a weak solution of the conservation law (2.5).

*Proof.* The proof follows the same lines as the proof given for a conservative system. Consider  $\mathcal{C}_k(\mathbf{u}_-)$  the integral curve of  $\mathbf{r}_k$  through  $\mathbf{u}_-$ , parametrized by some parameter  $\varepsilon \mapsto \Phi_k(\varepsilon)$ , with  $\Phi_k(0) = \mathbf{u}_-$ ,  $\lambda_k$  is constant on  $\mathcal{C}_k(\mathbf{u}_-)$ ,  $\lambda_k = \lambda_k(\Phi_k(\varepsilon)) = \lambda_k(\mathbf{u}_-)$ , and any  $k$ -Riemann invariant is constant on  $\mathcal{C}_k$ . Let us differentiate wrt.  $\varepsilon$  the function

$$\mathcal{E}(\varepsilon) = F(\Phi_k(\varepsilon)) - F(\mathbf{u}_-) - \lambda_k(U(\Phi_k(\varepsilon)) - U(\mathbf{u}_-)).$$

Since  $U' \mathbf{A} = F'$ , we get easily  $\mathcal{E}'(\varepsilon) = 0$ , and since  $\mathcal{E}$  vanishes at  $\Phi_k(0) = \mathbf{u}_-$ ,  $\mathcal{E} = 0$ , so that for any state  $\mathbf{u}_+$  on  $\mathcal{C}_k$ , we have

$$F(\mathbf{u}_+) - F(\mathbf{u}_-) = \lambda_k(U(\mathbf{u}_+) - U(\mathbf{u}_-))$$

which is the desired result.  $\square$

Note again that the result does not need the fact that  $U$  is convex and holds for any additional conservation law satisfied by smooth solutions of (2.3).

In particular, if the nonconservative system (2.2) has additional conservation laws, such as an entropy conservation law for smooth solutions or any other conservation law  $\partial_t U(\mathbf{v}) + \partial_x F(\mathbf{v}) = 0$ , then we may use the results of Proposition 2.1 in case of a  $\lambda_0$ -contact discontinuity: any such  $F$  is a Riemann invariant associated to the  $\lambda_0$ -contact wave. Conversely, if one can find enough pairs  $(U, F)$ , the associated Riemann invariants  $F$  can determine the corresponding wave curve.

In what sense is a contact discontinuity a weak solution of (2.3)? We have to define the nonconservative product, i.e., a family of path  $\Phi$  such that

$$\lambda_k(\mathbf{u}_+ - \mathbf{u}_-) = \int_0^1 \mathbf{A}(\Phi_k(s; \mathbf{u}_-, \mathbf{u}_+)) \partial_s \Phi(s; \mathbf{u}_-, \mathbf{u}_+) ds.$$

*Lemma 2.2*

Let  $(s, \mathbf{u}_-, \mathbf{v}) \mapsto \Phi_k(s; \mathbf{u}, \mathbf{v})$  be a parametrization of the curve  $\mathcal{C}_k(\mathbf{u}_-)$  such that  $\Phi_k(0; \mathbf{u}_-, \mathbf{u}_+) = \mathbf{u}_-$ ,  $\Phi_k(1; \mathbf{u}_-, \mathbf{u}_+) = \mathbf{u}_+$ . The  $k$ -contact discontinuity is a weak solution of (2.3) for the nonconservative product defined by the path  $\Phi_k$ . This will be the case for the nozzle model and the Saint-Venant system.

This result is easily proved since, by definition, the curve  $\mathcal{C}_k$  is an integral curve of the eigenvector  $\mathbf{r}_k$ . It may be useful for a system of the form (2.2) where the contact discontinuity associated to  $\lambda_0$  is the one involved in the nonconservative product, the other waves are associated to a conservative form for which any path gives the same result, and for which any conservative flux-component corresponding to a conservation law (i.e., without source term) provides a  $\lambda_0$ -Riemann invariant.

## 2.2 Stationary Waves and Resonance

We want to define the nonconservative product  $\mathbf{s}(\mathbf{u})\partial_x K$  in (2.2) when  $K$  is discontinuous which means across  $x = 0$ .

If  $\mathbf{u}$ , and thus  $\mathbf{s}(\mathbf{u})$ , is continuous, there is no difficulty to define  $\mathbf{s}(\mathbf{u})\partial_x K$ . So we consider the case when  $\mathbf{u}$  and thus  $\mathbf{v} = (\mathbf{u}, K)$  is discontinuous.

First, if the corresponding nonconservative system (2.2) has a basis of eigenvectors, which means outside resonance (see Lemma 2.1), we may apply the above results and consider the integral curve of the eigenvector  $\tilde{\mathbf{r}}_0$ . Provided the associated nonconservative system (2.3) has an additional conservation law,  $\partial_t U(\mathbf{v}) + \partial_x F(\mathbf{v}) = 0$ , such as an entropy conservation law for smooth solutions, then we may use the results of Proposition 2.1 for a contact discontinuity: any such  $F$  is a Riemann invariant associated to the  $\lambda_0$ -wave. This determines the state  $\mathbf{u}_+$ , and the nonconservative product is given a sense through Lemma 2.2. The Riemann invariants determine the wave curve which allows to define a nonconservative product. The left and right states  $\mathbf{u}_\pm$  which will define the weight of the Dirac mass  $\delta_0$  in the nonconservative product satisfy the equation  $F(\mathbf{u}_+) = F(\mathbf{u}_-)$ , for any pair  $(U, F)$  such that (2.5) holds.

The problem occurs in the case of resonance, when an eigenvalue of the conservative system  $\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, K) = \mathbf{0}$  vanishes, in particular for an eigenvalue  $\lambda_{j_0}$  of  $\partial_u \mathbf{f}(\mathbf{v})$  associated to a GNL field, which can vanish at some sonic state  $\bar{\mathbf{v}}$ . In that case, starting from the assumptions of Lemma 2.1, we know that if (2.4) holds, one cannot find at  $\bar{\mathbf{v}}$  an eigenvector  $\tilde{\mathbf{r}}_0$  independent from the vectors  $\tilde{\mathbf{r}}_j$ . We can precise a little what happens; in particular, the resonant state is not isolated, it lies on a *transition* manifold  $\mathcal{T}$  (for a scalar

problem, for instance, it lies on a curve). If one cannot use the proof of the Lax Theorem (Chap. II, Theorem 6.1), since it needs a basis of eigenvectors, to obtain the existence of a solution of the Riemann problem, one still works with curves in the set of states, the standard wave curves on each side of  $\mathcal{T}$ . The analysis aims at precising the crossing of  $\mathcal{T}$ . The solution is built with standard waves and the standing contact discontinuity; however, composite waves are needed, linked to some *admissible* criteria, introduced for uniqueness. We just give some hints and will not give the whole results, for which we refer to the first papers of [646, 647, 825, 1111], and [536] for the Riemann problem and [34, 1162] for existence and uniqueness or in the contrary ill-posedness results in the scalar case according to the hypothesis on  $K$  and  $\partial_u f$ .

*Lemma 2.3*

Assume  $\exists \bar{\mathbf{v}}$  such that (2.4) holds, where  $\lambda_{j_0}$  is a genuinely nonlinear (GNL) field. There exists locally around  $\bar{\mathbf{v}}$  a smooth manifold  $\mathcal{T}$  of states satisfying  $\lambda_{j_0}(\mathbf{v}) = 0$ . Moreover, there exists a continuous mapping  $\mathbf{v} \mapsto \tilde{\mathbf{r}}_0(\mathbf{v})$  such that, outside  $\mathcal{T}$ , the vectors  $\tilde{\mathbf{r}}_j(\mathbf{v}), 1 \leq j \leq p$  and  $\tilde{\mathbf{r}}_0(\mathbf{v})$  form a basis of  $\mathbb{R}^{p+1}$  and  $(\frac{1}{\|\tilde{\mathbf{r}}_{j_0}\|} \tilde{\mathbf{r}}_{j_0} - \tilde{\mathbf{r}}_0)(\mathbf{v}) \rightarrow 0$  as  $\mathbf{v} \rightarrow \bar{\mathbf{v}}, \mathbf{v} \notin \mathcal{T}$ .

*Proof.* Let us normalize  $\mathbf{r}_{j_0}$ , as usual for the eigenvector of a GNL field, by  $\forall \mathbf{v}, (D_{\mathbf{u}}\lambda_{j_0} \cdot \mathbf{r}_{j_0})(\mathbf{v}) = 1$ . Then, using the implicit function theorem, we obtain that  $\lambda_{j_0}(\mathbf{v}) = 0$  defines locally around  $\bar{\mathbf{v}}$  a smooth  $p$ -dimensional manifold  $\mathcal{T}$  in  $\mathbb{R}^{p+1}$ , containing  $\bar{\mathbf{v}}$ , which is called the *transition* manifold (for  $p = 1$ , it is called the transition curve). Recall that  $\tilde{\mathbf{r}}_{j_0}(\mathbf{v}) = (\mathbf{r}_{j_0}(\mathbf{v}), 0)$ , and note that the integral curve  $\mathcal{C}_{j_0}$  of  $\tilde{\mathbf{r}}_{j_0}$  passing by  $\bar{\mathbf{v}}$  cuts the tangent surface to  $\mathcal{T}$  transversally, since  $(D_{\mathbf{u}}\lambda_{j_0}, D_K\lambda_{j_0})\tilde{\mathbf{r}}_{j_0} = 1$ . Hence, choosing an orientation of the transition surface, associated to the gradient  $D_{\mathbf{v}}\lambda_{j_0}$ , we can distinguish two sides,  $\mathcal{T}^\pm$ , with  $\lambda_{j_0} < 0$  (resp.  $> 0$ ) in  $\mathcal{T}^-$  (resp.  $\mathcal{T}^+$ ).

Now, away from  $\mathcal{T}$ , we may take for  $\tilde{\mathbf{r}}_0(\mathbf{v})$  a unit vector in the form  $\tilde{\mathbf{r}}_0(\mathbf{v}) = \frac{1}{n}(\mathbf{r}_0(\mathbf{v}), -\lambda_{j_0}(\mathbf{v}))$ . Since the eigenvectors  $\mathbf{r}_j$  span  $\mathbb{R}^p$ ,  $\mathbf{r}_0 \in \mathbb{R}^p$  can be decomposed as  $\mathbf{r}_0(\mathbf{v}) = \sum_{j=1}^p \alpha_j \mathbf{r}_j(\mathbf{v})$ , and since  $\tilde{\mathbf{r}}_0$  should be such that the identity  $\mathbf{f}'(\mathbf{v})\mathbf{r}_0(\mathbf{v}) - \lambda_{j_0}(\partial_K \mathbf{f} - \mathbf{s}(\mathbf{u})) = \mathbf{0}$  holds, we get  $\sum_{j=1}^p \alpha_j \lambda_j \mathbf{r}_j(\mathbf{v}) - \lambda_{j_0}(\partial_K \mathbf{f} - \mathbf{s}(\mathbf{u})) = \mathbf{0}$ . The coefficient  $\alpha_j$  is determined by taking the scalar product with the “left” eigenvector  $\mathbf{l}_j$  of  $\mathbf{f}'(\mathbf{v})$ . The eigenvalues are distinct; hence, in a neighborhood of  $\bar{\mathbf{v}}$ ,  $\lambda_j(\mathbf{v}) \neq 0$  for  $j \neq j_0$  and does not change sign. Then, away from  $\mathcal{T}$ ,  $\alpha_j(\mathbf{v}) = \frac{\lambda_{j_0}}{\lambda_j} \mathbf{l}_j^T(\mathbf{v})(\partial_K \mathbf{f} - \mathbf{s}(\mathbf{u}))$ , and  $n$  follows so as to have a unit vector, which completely determines  $\tilde{\mathbf{r}}_0$ . In particular, by assumption (2.4), for  $j \neq j_0$ ,  $\alpha_j(\mathbf{v}) \rightarrow 0$  as  $\mathbf{v} \rightarrow \bar{\mathbf{v}}$  but  $\alpha_{j_0}(\mathbf{v}) \neq 0$  and does not vanish, so that in the limit,  $\tilde{\mathbf{r}}_0$  is parallel to  $\tilde{\mathbf{r}}_{j_0}(\bar{\mathbf{v}})$ .

If  $\partial_K \mathbf{f} - \mathbf{s}(\mathbf{u}) > 0$ , we define  $\tilde{\mathbf{r}}_0$  on  $\mathcal{T}$  by the normalized eigenvector  $\frac{\mathbf{r}_{j_0}}{\|\mathbf{r}_{j_0}\|}$  (otherwise we take the opposite unit vector), and the result holds.  $\square$

Then the integral curve of  $\tilde{\mathbf{r}}_0$ ,  $\mathcal{C}_0(\bar{\mathbf{v}})$  which is parallel to  $\tilde{\mathbf{r}}_{j_0}(\bar{\mathbf{v}})$  at  $\bar{\mathbf{v}}$ , crosses  $\mathcal{T}$  transversally; the last component of the tangent  $\tilde{\mathbf{r}}_0$  changes sign as  $\mathcal{C}_0$

crosses  $\mathcal{T}$ , and thus the curve stays on one side of the hyperplane  $K = \bar{K}$  (the one which contains  $\tilde{\mathbf{r}}_0(\bar{\mathbf{v}})$ ), and  $\bar{K}$  is a local extrema (of the last component along  $\mathcal{C}_0$ ).

If two states lie on  $\mathcal{C}_0$  on the same side, without crossing  $\mathcal{T}$ , they are connected by a standing contact discontinuity as already introduced. If they are not on the same side, allowing a unique standing contact discontinuity would increase the total variation in  $K$ . One rather constructs a composite wave, surimposing a standing  $j_0$ -shock. We do not go further on and refer to [646, 825], and [536] where resonance is studied for system (2.2); see also [498]. We illustrate the above computations in the scalar case.

*Example 2.2. Resonance in the scalar case.* First recall that for a scalar equation, i.e., assuming  $p = 1$  and  $f$  a smooth scalar flux, if the source term  $k$  is smooth, (2.1) has a unique entropy solution, as proved by Kruzhkov.

Now, let us explicit some computations for (2.2) which writes in the scalar case for smooth solutions

$$\begin{cases} \partial_t u + \partial_u f(u, K) \partial_x u + (\partial_K f(u, K) - s(u)) \partial_x K = 0, \\ \partial_t K = 0. \end{cases} \quad (2.6)$$

There are some existence and uniqueness results for the Cauchy problem, away from resonance, assuming  $\partial_u f > 0$  for instance (see [564, 1162]), and existence results for the Riemann problem (see [34, 646, 647], and [536, 629, 677] which also treats resonance). If resonance is not avoided, there is no uniqueness result, unless some interface entropy condition is added (see also [49]); no criterion seems to be proposed to answer the problem of nonuniqueness in the general case (see however Remark 2.2 below).

For ease of notation, we set again  $\mathbf{v} = (u, K)$ , then the field  $\lambda_1(\mathbf{v}) = \partial_u f(\mathbf{v})$ ,  $\tilde{\mathbf{r}}_1 = (1, 0)^T$  is GNL if  $\partial_{u,u} f(\mathbf{v}) \neq 0$ ; and the eigenvector  $\tilde{\mathbf{r}}_0$  is given by  $\tilde{\mathbf{r}}_0(\mathbf{v}) = (\partial_K f - s(u), -\partial_u f(\mathbf{v}))^T$ . Resonance occurs at a state  $\bar{\mathbf{v}} = (\bar{u}, \bar{K})$  if  $\partial_u f(\bar{\mathbf{v}}) = 0$ . Then the eigenvalues are not globally ordered, the speed  $\lambda_1$  may be smaller or greater than  $\lambda_0$ , and the states where they coincide are not isolated.

For a state  $\bar{\mathbf{v}}$  such that  $\partial_u f(\bar{\mathbf{v}}) = 0$ , the hypothesis (2.4) writes:  $s(\bar{u}) - \partial_K f(\bar{\mathbf{v}}) \neq 0$ , for instance,  $(s - \partial_K f)(\bar{\mathbf{v}}) > 0$ ; this assumption yields nondegeneracy of the linearized problem at state  $\bar{\mathbf{v}}$  (see [646, 803])

$$\begin{cases} \partial_t u + (\partial_K f(\bar{\mathbf{v}}) - s(\bar{u})) \partial_x K = 0, & x \in \mathbb{R}, t > 0, \\ \partial_t K = 0, \end{cases}$$

the solution of which, for initial data  $u_0(x), K(x)$ , writes

$$u(x, t) = u_0(x) + (s(\bar{u}) - \partial_K f(\bar{\mathbf{v}})) K'(x) t.$$

Note that if  $K$  is only BV,  $u$  and the partial derivatives in  $x$  explode as  $t \rightarrow \infty$ , and there is no BV estimate on  $u$  either for the linearized equation.

In order to solve the Riemann problem, we introduce, as explained above, some curves in the plane of states  $\mathbf{v} = (u, K)$ , namely, the wave curves and the transition curve. Then, for a given Riemann data,  $\mathbf{v}_L = (u_L, K_L)$  and  $\mathbf{v}_R = (u_R, K_R)$ , we try to connect the two states by some “admissible” path following these curves. We detail the computations.

First, the *transition* curve  $\mathcal{T}$  is defined by  $\{(u, K); \partial_u f(u, K) = 0\}$ . Since  $\partial_{u,u} f(u, K) \neq 0$ , by the implicit function theorem, the transition curve  $\mathcal{T}$  can be written  $u = u_{\mathcal{T}}(K)$  in the neighborhood of  $\bar{\mathbf{v}}$ , with

$$\frac{du_{\mathcal{T}}}{dK} = -\frac{\partial_{u,K} f}{\partial_{u,u} f}.$$

For instance, if  $f(u, K) = f(u)$  does not depend on  $K$ , and  $u = \bar{u}$  is the sonic point, i.e.,  $f'(\bar{u}) = 0$ ,  $\mathcal{T}$  is the straight line  $u = \bar{u}$ .

Outside resonance, in  $\mathcal{T}^-$  (where  $\partial_u f < 0$ ) or in  $\mathcal{T}^+$  ( $\partial_u f > 0$ ), the  $\lambda_1$ -wave curves are the lines  $K = cst$ , say  $K = K_1$ , corresponding to (nonstationary) shocks ( $[K] = 0$ ,  $[f] = \sigma[u]$ ,  $\sigma \neq 0$ ) or rarefactions where  $u$  is determined by the scalar law

$$\partial_t u + \partial_x f(u, K_1) = 0,$$

supplemented by the usual entropy condition.

#### *Lemma 2.4*

*The set of states that can be connected to a state  $\bar{\mathbf{v}}$  by a contact discontinuity associated to  $\lambda_0$  is a curve  $\mathcal{C}_0(\bar{\mathbf{v}})$  in the  $(u, K)$ -plane with equation  $K = K_0(u)$  where  $K_0$  is solution of the ODE*

$$\frac{dK_0}{du} = \frac{\partial_u f}{s - \partial_K f}, \quad K_0(\bar{u}) = \bar{K}.$$

Moreover,  $w(\mathbf{v}) = K_0(u) - K$  is a  $\lambda_0$ -Riemann invariant.

*Proof.* According to Definition 2.1, a contact discontinuity curve associated to the characteristic field  $\lambda_0$  is given by an integral curve of  $\tilde{\mathbf{r}}_0(\mathbf{v}) = (\partial_K f - s(u), -\partial_u f(\mathbf{v}))^T$ . Equivalently, since  $\tilde{\mathbf{r}}_0$  satisfies by definition  $(\partial_u f, \partial_K f - s)^T \tilde{\mathbf{r}}_0 = 0$ , we can find a contact discontinuity as the limit of smooth stationary waves, i.e., smooth solutions of

$$\frac{d}{dx} f(u, K) = s K'(x).$$

Indeed, we can write the above equation in the form of an autonomous system, defining in the  $(u, K)$ -plane the curve  $\mathcal{C}_0 = \{(u, K), K = K_0(u)\}$ , through  $(\bar{u}, \bar{K})$ , from

$$\frac{d}{dx} f(u(x), K(u(x))) = (s \frac{d}{dx} K)(u(x)).$$

We get

$$(\partial_u f + \partial_K f \frac{dK}{du}) \frac{du}{dx} = s \frac{dK}{du} \frac{du}{dx}$$

and

$$\frac{dK}{du} = \frac{\partial_u f}{s - \partial_K f}.$$

Assuming (2.4), in a neighborhood of a state  $\bar{\mathbf{v}} = (\bar{u}, \bar{K})$ , the above ODE defines a curve  $K = K_0(u)$  through  $\bar{\mathbf{v}}$ . It is easy to check that the function  $w(\mathbf{v}) = K_0(u) - K$  is such that  $(\partial_u w, \partial_K w) \cdot \tilde{\mathbf{r}}_0 = 0$  and is thus a  $\lambda_0$ -Riemann invariant. Again if  $f(u, K) = f(u)$  does not depend on  $K$ , noting now  $\varphi$  the function  $K_0$ , i.e., a primitive of  $\frac{f'(u)}{s(u)}$ , we find that  $w = \varphi(u) - K$  is a Riemann invariant (see Example 2.3 and (2.7) below).  $\square$

If the state  $\bar{\mathbf{v}}$  lies on  $\mathcal{T}$ , the tangent to  $\mathcal{C}_0$  at  $\bar{\mathbf{v}}$  is horizontal (and  $\frac{dK}{du} = 0$  iff  $\partial_u f = 0$ ). Thus the curve  $\mathcal{C}_0$  crosses  $\mathcal{T}$  in a transverse way (recall that the transition curve  $u = u_{\mathcal{T}}(K)$  cannot have an horizontal tangent in plane  $(u, K)$ ). Assuming, for instance,  $s - \partial_K f < 0$ , we get  $\frac{d^2 K}{du^2}(\bar{\mathbf{v}}) < 0$  if  $\partial_{u,u} f > 0$ , the curve lies under the line  $K = \bar{K}$ .

In fact, in order to bound the total variation on  $K$ , only stationary waves which do not cross  $\mathcal{T}$  will be considered.

And following Isaacson and Temple (the definition is given in Section 3 in [646]), a 0-wave curve connecting the states  $\mathbf{v}_0$  and  $\mathbf{v}_1$  is called *admissible* if it does not cross the transition curve  $\mathcal{T}$  between  $\mathbf{v}_0$  and  $\mathbf{v}_1$ .

Let us give some hints on the solution of the Riemann problem for (2.6). There is existence (we refer to [536, 646]); however, resonance brings nonuniqueness, and we will not give the full result. We first consider some easy cases.

First,  $K$  can only change across  $x = 0$ , and if  $K_L = K_R \equiv K_{LR}$ , and  $\mathbf{v}_L, \mathbf{v}_R$  lie on the same side of  $\mathcal{T}$ , in  $\mathcal{T}^+$  or in  $\mathcal{T}^-$ , one solves  $\partial_t u + \partial_x f(u, K_{LR}) = 0$ . Assuming  $\partial_{u,u} f > 0$ , it gives a rarefaction if  $u_L < u_R$  and a shock if  $u_L > u_R$ . The solution of the Riemann problem is then a unique  $\lambda_1$ -wave, connecting  $\mathbf{v}_L = (u_L, K_{LR})$  and  $\mathbf{v}_R = (u_R, K_{LR})$ .

If  $K_L \neq K_R$ , there is necessarily a stationary discontinuity involved. Again outside  $\mathcal{T}$ , in  $\mathcal{T}^+$  or in  $\mathcal{T}^-$ , which means that  $\partial_u f'(\mathbf{v}_L)$  and  $\partial_u f'(\mathbf{v}_R)$  have the same sign, the solution is found by linking  $\mathbf{v}_L$  and  $\mathbf{v}_R$  by a  $\lambda_0$ - and a  $\lambda_1$ -wave, the order depends on the sign of  $\partial_u f'(\mathbf{v}_{L/R})$ . Assume, for instance, it is negative. Then a  $\lambda_1$ -wave connects  $\mathbf{v}_L$  with an intermediate state  $\mathbf{v}^* = (u^*, K_L)$ , and a stationary wave connects  $\mathbf{v}^* = (u^*, K_L)$  and  $\mathbf{v}_R$ . The state  $\mathbf{v}^*$  is found by intersecting in the  $(u, K)$ -plane the curve  $\mathcal{C}_0(\mathbf{v}_R)$  and the line  $K = K_L$ ; the (negative)  $\lambda_1$ -wave is a shock if  $u_L > u^*$ , a rarefaction otherwise, similarly if both signs are positive.

As already written, we restrict to *admissible* stationary waves which do not cross  $\mathcal{T}$ . Then one has to consider stationary waves, also associated to  $\lambda_1$ , which allow the crossing of  $\mathcal{T}$ . We introduce a new curve, a zero shock curve corresponding to a standing wave. Given a state  $\mathbf{v}$  on the discontinuity curve  $\{\mathbf{v} = (u, K); K = K_0(u)\}$  in  $\mathcal{T}^+$ , we associate the set of states  $(\tilde{u}, K)$

in  $\mathcal{T}^-$  with  $f(\tilde{u}, K) = f(u, K)$ ,  $K = K_0(u)$ ; note that  $\tilde{u}$  exists ( $\tilde{u} \in \mathcal{T}^-$ , i.e.,  $u \leq u_{\mathcal{T}}(K)$ ), thanks to the assumption  $\partial_{u,u}f > 0$ . Again if  $f(u, K) = f(u)$  does not depend on  $K$ ,  $\tilde{u}$  satisfies  $f(\tilde{u}) = f(u)$ .

For example if  $\mathbf{v}_L = (u_L, K_L) \in \mathcal{T}^+$  is given, and  $\mathbf{v}^* = (u^*, K_0(u^*)) \in \mathcal{T}^+$  lies on the (admissible) discontinuity curve through  $\mathbf{v}_L$ , and  $\mathbf{v}_R = (\tilde{u}^*, K_R = K_0(u^*))$ , one can connect  $\mathbf{v}_L$  and  $\mathbf{v}_R$  by a contact discontinuity between  $\mathbf{v}_L$  and  $\mathbf{v}^*$  superimposed with a stationary shock between  $\mathbf{v}^*$  and  $\mathbf{v}_R$ ; the two waves coincide in the  $(x, t)$  plane but not in the plane  $(u, K)$ .

In fact uniqueness is still not ensured, and for some data, up to three configurations for the solution of the Riemann problem can be constructed. A criterion is introduced in [647], corresponding to the minimization of some functional (and the convergence of the Godunov scheme toward the admissible solution is proved).  $\square$

*Remark 2.2.* Solving the Riemann problem for (2.6) is obviously connected with solving the Riemann problem for a conservation law with a discontinuous flux function  $f(u, x)$  with

$$f(u, x) = \begin{cases} f_l, & x < 0, \\ f_r, & x > 0. \end{cases}$$

Indeed, given a Riemann data  $U_L = (u_L, K_L)$  and  $U_R = (u_R, K_R)$ , one solves

$$\begin{cases} \partial_t u + \partial_x f(u, K_L) = 0, & x < 0, \\ \partial_t u + \partial_x f(u, K_R) = 0, & x > 0, \end{cases}$$

with

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

Moreover, we impose as interface condition the continuity of the  $\lambda_0$ -Riemann invariant which writes  $K_0(u(0-, t)) - K_L = K_0(u(0+, t)) - K_R$ . However, if  $\partial_u f'$  changes sign between  $U_L$  and  $U_R$ , the two states lie on either side of the curve  $\mathcal{T}$ ; it is easy to prove nonuniqueness if one does not add an admissibility criterion. There is an extensive literature on conservation laws with discontinuous flux (let us mention one [498] where the link with resonance is explicit). After [928] the work by Andreianov, Karlsen, and Risebro [45] provides a general framework for the study of *admissible* solutions to conservation laws with discontinuous flux, unifying perspective on many of the known entropy conditions. The whole admissibility issue is reduced to the selection of a family of “elementary solutions” which are admissible solutions to the problem with Riemann data; a complete theory for  $\mathbf{L}^1$  contractive semigroups is developed.

Let us also mention that the problem for a scalar equation in two dimensions, writing  $\partial_t u + \partial_x f(u, k) + \partial_y g(u, l) = 0$ , together with  $\partial_t k = 0$  and  $\partial_t l = 0$  is studied in [676].  $\square$

*Example 2.3. A simplified scalar case.* Consider with a little more details the scalar case (2.6) assuming moreover  $\partial_K f = 0$ , i.e.,  $f = f(u)$ , the equation writes

$$\partial_t u + \partial_x f(u) = k(x)s(u).$$

This case has been much studied; the function noted above  $K_0(u)$ , primitive of  $\frac{f'}{s}(u)$ , which is often noted  $\varphi$

$$\varphi(u) = \int^u \frac{f'}{s}(v)dv, \quad (2.7)$$

is introduced in many works: [158, 552, 677], and references therein. The transition curve  $\mathcal{T}$ , set of states such that  $\lambda_1 = \lambda_0$ , is the straight line  $u = \bar{u}$  (where  $\bar{u}$ , such that  $f'(\bar{u}) = 0$ , is the sonic point). Since  $w = \varphi - K$  is a  $\lambda_0$ -Riemann invariant, the interface condition writes

$$\varphi(u(0-, t) - K_-) = \varphi(u(0+, t) - K_+).$$

If  $\varphi$  is invertible, for instance, if it is strictly increasing, which holds if we assume  $\frac{f'}{s}(v) > 0$ , the above relation defines  $u(0\pm, t)$  as a function of  $u(0\mp, t)$  and  $\Delta K = K_+ - K_-$ . We will come back on this example when designing well-balanced schemes in a following section.  $\square$

### 2.3 Case of a Nozzle with Discontinuous Section

We come back to Example 1.8 and give some hints on the solution of the Riemann problem for the nonconservative system (1.19)

$$\begin{cases} \partial_t(A\varrho) + \partial_x(A\varrho u) = 0, \\ \partial_t(A\varrho u) + \partial_x(A(\varrho u^2 + p(\varrho))) - p(\varrho)\partial_x A = 0, \\ \partial_t A = 0. \end{cases}$$

We will not go into the details of the solution, because of the resonance phenomenon which brings nonuniqueness, unless some *monotonicity criterion* is added [714], and the results are too complex (we refer to [51, 578, 757]).

First, in order to compute directly the characteristic fields of (1.19), not using the set of variables  $(\varrho, \varrho u)$  (as done after (1.17), relying on the results known for the barotropic Euler system), nor the results of Lemma 2.1, we rather introduce another set of variables, say  $\mathbf{v}$ , for which the system can be put in another conservative form, in order to follow the approach of the previous section, and use the results of Lemma 2.2. Indeed, if  $\varrho, u, A$  are smooth, we have a conservative form equivalent for smooth solutions. Writing

$\partial_x(Ap) - p\partial_x A = A\partial_x p$ , we obtain from the momentum equation, combined with the mass equation  $A\varrho\partial_t u + A\varrho u\partial_x u + A\partial_x p = 0$ , that smooth solutions satisfy

$$\partial_t u + u\partial_x u + \frac{p'(\varrho)}{\varrho}\partial_x \varrho = 0,$$

which naturally coincides with the equation on the velocity for the barotropic Euler system. Then, defining  $\pi$  (up to an additive constant) by  $\pi'(\varrho) = p'(\varrho)/\varrho$ , we get

$$\begin{cases} \partial_t(A\varrho) + \partial_x(A\varrho u) = 0, \\ \partial_t u + \partial_x(\frac{u^2}{2} + \pi) = 0, \\ \partial_t A = 0. \end{cases} \quad (2.8)$$

One can check that  $\pi = \varepsilon + p/\varrho$  (it is the enthalpy) where  $\varepsilon$  is the internal energy, defined by  $\varepsilon'(\varrho) = p(\varrho)/\varrho^2$  (see Chap. II, Sect. 7.3.1). For example, if  $p(\varrho) = \kappa\varrho^\gamma$ , with  $\gamma > 1$ , then  $\pi = \kappa\frac{\gamma}{\gamma-1}\varrho^{\gamma-1}$ ,  $\varepsilon = \frac{\gamma p}{\varrho}$ . For an isothermal flow,  $p = c^2\varrho$  and  $\pi = c^2 \log \varrho$ . In “primitive” variables  $\mathbf{w} = (\varrho, u, A)^T$ , we have a nonconservative quasilinear form, equivalent for smooth solutions of (1.19) (or of (2.8)), for which the computations are easy

$$\partial_t \mathbf{w} + \mathbf{B}(\mathbf{w})\partial_x \mathbf{w} = \mathbf{0},$$

with matrix

$$\mathbf{B}(\mathbf{w}) = \begin{pmatrix} u & \varrho & \varrho u/A \\ \pi'(\varrho) & u & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Setting  $c^2 = \varrho\pi'(\varrho) = p'(\varrho)$ ,  $c > 0$ , we check that the eigenvalues are  $\lambda_\pm = u \pm c$  (those of the barotropic Euler system as expected) and  $\lambda_0 = 0$  and the eigenvectors of  $\mathbf{B}(\mathbf{w})$  are  $\mathbf{r}_\pm = (\varrho, \pm c, 0)^T$ ,  $\mathbf{r}_0 = (\varrho u, -c^2, A(c^2 - u^2)/u)^T$ . If  $p(\varrho) = \kappa\varrho^\gamma$ , we get  $c^2 = \gamma p/\varrho = \kappa\gamma\varrho^{\gamma-1}$ . The system is not strictly hyperbolic at a sonic state since the eigenvalues may coincide:

$$\lambda_\pm = \lambda_0 \text{ on } \mathcal{T}^\pm = \{\mathbf{w}; u = \pm c\},$$

and then  $\mathbf{r}_\pm$  and  $\mathbf{r}_0$  are not independent. The  $\lambda_0$ -field is LD, and the other fields are GNL; the resonance phenomenon comes from the superimposition of an acoustic wave (GNL) and the standing LD wave.

Then, we also have an additional entropy conservation law, with the energy playing the role of mathematical entropy.

*Proposition 2.2*

*Smooth solutions of (1.19) satisfy an entropy relation*

$$\partial_t(A\varrho e) + \partial_x(A(\varrho e + p)u) = 0, \quad (2.9)$$

*with  $e = u^2/2 + \varepsilon$  and  $\varepsilon'(\varrho) = p(\varrho)/\varrho^2$ . The function  $(A\varrho, A\varrho u, A) \mapsto A\varrho e$  is convex.*

*Proof.* Note that (2.9) is the analog of the energy for the barotropic Euler system, and this equation is formally the energy equation in the full system (1.17) (as expected). The proof is simply obtained by developing the partial derivatives and combining with the other equations for  $A\varrho$  and  $A\varrho u$ .

Note also that  $\varepsilon$  is a function of the conservative variables  $(A\varrho, A\varrho u, A)$ , since we may write  $\varepsilon(\varrho) = \varepsilon(\frac{A\varrho}{A}) = \tilde{\varepsilon}(r, A)$ ,  $r = A\varrho$ . Then  $\varepsilon'(\varrho) = A\partial_r \tilde{\varepsilon}(A\varrho, A)$  and writing  $p(\varrho)/\varrho^2 = Ap(\varrho)/A\varrho^2 = P(A\varrho, A)/A\varrho^2$  we get  $\partial_r \tilde{\varepsilon} = P(r, A)/r^2$ , while for the barotropic Euler system  $\varepsilon'(\varrho) = p/\varrho^2$ .

In case of discontinuous solutions, the equality (2.9) becomes an inequality ( $\leq$ ); this is formally justified by the vanishing viscosity method.  $\square$

Let us be given a Riemann data  $\mathbf{v}_{L,R} = (A\varrho, A\varrho u, A)_{L,R}^T$ , in particular  $A$  is piecewise constant

$$A = \begin{cases} A_L, & x < 0 \\ A_R, & x > 0. \end{cases}$$

*Remark 2.3.* Note that the assumptions which lead to write the system (1.18) are no longer valid in case of a discontinuous section; nevertheless, we are interested in this study for numerical purposes, when  $A$  is discretized by a piecewise constant function, for instance, if we use the Godunov method. From a physical standpoint, geometrical discontinuities induce multidimensional vortices that modify the surface pressure integral. An improved 1D-averaged flow model is proposed in [863]. An extra energy (or entropy) equation is added to the Euler system, expressing the energy and turbulent pressure stored in the vortices generated by the abrupt area variation. The turbulent energy created by the flow-area change interaction is determined by a specific estimate of the surface pressure integral. The predictions given by this 1D-averaged model are compared with 2D-averaged results from numerical solution of the Euler equations.  $\square$

In order to solve the Riemann problem for (1.19), we have to characterize the discontinuous solutions. In the solution of the Riemann problem,  $A$  can be discontinuous only across  $x = 0$ . In each domain  $x < 0$  and  $x > 0$  where  $A$  is constant,  $\partial_x A = 0$ , and we can divide the other terms by the constant area  $A$  so that we are left with the usual barotropic Euler system for which the Rankine-Hugoniot conditions for a shock are

$$\begin{cases} [\varrho u] = \sigma[\varrho], \\ [\varrho u^2 + p] = \sigma[\varrho u], \end{cases} \quad (2.10)$$

where we must assume here that the speed  $\sigma$  is  $\neq 0$ . Recall also that when  $A$  is smooth, the Rankine-Hugoniot conditions for system (1.18) are

$$\begin{cases} [A\varrho u] = \sigma[A\varrho], \\ [A(\varrho u^2 + p)] = \sigma[A\varrho u], \end{cases} \quad (2.11)$$

and since  $A$  is continuous  $A_+ = A_- = A$ , we may divide by  $A$  so that the jump conditions are naturally the same as (2.10). Moreover, we note that these conditions are not those of system (2.8).

Now, we need to consider a stationary discontinuity for (1.19), away from resonance, and we focus on the term  $p\partial_x A$  which has to be given a precise definition; the definition is not “natural,” in the sense that there is no reason for  $p$  to be continuous when the section is discontinuous. However, following the previous section, we can now use system (2.8) which provides two  $\lambda_0$ -Riemann invariants,  $A\varrho u$  and  $\frac{u^2}{2} + \pi$ , which define the  $\lambda_0$ -contact discontinuity curve (see Definition 2.1). Thus, across  $x = 0$ , we have the jump relations

$$\begin{cases} [A\varrho u] = 0, \\ [\frac{u^2}{2} + \pi] = 0, \end{cases} \quad (2.12)$$

for two states  $\mathbf{w}_- = (\varrho_-, u_-, A_L)^T$  and  $\mathbf{w}_+ = (\varrho_+, u_+, A_R)^T$  connected by such a discontinuity. If  $\varrho_-, u_-$  are known, since  $A_{L,R}$  are also given, the jump relations (2.12) provide the value of  $\varrho_+, u_+$  and then also  $p_\pm = p(\varrho_\pm)$ . The nonconservative product  $p\partial_x A$  is a Dirac measure concentrated on the axis  $x = 0$  with known weight since  $p_+, p_-$  can be computed.

We are going to show that this stationary solution between  $\mathbf{w}_\pm$  is a limit of time-independent smooth solutions of (1.19) (as already seen in the scalar case in the proof of Lemma 2.4). In general, in order to define a nonconservative product, one can look for traveling waves whose limit defines an admissible discontinuity (see Chap. II, Remark 5.2). Here we have a standing discontinuity, and we look for stationary waves.

Considering smooth solutions, they are solution of the equivalent system (2.8), and if they are moreover stationary, they satisfy the system of ODE obtained from (2.8)

$$\begin{cases} (A\varrho u)' = 0, \\ (\frac{u^2}{2} + \pi)' = 0, \end{cases} \quad (2.13)$$

where the derivatives are wrt. to the variable  $x$ .

### Lemma 2.5

*Equations (2.13) define a curve in the set of states, which is an integral curve of the (eigen)vector field associated to the eigenvalue  $\lambda_0 = 0$ .*

*Proof.* Equations (2.13) provide a parametrization by  $x$  of a curve in the set of states described by  $\mathbf{w} = (\varrho, u, A)^T$ . Let us be given a state  $\mathbf{w}_0 = (\varrho_0, u_0, A_0)^T$ ; integrating (2.13) gives a set of states  $\varrho, u, A$ , depending on  $x$  which defines a curve  $\mathcal{C}_0$  in the state space, passing through  $\mathbf{w}_0$ . This curve can be parametrized by  $u$  in the  $\mathbf{w}$ -space. Indeed, write

$$\pi(\varrho(u)) = u_0^2/2 + \pi(\varrho_0) - u^2/2, \quad A(u) = A_0\varrho_0 u_0 / \varrho(u)u,$$

since  $\pi' > 0$ , we can invert the first equation to obtain  $\varrho(u)$ . For instance, assume that  $p(\varrho) = \kappa\varrho^\gamma$ ,  $\gamma > 1$ , then  $\pi(\varrho) = \frac{\kappa\gamma}{\gamma-1}\varrho^{\gamma-1}$ , and the first equation gives  $\varrho(u)$ , the second  $A(u)$ .

Now, let us note  $\mathbf{h}(\mathbf{V}) = (A\varrho u, u^2/2 + \pi, 0)^T$  the conservative flux of (2.8) associated to the other conservative variable  $\mathbf{V} = (A\varrho, u, A)^T$ . We easily check that the curve, considered now in the set of these conservative variables, is an integral curve of  $\mathbf{s}_0(\mathbf{V})$  where  $\mathbf{s}_0$  denotes an eigenvector of  $\mathbf{h}'(\mathbf{V})$  corresponding to the eigenvalue  $\lambda_0 = 0$ . Indeed, differentiating the relation  $\mathbf{h}(\mathbf{v}(x)) = \mathbf{h}(\mathbf{v}_0)$  implies  $\mathbf{h}'(\mathbf{v}) \frac{d\mathbf{v}}{dx} = 0$ ; hence,  $\frac{d}{dx}\mathbf{v}$  is an eigenvector associated to the eigenvalue  $\lambda_0 = 0$ . Another way to look at the result is that we deduce from (2.8) that  $h_1 = A\varrho u$ ,  $h_2 = \frac{u^2}{2} + \pi$  are  $\lambda_0$ -Riemann invariants. Indeed, the lines of the Jacobian matrix of (2.8) are  $Dh_1$ ,  $Dh_2$  and  $(0, 0, 0)$ , where we denote  $Dh_i$  the differential of  $h_i$ , with components the partial derivatives of  $h_i$  wrt. the components of conservative variable  $\mathbf{V} = (A\varrho, u, A)$ . Then, by definition of an eigenvector associated to  $\lambda_0$ , we have necessarily  $Dh_1 \cdot \mathbf{s}_0 = Dh_2 \cdot \mathbf{s}_0 = 0$ .  $\square$

### Lemma 2.6

If  $\mathbf{w}_-$  and  $\mathbf{w}_+ = (\varrho_+, u_+, A_+)^T$  are such that the jump relations (2.12) hold, there exists a sequence of smooth stationary solutions of (2.13) which converges as  $\varepsilon \rightarrow 0$  to

$$\mathbf{w}(x) = \begin{cases} \mathbf{w}_-, & x < 0 \\ \mathbf{w}_+, & x > 0. \end{cases}$$

*Proof.* Let us be given two such states  $\mathbf{w}_\pm = (\varrho(u_\pm), u_\pm, A_\pm)^T$  and introduce for  $\varepsilon > 0$  a smooth monotone function  $u_\varepsilon(x)$  such that

$$u_\varepsilon(x) = \begin{cases} u_-, & x < 0 \\ u_+, & x > \varepsilon. \end{cases}$$

Consider the corresponding states  $\mathbf{w}_\varepsilon(x) = (\varrho(u_\varepsilon(x)), u_\varepsilon(x), A(u_\varepsilon(x)))^T$ , these states belong to the same integral curve, say  $\mathcal{C}_-$ , through  $\mathbf{w}_- = (\varrho(u_-), u_-, A_-)^T$  defined in the previous lemma. Taking the limit of these stationary smooth solution  $(\varrho(u_\varepsilon(x)), u_\varepsilon(x), A(u_\varepsilon(x)))$  as  $\varepsilon \rightarrow 0$  defines a state  $(\varrho(u_+), u_+, A_+)^T$  which belongs to the curve  $\mathcal{C}_-$ . Since we have assumed that the jump relations (2.12) hold, this state coincides with  $\mathbf{w}_+$ . Thus a stationary contact discontinuity connecting the states  $\mathbf{w}_-$  and  $\mathbf{w}_+$  is the limit of smooth stationary solutions.  $\square$

Let us emphasize again that the resulting definition of the nonconservative product  $p(\varrho)\partial_x A$  does not imply that  $p$  is continuous when  $A$  is discontinuous:  $p$  is not a Riemann invariant associated to  $\lambda_0$ . From (2.8) we have seen that the Riemann invariants may be chosen as  $h_1 = A\varrho u$  and  $h_2 = u^2 + \pi(\varrho)$ , the two invariants which are involved in the definition of the jump relations (2.12).

The above results, which was already proved in the scalar case (see Lemma 2.4) can also be extended to systems of the form (2.2) provided they have an equivalent conservative form (for smooth solutions) and away from resonance (see Lemma 2.2).

This proves that the definition of the nonconservative product for (2.2) which we have given is coherent: find enough conservation laws satisfied by smooth solutions, and use the corresponding  $\lambda_0$ -Riemann invariants which they provide. Equivalently, the contact discontinuity connects two states on the same integral curve of the associated vector field; eventually they are also obtained as limit of smooth stationary solutions.

*Remark 2.4.* We might equivalently have started from the entropy conservation for smooth solutions (2.9): it is easy to check that if the jump conditions (2.12) are satisfied, then

$$[A(\varrho e + p)u] = 0. \quad (2.14)$$

Indeed  $A(\varrho e + p)u$  is another Riemann invariant for  $\lambda_0$ , but it is not independent from the two others.  $\square$

*Remark 2.5.* Among the solutions of (2.12), we find steady states at rest  $u = 0$  and  $[\varrho] = 0$  (no jump across  $x = 0$ ).  $\square$

The Riemann problem is however much more delicate to solve unless the initial states are nearby and in the same region wrt. to resonance (recall the 3 regions  $\lambda_- < \lambda_+ < 0$ ,  $\lambda_- < 0 < \lambda_+$ ,  $0 < \lambda_- < \lambda_+$ ). In that case, one derives as usual the wave curves (rarefactions and shocks satisfying the Lax entropy condition) associated with the characteristic fields  $\lambda_{\pm}$ , which can be parametrized by  $\varrho$  and are proved to be monotone increasing for  $\lambda_+$  (resp. decreasing for  $\lambda_-$ ). The solution is composed of at most three waves, those associated to  $\lambda_{\pm}$  across which  $A$  is constant, and a standing wave across which  $A$  jumps. The precise computations follow exactly those of the barotropic Euler system.

Resonance occurs in transonic regime with possible stationary shocks, and one has to analyze the situation when a shock velocity can vanish. Neither uniqueness nor existence is ensured for general data, and one has to add some criterion [579]. Here a first monotonicity criterion is natural which is the monotonicity of the cross section  $A$ , but it does not bring uniqueness. It is even possible to construct solutions with two stationary waves (superimposing a stationary shock with a contact discontinuity) [51, 978].

*Remark 2.6.* In particular, if the standing wave (LD field) superimposes with a stationary shock (GNL field), the conservation of energy (2.14) can no longer hold, since conservation of energy does not hold for the GNL field and is replaced by an inequality

$$[A(\varrho e + p)u] \leq \sigma[A\varrho e] \quad (2.15)$$

which will give a strict inequality when  $\sigma = 0$  (see [361]).  $\square$

*Remark 2.7.* The above model of PDE is also involved in other physical situations and thus is interesting to study. As already written, it is a simple model for flow in a porous medium [312, 606] and also for some models of two-phase flow (the so-called homogeneous equilibrium models (HEM [37])). The term  $A$  will have a different meaning; it represents the local porosity in the first case, or  $A$  represents the volume fraction of one phase in the last one; the pressure law will differ. It is also involved in the derivation of numerical schemes for two-phase flow modeled by a Baer-Nunziato system [52, 352]; see also [350].  $\square$

## 2.4 The Example of the Shallow Water System

The shallow water equations, also referred to as the Saint-Venant system, govern the flow of an incompressible fluid with free surface when the depth of the fluid is small when compared to the characteristic dimensions of the problem. As seen in Example 1.7 above, these equations are obtained from averaging the incompressible Euler equations assuming that the pressure is hydrostatic and neglecting dissipative effects. Consider the system (1.16)

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + \frac{1}{2}gh^2) = -ghZ'(x), \end{cases} \quad (2.16)$$

where  $h$  is the water height,  $u$  the horizontal velocity (obtained by averaging the fluid velocity in the vertical direction)),  $g$  the gravity constant, and  $Z(x)$  is the fixed bottom (or bathymetry). System (2.16) is a system in variable  $\mathbf{u} = (h, hu)^T$  with geometric source term which enters the frame (2.1), with  $\mathbf{f} = \mathbf{f}(\mathbf{u}) = (hu, hu^2 + \frac{1}{2}gh^2)^T$ , not depending on  $Z$ . The homogenous part is a barotropic Euler system provided we identify the height of the column of water  $h$  with  $\varrho$  and we define the pressure by  $p(h) = \frac{1}{2}gh^2$ . Hence, the eigenvalues are easily computed  $\lambda_1(\mathbf{u}) = u - \sqrt{gh}$ ,  $\lambda_2(\mathbf{u}) = u + \sqrt{gh}$ ; they give GNL fields.

It has also obviously some common features with flow in a nozzle (1.18), if we set  $h = \varrho A$ , and because of the geometric source term obtained with the averaging procedure, though the section  $A$  and the topography  $Z$  do not play exactly the same role.

As for the flow in a nozzle, it is usual to introduce the equation  $\partial_t Z = 0$  and to consider the larger system

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + \frac{1}{2}gh^2) + gh \partial_x Z = 0, \\ \partial_t Z = 0, \end{cases} \quad (2.17)$$

involving again a nonconservative product  $h \partial_x Z$ .

We are faced with the problem of resonance when  $u = \pm\sqrt{gh}$ , since the eigenvectors are no longer a basis, and again a GNL field superimposes a LD one. The flow is *torrential* if  $|u| > \sqrt{gh}$  (the analog of supersonic flow in hydrodynamics),  $Fr = |u|/\sqrt{gh}$  is called the Froude number, and  $Fr < 1$  is the subcritical or *fluvial* regime. Moreover, the eigenvalues coincide when the water depth vanishes  $h = 0$ , which corresponds to vacuum in gas dynamics.

We have the analog of Proposition 2.2.

*Proposition 2.3*

*Smooth solutions of (2.17) satisfy the following (energy) conservation law*

$$\partial_t \mathcal{E} + \partial_x \mathcal{F} = 0, \text{ with } \mathcal{E} = \frac{1}{2}hu^2 + \frac{1}{2}gh^2 + ghZ, \quad \mathcal{F} = (\mathcal{E} + \frac{1}{2}gh^2)u. \quad (2.18)$$

The function  $(h, hu, Z) \mapsto \mathcal{E}$  is convex.

If we again compare with a barotropic Euler flow, we can define the “internal energy”  $e'(h) = p(h)/h^2 = g/2$ , then the total energy  $E = \frac{1}{2}hu^2 + he(h)$ , and  $\mathcal{E}$  writes  $\mathcal{E} = E + ghZ$ , while the energy flux  $F = (E + p(h))u$  and  $\mathcal{F} = F + ughZ$ . The entropy (in fact energy) conservation is obtained as for the entropy relation for Euler, taking the source term into account. The convexity of  $\mathcal{E}$  comes from the result proved for the barotropic Euler system and the strict convexity of the energy  $E$ ;  $\mathcal{E}$  is not strictly convex wrt.  $Z$  since  $\partial_{ZZ}^2 \mathcal{E} = 0$ ; note also  $\partial_{Z,h}^2 \mathcal{E} = \partial_{Z,(hu)}^2 \mathcal{E} = 0$ ; however, since  $Z$  is linearly degenerate, there is no problem linked: the stationary contact discontinuities will satisfy (by definition) the energy conservation law.

Discontinuous solutions of (2.17) satisfy the following entropy inequality

$$\partial_t \mathcal{E} + \partial_x \mathcal{F} \leq 0,$$

which can be written directly from the inequality for the barotropic Euler system with a source term,

$$\partial_t E + \partial_x F \leq ghu \partial_x Z,$$

multiplying on the left system (2.16) by  $E' = (\partial_h E, \partial_{hu} E)$ . The nonconservative product in the right-hand side has a natural definition. Indeed,  $Z$  has only stationary discontinuities, and then  $hu$  is continuous across such a stationary discontinuity as results from the Rankine-Hugoniot relation for (2.17).

One is interested in preserving some stationary solutions or equilibria. Again, we intend to use the results of Sect. 2.2. Let us first derive another conservation law for smooth solutions. One can easily check that

$$\partial_t u + \partial_x(u^2/2 + g(h + Z)) = 0.$$

Thus smooth equilibria satisfy

$$hu = cst, \quad u^2/2 + g(h + Z) = cst. \quad (2.19)$$

Equivalently,  $w_1 = hu$ ,  $w_2 = u^2/2 + g(h + Z)$  are Riemann invariants for the LD field  $\lambda_0 = 0$  of system (2.17). From system (2.19), if  $h$  does not vanish, we may write  $u$  in terms of  $h$ , and the second equation then gives a third-order polynomial in  $h$ .

Among physically relevant equilibria are states at rest

$$u = 0, \quad h + Z = cst;$$

we get the so called “lake at rest” test case; the term  $H = h + Z$  is the total height from some reference horizontal level. It is important for a scheme to preserve such solutions: well-balanced schemes are designed in that aim.

Now, even if the system (2.16) was derived assuming the function  $Z$  is regular, when solving the Riemann problem for (2.17), we need to define the nonconservative product  $h\partial_x Z$  in the second equation of (2.17), at a discontinuity of  $Z$  (which corresponds to a stationary wave). The above derivation of smooth stationary waves gives a possible definition for a contact discontinuity; from (2.19) the jumps at  $x = 0$  are required to satisfy

$$[hu] = 0, \quad [u^2/2 + g(h + Z)] = 0,$$

which expresses the fact that  $hu$  and  $u^2/2 + g(h + Z)$  are  $\lambda_0$ -Riemann invariants: these two relations define the  $\lambda_0$ -wave curves according to Definition 2.1. Note that if we use the entropy conservation law also satisfied by smooth solutions, we get  $[\mathcal{F}] = 0$ . Computing  $\mathcal{F} = (\mathcal{E} + \frac{1}{2}gh^2)u = hu(\frac{1}{2}u^2 + g(h + Z))$ , if  $hu$  and  $u^2/2 + g(h + Z)$  are constant, then  $\mathcal{F}$  is indeed constant. Again, this is a general result (see Proposition 2.1).

Away from resonance, the Riemann problem can be solved in the usual way, following the computations done for the isentropic Euler equations, including a stationary wave if necessary. We do not consider here the solution of the Riemann problem when resonance occurs. A GNL wave may be superposed with the stationary wave; we refer to the existing literature, and see, for instance, [25, 301, 912] where a regularization of the topography is considered, replacing the step by a continuous profile in order to select a solution (one says the interface is *thickened*) with also the use of a continuation method; the authors discuss the case of nonuniqueness, [122], and [758] where solutions containing more than one wave of each characteristic family are constructed; we also refer to the recent work [580].

### 3 Specific Numerical Treatment of Source Terms

We will not go into details for all possible schemes; we will try to give some general ideas starting from a few examples, which help understand the chosen approach and possibly extend to other cases.

We consider in this section a class of nonconservative schemes resulting from the above study of systems with a geometric source term. A different approach, with the construction of *simple Riemann solvers*, will be considered in the following section; the resulting schemes may coincide.

We begin by an example; we then consider a property which may be required for the simulation of some physical problems. Indeed, it is important to derive schemes which behave well on some particular solutions, for instance, which preserve states at rest and do not create spurious waves. This needs some more precision on how these states are discretized, i.e., the definition of *discrete equilibria*, as emphasized when we revisited Example 1.3 (see Remark 1.2).

#### 3.1 Some Numerical Considerations for Flow in a Nozzle

We restrict ourselves to the barotropic case. The numerical approach is directly linked to the fact that we have transformed the source term from (1.18) in some partial derivative term, so as to use the Godunov scheme directly on (1.19).

We first discretize the initial data  $(\varrho, u, A)_0(x)$  by a piecewise constant function  $(\varrho, u, A)_\Delta(x, 0) = (\varrho, u, A)_j^0$  on each cell  $C_j = (x_{j-1/2}, x_{j+1/2})$ . If we follow the approach of Godunov's scheme, we integrate the system of PDE on a cell  $(x_{j-1/2}, x_{j+1/2}) \times (0, \Delta t)$ . Since  $A = A_j$  is constant inside the cell, the nonconservative term disappears. Note that the last component  $A$  remains constant,  $A = A_j$  in  $(x_{j-1/2}, x_{j+1/2})$ , the two first components enable the updating of  $(\varrho, \varrho u)$ , and keeping the two first equations leads to the scheme

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda(\mathbf{f}(\mathbf{W}_R(0-; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n)) - \mathbf{f}(\mathbf{W}_R(0+; \mathbf{v}_{j-1}^n, \mathbf{v}_j^n))),$$

where  $\mathbf{u} = (A\varrho, A\varrho u)^T$ ,  $\mathbf{v} = (A\varrho, A\varrho u, A)^T = (\mathbf{u}, A)$ ,  $\mathbf{f}(\mathbf{v}) = (A\varrho u, A(\varrho u^2 + p))^T$ , and the numerical fluxes at each interface involve the solution of the Riemann problem for (1.19), noted  $\mathbf{W}_R$ .

If for a conservative system  $\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{0}$ , the interface flux of Godunov's method is continuous  $\mathbf{f}(\mathbf{w}_R(0+; \mathbf{u}_j, \mathbf{u}_{j+1})) = \mathbf{f}(\mathbf{w}_R(0-; \mathbf{u}_j, \mathbf{u}_{j+1}))$ , thanks to the Rankine-Hugoniot condition, for the present nonconservative system  $\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{v}) - \mathbf{p} \partial_x A = \mathbf{0}$ ,  $\mathbf{p} = (0, p)^T$ , the term  $p \partial_x A$  yields a new stationary wave across which the flux  $\mathbf{f}$  is not continuous. Hence, assuming we know how to compute the exact solution of the Riemann problem for the

enlarged system (1.19), the fluxes may differ on each side of the interface, and one defines  $\mathbf{g}_\pm(\mathbf{u}_j, \mathbf{u}_{j+1}) = \mathbf{f}(\mathbf{W}_R(0\pm; \mathbf{v}_j, \mathbf{v}_{j+1}))$ , which in fact depends also on  $A_j, A_{j+1}$  and should be noted  $\mathbf{g}_\pm(\mathbf{u}_j, \mathbf{u}_{j+1}, A_j, A_{j+1})$ .

Then, in general, the formulation of a finite volume scheme for (1.18), resulting from the discretization of (1.19), is naturally nonconservative and writes

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda(\mathbf{g}_{j+1/2,-}^n - \mathbf{g}_{j-1/2,+}^n), \quad (3.1)$$

with two numerical fluxes  $\mathbf{g}_{j+1/2,\pm}$  at an interface  $x = x_{j+1/2}$ , which differ in case of a nonconstant section. But in case of a constant section, they should be equal, and we require moreover a consistency condition with  $\mathbf{f}$ , the conservative flux (meaning when  $A' = 0$  in (1.18)).

More precisely, as for the exact solution of the Riemann problem  $\mathbf{W}_R(0\pm)$ , we define states on each side of the interface by involving the values of  $A_\pm$  of  $A$ . The left value, say  $\mathbf{v}_{j+1,-}$  at  $x_{j+1/2-}$  (the right interface of  $C_j$ ), is defined by  $A_{j+1,-} = A_j$  and

$$(A\varrho u)_{j+1,-} = (A\varrho u)_{j+1} \\ (u^2/2 + \pi(\varrho))_{j+1,-} = (u^2/2 + \pi(\varrho))_{j+1},$$

with analog formula for the right value at the left interface  $x_{j-1/2+}$  of  $C_j$ , noted with index  $j-1, +$ :  $A_{j-1,+} = A_j$ ,  $(A\varrho u)_{j-1,+} = (A\varrho u)_{j-1}$ ,  $(u^2/2 + \pi(\varrho))_{j-1,+} = (u^2/2 + \pi(\varrho))_{j-1}$ . If the two states  $\mathbf{v}_j$  and  $\mathbf{v}_{j+1}$  are connected by a stationary wave, they satisfy (2.12); it is easy to check that  $\mathbf{v}_{j+1,-} = \mathbf{v}_j$ . Similarly if  $\mathbf{v}_{j-1}$  and  $\mathbf{v}_j$  are connected by a stationary wave, we have  $\mathbf{v}_{j-1,+} = \mathbf{v}_j$ . The two interface states *reconstructed* from the state inside the cell, taking into account the values of the section  $A$ , in order to ensure the continuity of Riemann invariants, are plugged in a consistent numerical flux for the system without source,  $\mathbf{g}_{j-1/2,+} = \mathbf{g}(\mathbf{u}_{j-1,+}, \mathbf{u}_j)$  and  $\mathbf{g}_{j+1/2,-} = \mathbf{g}(\mathbf{u}_j, \mathbf{u}_{j+1-})$ ; the well-balanced property follows (see [718] for details and also [552]).

Note however that, as already pointed out, problems arise with *resonance*, when an eigenvalue of the original system corresponding to a GNL field vanishes, i.e., at a sonic state  $u = \pm c$ . In this case, the two eigenvalues coincide, and the other being associated to an LD scheme, multiple solutions can be constructed. A numerical scheme will select one solution which must be characterized (for instance, in [978], the solution is compared with a 3D axisymmetric solution); one can modify a given scheme in order that it does compute the solution which some admissibility criterion has selected. Besides, a scheme may behave in some unstable way and present oscillations; some correction is derived in [51, 1119], and [536]. In particular [1119], following [718], treats the system with energy also in the resonant regime; it introduces an admissibility criterion and designs a well-balanced scheme, with a computing corrector that selects the admissible equilibrium state.

Let us mention a few more references for the numerical simulation of a flow in a nozzle: a pioneering paper [638], then [601, 667] with a VFRoe-ncv scheme, and a recent original approach with a relaxation scheme in [361] for

which some properties (consistency, mass conservation, positivity of  $\varrho$  and  $A$ , entropy and well-balanced for steady states at rest) are proved.

The previous considerations, leading to the construction of a nonconservative scheme, where the source term is incorporated in the definition of the interface fluxes, will be naturally extended below to more general situations.

### 3.2 Preserving Equilibria, Well-Balanced Schemes

Consider a system of balance laws, which we write

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{s}(\mathbf{u}, x) \quad (3.2)$$

where the source may depend on  $x$ . An *equilibrium* state is a steady state solution; there is an exact balance between the flux gradient and the source. Thus by definition an *equilibrium* solution of (3.2) satisfies

$$\partial_t \mathbf{u} = 0;$$

this is equivalent to

$$\partial_x \mathbf{f}(\mathbf{u}) = \mathbf{s}(\mathbf{u}). \quad (3.3)$$

The concept of well-balanced numerical scheme introduced in [560] is related to the question of preserving at the discrete level the steady solutions of (3.2), that is, discrete solutions satisfying

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n, \quad j \in \mathbb{Z}; \quad (3.4)$$

however there are many ways to obtain a discrete counterpart of (3.3). For instance, in a finite difference approach, if we localize around the interface  $x_{j+1/2}$ , we get

$$\frac{1}{\Delta x} (\mathbf{f}(\mathbf{u}_{j+1}^n) - \mathbf{f}(\mathbf{u}_j^n)) = \mathbf{s}_{j+\frac{1}{2}}^n, \quad j \in \mathbb{Z}, \quad (3.5)$$

where  $\mathbf{s}_{j+\frac{1}{2}}$  represents the  $x_{j+1/2}$ -interfacial contribution of the source term  $\mathbf{s}$ .

This notion of discrete equilibrium can be defined for any finite difference-type scheme, where  $\mathbf{s}_{j+\frac{1}{2}}^n$  is some consistent approximation of the interfacial contribution of the source term up to second-order (or higher-order) terms in  $\Delta x$ . However, there is no evidence that, in the general case, this definition is relevant. Nothing ensures that every solution of (3.3) can be approximated by a discrete solution of (3.5), where the  $\mathbf{u}_j^n$ 's are computed by a finite volume scheme. Thus, in the applications, it is important to have a proper definition of discrete equilibria, and of schemes preserving these equilibria (if we start from an initial data  $\mathbf{u}_j^0, \forall j \in \mathbb{Z}$  which is a discrete equilibria, we expect

$\mathbf{u}_j^n = \mathbf{u}_j^0, \forall j \in \mathbb{Z}, \forall n > 0$ ), which means that no spurious moving wave is created at the discrete level.

*Remark 3.1.* Moreover, in some situations, we intend to respect the asymptotic behavior wrt. some parameter, say  $\alpha$ , in the source term such as a friction coefficient; the dependence of the abovementioned  $\mathcal{O}(\Delta x)$  error terms on  $\alpha$  will have to be precisely analyzed.  $\square$

Among the many attempts that have been made lately to design so-called *well-balanced* schemes, the equilibria one wants to preserve are characterized at the discrete level so as to include their preservation directly in the numerical method. We begin by illustrating the approach on some examples.

### 3.2.1 Upwinding the Source: A Scalar Example

We have already encountered in the scalar simple Example 1.5 a first device to exactly preserve discrete equilibria: the upwinding of source term; see (1.12). Note that discrete equilibria were precisely defined. This construction can be interpreted so as to extend it to more general systems. We begin by the scalar case.

The idea again consists in introducing a new formulation which transforms the source term in a differential one. For (1.7), it gives simply

$$\partial_t u + \partial_x(au - S) = 0,$$

where  $S(x) = \int^x s(y)dy$  is a primitive of the source term. Thus, equilibria are characterized by  $au - S = \text{const}$ .

For more general equations, one tries first to give some other “explicit” form of the equilibria, for systems it will involve Riemann invariants, and exploit the resulting equation in the derivation of the scheme. Consider for instance Eq. (2.1) in the scalar case (see Example 2.3)

$$\partial_t u + \partial_x f(u) = k(x)s(u),$$

with  $s(u) \neq 0$ . Smooth equilibria satisfy

$$\frac{d}{dx}f(u) = k(x)s(u),$$

equivalently, if  $s$  does not vanish,  $\frac{f'}{s}(u)\frac{du}{dx} = k(x)$ . Let us introduce again a primitive of  $\frac{f'}{s}$  (see (2.7))  $\varphi(u) = \int^u \frac{f'}{s}(v)dv$ , and let  $K$  be a primitive of  $k$ ,  $k = K'$ ; then  $\partial_x \varphi(u(x, t)) = \frac{f'(u)}{s(u)}\partial_x u$ . Set  $w = \varphi(u) - K(x)$ , so that equilibria are characterized by:

$$\frac{d}{dx}f(u) = k(x)s(u) \Leftrightarrow w = \varphi(u) - K(x) = \text{cst}.$$

Then one defines a *discrete equilibria* for (3.2) as

$$\forall j \in \mathbb{Z}, w_j = cst \Leftrightarrow \varphi(u_j) - K_j = cst.$$

For the scheme, let  $g$  be some two-point numerical flux, consistent with  $f$ . One discretizes  $K$  by a piecewise constant function ( $K_j, K_{j+1}$  denote the values of  $K$  respectively on the left and right side of the interface  $x_{j+1/2}$ ). This piecewise constant source term is “distributed” between the interfaces; the formula for updating  $u_j^n$  writes

$$u_j^{n+1} = u_j^n - \lambda(g(u_j^n, u_{j+1,-}^n) - g(u_{j-1,+}^n, u_j^n)), \quad (3.6)$$

where the  $K_j$ 's are involved in the definition of the interface states  $u_{j\pm 1,\mp}^n$  which are constructed in order to preserve a discrete equilibrium at an interface. At each interface  $x_{j\pm 1/2}$ , one writes the conservation of  $w = \varphi(u) - K$

$$\varphi(u_{j+1,-}) - K_j = \varphi(u_{j+1}) - K_{j+1}, \quad \varphi(u_{j-1,+}) - K_j = \varphi(u_{j-1}) - K_{j-1}, \quad (3.7)$$

and assuming that  $\varphi$  is increasing, the states  $u_{j,\pm}$  are well defined. Note that even if it does not directly appear in the formula (3.6), the resulting interface fluxes  $g_{j+1/2-} = g(u_j, u_{j+1,-})$ ,  $g_{j+1/2+} = g(u_{j+1,+}, u_{j+1})$  depend not only on  $(u_j, u_{j+1})$  but also on  $(K_j, K_{j+1})$ , so that  $g_{j+1/2\pm}$  is indeed a function of  $\mathbf{v}_j, \mathbf{v}_{j+1}$  where  $\mathbf{v} = (u, K)$ . In fact  $u_{j+1,-}$  (resp.  $u_{j-1,+}$ ) depends on the increment  $\Delta K_{j+1/2} \equiv K_{j+1} - K_j$  (resp.  $\Delta K_{j-1/2}$ ) which is natural since  $K$  is defined up to an additive constant. Indeed, from (3.7), we may write

$$u_{j+1,-} = r_-(u_{j+1}, \Delta K_{j+1/2}), \quad u_{j-1,+} = r_+(u_{j-1}, \Delta K_{j-1/2}),$$

for some function  $r_-$  (resp.  $r_+$ ) obtained by inverting  $\varphi$ . With these notations, the interface fluxes read

$$g_{j+1/2-} = g(u_j, r_-(u_{j+1}, \Delta K_{j+1/2})), \quad g_{j-1/2+} = g(r_+(u_{j-1}, \Delta K_{j-1/2}), u_j),$$

which allows us to define two new numerical flux functions  $g_-$  (resp.  $g_+$ ) for the right (resp. left) boundary of  $C_j$

$$g_{j+1/2-} = g_-(u_j, u_{j+1}, \Delta K_{j+1/2}), \quad g_{j-1/2+} = g_+(u_{j-1}, u_j, \Delta K_{j-1/2}),$$

satisfying  $g_-(u, v, 0) = g_+(u, v, 0) = g(u, v)$ , yielding in turn the consistency condition  $g_-(u, u, 0) = g_+(u, u, 0) = f(u)$ .

By construction, a discrete equilibrium state  $w_j = cst$  is exactly preserved. Starting from an initial data  $(\mathbf{v}_j^0)_{j \in \mathbb{Z}}$  such that  $\forall j, w_j^0 = w(\mathbf{v}_j^0) = cte$ , we obtain

$$\varphi(u_{j+1,-}^0) - K_j = \varphi(u_{j+1}^0) - K_{j+1} = \varphi(u_j^0) - K_j \Rightarrow u_{j+1,-}^0 = u_j^0,$$

then  $g_{j+1/2-} = g(u_j^0, u_{j+1,-}^0) = f_j^0$ , similarly  $u_{j-1,+}^0 = u_j^0$ ,  $g_{j-1/2+} = f_j^0$ , so that  $\forall j, u_j^1 = u_j^0$ .

Convergence results can be found in [552], under some assumptions, and also in [158] when  $g$  is the Engquist-Osher numerical flux, for which the associated scheme has a kinetic interpretation.

*Remark 3.2.* For the example of Eq. (1.7), or more generally for the equation  $\partial_t u + \partial_x f(u) = s(x)$ , we have  $\varphi = f$ . If  $f' > 0$  and if we take for  $g$  the upwind flux,  $g(u, v) = f(u)$ , (3.6) writes

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

then (3.7) gives  $f(u_j) - f(u_{j-1,+}) = f(u_j) - f(u_{j-1}) - (S_j - S_{j-1})$  so that the scheme can also be written with explicit mention of the source term

$$u_j^{n+1} = u_j^n - \lambda(f(u_j^n) - f(u_{j-1}^n)) + \frac{\Delta t}{\Delta x}(S_j - S_{j-1}).$$

The above scheme preserves exactly the discrete equilibria defined by  $w(u_j) \equiv f(u_j) - S_j = cst$  (constant value).

The construction of the states  $u_{j,\pm}$  in (3.7) can be interpreted equivalently in terms of Riemann invariants for the added standing wave, obtained by adding to (3.2) the equation  $\partial_t S = 0$ . Indeed, (3.7) says that the states  $(u_{j+1,-}, S_j)$  and  $(u_{j+1}, S_{j+1})$ , on each side of the interface  $x = x_{j+1/2}$ , are two states connected by a standing wave, since the Riemann invariant  $w$  is constant. Similarly  $(u_{j-1}, S_{j-1})$  and  $(u_{j+1,+}, S_j)$  at  $x = x_{j-1/2}$ .

For the simple example of Eq. (1.7), i.e., in the case  $f(u) = au$ , the above scheme coincides with the scheme (1.12) provided  $S_j - S_{j-1} = \int_{x_{j-1}}^{x_j} s(y)dy$  is discretized by  $\frac{\Delta x}{2}(s_j + s_{j-1})$  (trapezoid rule, which is second-order accurate). It then preserves the discrete equilibria previously defined (see the lines following (1.12)). Recall that these discrete equilibria had been defined by  $a(u_j - u_{j-1}) = \frac{1}{2}(s_j + s_{j+1})$ , i.e., from a finite difference approximation of the ODE  $a \frac{du}{dx} = s$ , while above, the discrete equilibria are obtained from the discretization of the integral form of the equation:  $au = S + cte$ . The two definitions coincide at second order.

It is interesting to see that both approaches, upwinding the source term or distributing it on each side of the interface so as to preserve the Riemann invariants associated to the standing wave, coincide in this simple case. We also mention [1062] for a preliminary study of convergence of this scheme.  $\square$

We now extend this construction to a general system with geometric source term.

### 3.2.2 Well-Balanced Schemes for Geometric Source Terms

In general, well-balanced schemes are obtained by trying to “incorporate the source” as a differential term and/or by enlarging the system. This approach leads to a definition of equilibria involving Riemann invariants, and at the discrete level also, the approximate Riemann solver is constructed by involving Riemann invariants. For instance, in the case of a geometric source term, starting from (2.1)

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, K) = k(x) \mathbf{s}(\mathbf{u}), \quad (3.8)$$

where the Jacobian matrix  $\mathbf{f}'(\mathbf{u})$ , a shorthand notation for  $\partial_{\mathbf{u}} \mathbf{f}(\mathbf{u}, K)$ , is diagonalizable on  $\mathbb{R}$ ; one may consider the enlarged nonconservative system (2.2), adding the equation for  $K$  with  $K'(x) = k(x)$

$$\begin{cases} \partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, K) - \mathbf{s}(\mathbf{u}) \partial_x K = 0, \\ \partial_t K = 0, \end{cases} \quad (3.9)$$

equivalent with the previous one if  $k$  is smooth.

*Lemma 3.1*

$\mathbf{u}$  is an equilibrium of (3.8) iff  $\mathbf{v} = (\mathbf{u}, K)$  is a stationary solution of (3.9).

The proof is obvious since, by definition of an equilibrium,  $\partial_t \mathbf{u} = 0$  or  $\partial_x \mathbf{f}(\mathbf{u}) = k(x) \mathbf{s}(\mathbf{u})$ , thus  $\partial_t \mathbf{v} = 0$ . This characterization enables one to use Definition 2.1.

In the remaining part of this section, we use the notations of Sect. 2.1.1; in particular, we denote by  $\tilde{\mathbf{r}}_0$  an eigenvector associated to  $\lambda_0 = 0$  (see Lemma 2.1). Moreover, we assume that we are away from resonance.

*Definition 3.1*

A discrete equilibrium  $\mathbf{v}_j^e = (\mathbf{u}_j^e, K_j^e)$ ,  $j \in \mathbb{Z}$ , is defined by  $\forall j$ , the states  $\mathbf{v}_j^e$  lie on the same integral curve  $\mathcal{C}^0$  of  $\tilde{\mathbf{r}}_0$ .

In the set of states, this integral curve may be defined by  $\frac{d\mathbf{v}}{d\xi} = \tilde{\mathbf{r}}_0(\mathbf{v}(\xi))$ ,  $\mathbf{v}(\xi_0) = \mathbf{v}_j^e$ , for some parametrization, and where  $j$  is any index. Then, for any  $\lambda_0$ -Riemann invariant  $w$ ,  $w(\mathbf{v}_j^e)$  is constant. This definition is of practical use only if the equations of the Riemann invariants are easy to invert, meaning one has an explicit expression for the states, leading to effective formula which can be used in an algorithm. Otherwise, the formula has to be discretized at some order, as noticed in Remark 3.2.

For the scheme, let us follow the ideas of Godunov’s method and the approach initiated above for flow in a nozzle (Sect. 3.1). Assume for a while that one knows the exact solution of the (nonconservative) Riemann problem for (3.9), noted  $\mathbf{W}_R(0^\pm; \mathbf{v}_L, \mathbf{v}_R)$ , and put apart the problem of resonance. One discretizes  $K$  by a piecewise constant function  $K_\Delta(x, t) = K_j$ ,  $x \in$

$C_j = (x_{j-1/2}, x_{j+1/2}), \forall t \geq 0$ , with  $K_j = \frac{1}{\Delta x} \int_{C_j} k(y) dy$  (or one may use a quadrature rule) and integrate the equation on a cell  $C_j \times [0, \Delta t]$ .

Since  $K_\Delta$  is now piecewise constant,  $\partial_x K_\Delta = 0$  for  $x \in ]x_{j-1/2}, x_{j+1/2}[$ , and inside the cell, one integrates  $\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, K_j) = \mathbf{0}$ . However, the system in  $\mathbf{v} = (\mathbf{u}, K)$  is nonconservative, and the flux is in general not continuous at an interface (while it is in the classical Godunov's method for a conservation law thanks to Rankine-Hugoniot condition); we take  $\mathbf{g}_{j+1/2,\pm}^n = \mathbf{f}(\mathbf{W}_R(0\pm; \mathbf{v}_j^n, \mathbf{v}_{j+1}^n))$  and one has two fluxes at each interface  $x_{j+1/2}$ , the scheme writes

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda(\mathbf{g}_{j+1/2-}^n - \mathbf{g}_{j-1/2+}^n). \quad (3.10)$$

It is built in order to preserve exactly discrete equilibria, provided they are defined as above.

### Proposition 3.1

Consider as initial data a discrete equilibrium  $\mathbf{v}_j^0 = (\mathbf{u}_j^0, K_j^0)$  in the sense of Definition 3.1, i.e.,  $\forall j, \mathbf{v}_j^0 \in \mathcal{C}^0(\mathbf{v}_0^0)$ . If  $\mathbf{u}_j^n$  is computed by (3.10), with  $\mathbf{g}_{j+1/2\pm} = \mathbf{f}(\mathbf{W}_R(0\pm; \mathbf{v}_j, \mathbf{v}_{j+1}))$ , then  $\forall j \in \mathbb{Z}, n > 0$ ,  $\mathbf{u}_j^n = \mathbf{u}_j^0$ .

*Proof.* If  $\mathbf{v}_j^0$  is a discrete equilibrium, at each interface  $x_{j+1/2}$ , the solution of the Riemann problem is a stationary contact discontinuity and  $\mathbf{f}(\mathbf{W}_R(0+; \mathbf{v}_{j-1}^0, \mathbf{v}_j^0)) = \mathbf{f}(\mathbf{W}_R(0-; \mathbf{v}_j^0, \mathbf{v}_{j+1}^0)) = \mathbf{f}(\mathbf{u}_j, K_j)$ . Then  $\mathbf{g}_{j-1/2+}^0 = \mathbf{g}_{j+1/2-}^0$ , which implies  $\mathbf{u}_j^1 = \mathbf{u}_j^0$ , and the result follows.  $\square$

More generally, the interface flux  $\mathbf{g}_{j+1/2\pm}$  can be defined as a numerical flux  $\mathbf{g}$ , which we assume continuous and consistent with  $\mathbf{f}$  when  $k' = 0$ , evaluated on reconstructed states  $\mathbf{v}_{j\pm}$ , where the states are defined in order to incorporate some information from the source. We have seen a first example in (3.7) above. The interface fluxes  $\mathbf{g}_{j+1/2\pm}$  then also depend on  $K_j, K_{j+1}$ .

### Definition 3.2

A scheme is well-balanced if it preserves exactly discrete equilibria in the sense of Definition 3.1: if the states  $\mathbf{v}_j^0$ , for all  $j \in \mathbb{Z}$ , lie on the same integral curve  $\mathcal{C}^0$  of  $\tilde{\mathbf{r}}_0$ , then  $\forall n > 0$ , the states  $\mathbf{v}_j^n$  lie on this same  $\mathcal{C}^0$ .

It is possible to restrict the property to a certain family of equilibria, typically the so-called lake at rest in the case of shallow water equations.

By Definition 3.1, discrete equilibria can be defined by the  $\lambda_0$ -Riemann invariants if there are enough Riemann invariants (in the sense that they have explicit expressions); this is the case for flow in a nozzle with the two invariants  $(A\varrho u, u^2/2 + \pi)$ , see (2.12), and will be illustrated more thoroughly with the shallow water equations.

### Proposition 3.2

Assume  $w_\ell, 1 \leq \ell \leq p$  are  $p$  independent Riemann invariants associated to the  $\lambda_0$ -characteristic field. Let the interface states be defined by

$$\begin{cases} \mathbf{v}_{j+1,-} = (\mathbf{u}_{j+1,-}, K_j), \mathbf{v}_{j-1,+} = (\mathbf{u}_{j,-}, K_j) \text{ where} \\ w_\ell(\mathbf{v}_{j-1,+}) = w_\ell(\mathbf{v}_j), w_\ell(\mathbf{v}_{j+1,-}) = w_\ell(\mathbf{v}_{j+1}), 1 \leq \ell \leq p. \end{cases} \quad (3.11)$$

Define two interface fluxes  $\mathbf{g}_\pm$  by the relations

$$\begin{cases} \mathbf{g}_{j-1/2+} = \mathbf{g}(\mathbf{u}_{j-1,+}, \mathbf{u}_j) \\ \mathbf{g}_{j+1/2-} = \mathbf{g}(\mathbf{u}_j, \mathbf{u}_{j+1,-}) \end{cases} \quad (3.12)$$

Then scheme (3.10) is well-balanced in the sense of Definition 3.2.

*Proof.* If we have a discrete steady state,  $w_\ell(\mathbf{v}_j^0) = w_\ell(\mathbf{v}_{j+1}^0)$  (the value does not depend on  $j$ ) and this holds for any  $\ell$ . Then the three states  $\mathbf{v}_j^0, \mathbf{v}_{j+1,-}^0$  and  $\mathbf{v}_{j+1}^0$  are on the same integral curve  $\mathcal{C}^0$ , and the first two ones involved in  $\mathbf{g}_{j+1/2-}^0$  have the same value  $K = K_j$ , so that  $\mathbf{v}_{j+1,-}^0 = \mathbf{v}_j^0$ , hence

$$\mathbf{g}_{j+1/2-}^0 = \mathbf{g}(\mathbf{u}_j^0, \mathbf{u}_j^0) = \mathbf{f}(\mathbf{u}_j^0, K_j).$$

Similarly the three states  $\mathbf{v}_{j-1}^0, \mathbf{v}_{j-1,+}^0$ , and  $\mathbf{v}_j^0$  are on the same integral curve  $\mathcal{C}^0$ , and the last two ones involved in  $\mathbf{g}_{j-1/2+}^0$  have the same value  $K = K_j$ , so that  $\mathbf{v}_{j-1,+}^0 = \mathbf{v}_j^0$ , hence

$$\mathbf{g}_{j-1/2+}^0 = \mathbf{g}(\mathbf{u}_j^0, \mathbf{u}_j^0) = \mathbf{f}(\mathbf{u}_j^0, K_j).$$

This implies in turn that  $\mathbf{u}_j^1 = \mathbf{u}_j^0$ , and the discrete equilibria is preserved.  $\square$

If we have  $p - 1$  independent Riemann invariants, the relations (3.11) can indeed define the interface states, at least from a theoretical point of view, which means we may write

$$\mathbf{u}_{j+1,-} = \mathbf{r}_-(\mathbf{u}_{j+1}, K_j, K_{j+1}), \mathbf{u}_{j-1,+} = \mathbf{r}_+(\mathbf{u}_{j-1}, K_{j-1}, K_j).$$

Then we can define two numerical flux functions  $\mathbf{g}_\pm$  by the relations

$$\begin{cases} \mathbf{g}_{j-1/2+} = \mathbf{g}(\mathbf{u}_{j-1,+}, \mathbf{u}_j) = \mathbf{g}_+(\mathbf{v}_{j-1}, \mathbf{v}_j), \\ \mathbf{g}_{j+1/2-} = \mathbf{g}(\mathbf{u}_j, \mathbf{u}_{j+1,-}) = \mathbf{g}_-(\mathbf{v}_j, \mathbf{v}_{j+1}). \end{cases}$$

In a number of cases, in particular for the shallow water system, the flux functions  $\mathbf{g}_\pm(\mathbf{v}_L, \mathbf{v}_R)$  depend only on  $\Delta K = K_R - K_L$ , not on each value  $K_{L/R}$ , which we may write  $\mathbf{g}_\pm(\mathbf{v}_L, \mathbf{v}_R) = \mathbf{g}_\pm(\mathbf{u}_L, \mathbf{u}_R, K_R - K_L)$ , where  $\mathbf{v}_{L/R} = (\mathbf{u}_{L/R}, K_{L/R})$ . This assumption is quite natural since in the original Eq. (3.8),  $K$  is only present through  $k = K'$ . This is not the case for the flux in a nozzle because  $A$  is involved in a different way in the source term and it is rather  $\mathbf{g}_\pm(\mathbf{v}_L, \mathbf{v}_R) = \mathbf{g}_\pm(\mathbf{u}_L, \mathbf{u}_R, \frac{A_R}{A_L})$  (as results from the sys-

tem after (1.17) and is clear from formulas (2.12)); let us also mention the Saint-Venant system with variable pressure in [163], section 4.11.1.

For instance, in the case of a scalar law with  $\partial_K f = 0$ , of Example 2.3, we have seen above after (3.7) that it was indeed the case, and we have written

$$u_{j+1,-} = r_-(u_{j+1}, \Delta K_{j+1/2}), \quad u_{j-1,+} = r_+(u_{j-1}, \Delta K_{j-1/2}).$$

We may also consider more general left and right fluxes  $\mathbf{g}_\pm$  which satisfy a consistency condition with (3.2) which writes

$$\forall \mathbf{v} = (\mathbf{u}, K), \quad \mathbf{g}_\pm(\mathbf{v}, \mathbf{v}) = \mathbf{f}(\mathbf{u}, K), \quad (3.13)$$

and are such that they coincide if  $K$  is constant (we have only one flux at each interface for a conservative system)

$$K_1 = K_2 \Rightarrow \mathbf{g}_+(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{g}_-(\mathbf{v}_1, \mathbf{v}_2).$$

One moreover assumes that  $\mathbf{g}_+ - \mathbf{g}_-$  is smooth enough ( $\mathcal{C}^1$ , for instance).

In order to make the consistency with the source term more explicit, we may also require a condition such as

$$\forall \mathbf{v}_1, \mathbf{v}_2, \quad \mathbf{g}_+(\mathbf{v}_1, \mathbf{v}_2) - \mathbf{g}_-(\mathbf{v}_1, \mathbf{v}_2) = (K_2 - K_1)\mathbf{s}(\mathbf{u}) + o(K_2 - K_1),$$

as  $\mathbf{v}_1, \mathbf{v}_2 \rightarrow \mathbf{v} = (\mathbf{u}, K)$ , which emphasizes the dependence on  $\Delta K = K_2 - K_1$ .

Eventually, one can define some stability properties (in general the preservation of some convex invariant domain, typically positive water height for the shallow water system) or a discrete entropy inequality if system (3.6) is endowed with an entropy-entropy flux pair, such that  $\partial_t U(\mathbf{v}) + \partial_x F(\mathbf{v}) = 0$  for smooth solutions, with  $U$  strictly convex in  $\mathbf{u}$ . The discrete entropy inequality then writes

$$U(\mathbf{u}_j^{n+1}, K_j) \leq U(\mathbf{u}_j^n, K_j) - \lambda(G_{j+1/2}^n - G_{j-1/2}^n), \quad (3.14)$$

with  $G_{j+1/2} = G(\mathbf{v}_j, \mathbf{v}_{j+1})$  and the numerical entropy flux  $G$  satisfying the usual consistency condition  $G(\mathbf{v}, \mathbf{v}) = F(\mathbf{v})$ . We will detail some general formalism for the *simple schemes* introduced below. Note that for an explicit scheme, stability properties always need some CFL condition which must be specified in each example.

We do not take up the subject of convergence; see [32, 677] (both concern the scalar case).

The above results give only the great lines of the approach. Moreover we assumed we were away from resonance, which occurs when a genuinely nonlinear eigenvalue of  $\mathbf{f}$  vanishes, and for systems, the regularity of the numerical fluxes becomes problematic. For the shallow water system, these critical points correspond to  $u = \pm\sqrt{gh}$ . We go a little further in this example, but not very far, and we will refer to the existing literature for a deeper study.

### 3.3 Schemes for the Shallow Water System

#### 3.3.1 Important Properties

There has been recent tremendous efforts to derive schemes with good properties for system (2.16)

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + \frac{1}{2}gh^2) = -ghZ'(x), \end{cases}$$

where  $z = Z(x)$  denotes the bottom topography and  $h$  is the height of water (between the bottom and the free surface). It enters in the general frame (3.8)(3.9) with  $K = Z$ ,  $\mathbf{u} = (h, hu)$ , a flux  $\mathbf{f}(\mathbf{u}) = (hu, hu^2 + gh^2/2)$  not depending on  $K$ , and a source  $\mathbf{s} = (0, -gh)$ . The required properties are preservation of equilibria, in particular the so-called lake at rest satisfying  $u = 0, h + Z = cst$  (see below), and robustness in case of dry zones, i.e., when  $h = 0$  (corresponding to vacuum in gas dynamics). Moreover, it is important to preserve the positivity of the water height:  $h \geq 0$ , and one requires the scheme to behave well if the topography has a steep gradient (even if the model is derived by assuming that the bottom  $Z$  is smoothly varying). It is quite difficult to derive a scheme having all these qualities. Many approaches for the derivation of *well-balanced* schemes follow the pioneer work of Greenberg and LeRoux [560] then Gosse and LeRoux [554].

First, equilibria for (2.16) satisfy (2.19)

$$hu = cst, \quad u^2/2 + g(h + Z) = cst,$$

and, according to Definition 3.1, discrete equilibria are naturally defined by

$$h_j u_j = cst, \quad u_j^2/2 + g(h_j + Z_j) = cst.$$

Expressing  $h$  in terms of  $u$  from the first equation, plugging in the second equation, leads to find the roots of a polynomial of degree 3. Most recent schemes for (2.16) are constructed to preserve exactly some particular discrete equilibria, the so-called lake at rest

$$u = 0, \quad h_j + Z_j = cst. \tag{3.15}$$

If we focus on the preservation of these specific equilibria, we define the reconstructed interface values by  $u_{j+1,-} = 0 = u_{j-1,+}$ ,  $Z_{j+1,-} = Z_j = Z_{j-1,+}$  and by

$$h_{j+1,-} = h_{j+1} + Z_{j+1} - Z_j, \quad h_{j-1,+} = h_{j-1} + Z_{j-1} - Z_j.$$

The last formulas depend only on  $\Delta Z_{j+1/2} = Z_{j+1} - Z_j$  for  $h_{j+1,-}$  (resp.  $\Delta Z_{j-1/2}$  for  $h_{j-1,+}$ ).

The consistency of the scheme imposes that the numerical fluxes say  $\mathbf{g}_\pm(\mathbf{u}_1, \mathbf{u}_2, \Delta Z)$  have their first component, say  $g_{h,\pm}$ , satisfying  $g_{h,\pm}(\mathbf{u}, \mathbf{u}, 0) = hu$ , while the second one satisfies  $g_{hu,\pm}(\mathbf{u}, \mathbf{u}, 0) = hu^2 + gh^2/2$ .

If we have a discrete equilibria at rest,  $h_j + Z_j = cst$ , we check easily that  $h_{j+1,-} = h_j = h_{j-1,+}$ , hence  $\mathbf{v}_{j+1,-} = \mathbf{v}_j = \mathbf{v}_{j-1,+}$  which implies by consistency of  $\mathbf{g}_\pm$ :  $\mathbf{g}_{j-1/2+} = \mathbf{f}(\mathbf{u}_j) = \mathbf{g}_{j+1/2-}$ , so that the scheme preserves exactly these equilibria. One can find in [1117] a corrector to treat the case of a resonant regime.

A recent method, which we now detail, introduces a more sophisticated reconstruction, the *hydrostatic reconstruction*; see [63]. It ensures the positivity of the water height; moreover, this reconstruction avoids the problem coming from the presence of critical points in the resolution of the system involved in the definition of the above defined reconstructed values.

### 3.3.2 Hydrostatic Reconstruction

The hydrostatic reconstruction method of [63] also involves the reconstruction of states on each side of the interface which are plugged in a numerical flux consistent with the continuous one—the method may be used in an arbitrary solver for the homogeneous problem—and the source is discretized and “distributed” to the cell interfaces. Moreover, it involves a particular discretization of the nonvanishing component of the source term, say  $S \equiv -gh\partial_x Z$ , taking into account the balance between flux and source obtained for static flows. Let us detail the different stages.

Since  $S$  balances the flux in case of a stationary solution, by integrating on a cell  $C_j$  the equation  $\frac{1}{2}\partial_x gh^2 = -gh\partial_x Z$  we get an approximation, say  $S_j$ , of the mean value of the source on the interval,  $\frac{1}{|C_j|} \int_{C_j} S dx$ , in the spirit of Finite Volume schemes  $-\int_{C_j} gh\partial_x Z dx = \frac{g}{2} (h^2(x_{j+1/2-}) - h^2(x_{j-1/2+}))$  which gives

$$|C_j|S_j \sim \frac{g}{2} (\bar{h}_{j+1/2-}^2 - \bar{h}_{j-1/2+}^2). \quad (3.16)$$

The reconstruction of the heights  $\bar{h}_{j+1/2\pm}$  on each side of the interface  $x = x_{j+1/2}$  follows first the above principle of keeping  $h + Z = cst$  (one of the Riemann invariant for a lake at rest), and one defines

$$h_{j+1/2-} + Z_{j+1/2} = h_j + Z_j, \quad h_{j+1/2+} + Z_{j+1/2} = h_{j+1} + Z_{j+1}, \quad (3.17)$$

with reconstructed cell interface bottom values

$$Z_{j+1/2} = \max(Z_j, Z_{j+1}).$$

Now, in order to preserve a positive water height, we must limit the values  $h_{j+1/2\pm}$ ; hence, we define the so-called hydrostatic reconstructed values of

the water height

$$\bar{h}_{j+1/2\pm} = \max(0, h_{j+1/2\pm}).$$

Last we define

$$U_{j+1/2-} = (\bar{h}_{j+1/2-}, \bar{h}_{j+1/2-} u_j)^T, \quad U_{j+1/2+} = (\bar{h}_{j+1/2+}, \bar{h}_{j+1/2+} u_{j+1})^T.$$

One uses the numerical flux  $\mathbf{g}$  of some 3-point scheme for the homogeneous shallow water equations without the nonconservative term due to the bottom topography. Then, the numerical flux  $\mathbf{g}_{j+1/2}^{(H)}$  is defined with the use of the above modified values,  $\mathbf{g}_{j+1/2}^{(H)} = \mathbf{g}(U_{j+1/2-}, U_{j+1/2+})$ .

Eventually, the discretized source term (3.16) is involved using the hydrostatic reconstructed values of the water height and the scheme writes (for a uniform mesh)

$$U_j^{n+1} = U_j^n - \lambda(\mathbf{g}_{j+1/2}^{(H)} - \mathbf{g}_{j-1/2}^{(H)}) + \Delta t \mathbf{S}_j^n.$$

The source (3.16) can be distributed to the cell interfaces

$$\mathbf{S}_j = \mathbf{S}_{j-1/2+} + \mathbf{S}_{j+1/2-},$$

with

$$\mathbf{S}_{j-1/2+} = (0, \frac{g}{2}(\bar{h}_j^2 - \bar{h}_{j-1/2+}^2))^T, \quad \mathbf{S}_{j+1/2-} = (0, \frac{g}{2}(\bar{h}_{j+1/2-}^2 - \bar{h}_j^2))^T,$$

in such a way that we can define left and right interface fluxes as in formulation (3.10) with

$$\mathbf{g}_{j-1/2+} = \mathbf{g}_{j-1/2}^{(H)} + \mathbf{S}_{j-1/2+}, \quad \mathbf{g}_{j+1/2-} = \mathbf{g}_{j+1/2}^{(H)} - \mathbf{S}_{j+1/2-}.$$

By construction, the fluxes  $\mathbf{g}_{j+1/2\pm}$  depend on  $U_j, U_{j+1}, Z_j, Z_{j+1}$  (in fact on the difference  $\Delta Z_{j+1/2}$ ).

Provided the original numerical flux  $\mathbf{g}$  is consistent and possesses some stability properties, the scheme with hydrostatic reconstruction is consistent and well-balanced, and stability properties are also derived. We refer to [63] for details and precise results concerning this scheme; see also Section 4.11 in the textbook of F. Bouchut [163].

There may still be some limitation to this technique, and we refer to [405, 876] for examples.

### 3.3.3 Some Complements on Shallow Water Models

We have chosen to detail a very small part of a large domain concerning the simulation of geophysical flows. Let us mention a few complementary subjects, either from the numerical or from the modeling point of view.

There are many numerical schemes for systems with geometric source terms which treat the particular example of shallow water (see, for instance, [888, 889]). As already said, many methods can be applied when a Riemann or approximate Riemann solver is available; they do not require the modification of the numerical fluxes for the nonlinear convection terms [660, 667]. Let us mention some schemes that come under classical approaches: Godunov-type [132], central schemes are found in [246, 723]; WAF (weighted average flux) methods are described in [1126], and extended in [472]; for path conservative, entropy stable path consistent (ESPC) see [484], path-conservative Osher-type scheme [242, 449], HLLE type [446]. Some schemes also use an hydrostatic reconstruction [141] but may involve the free surface instead of the water height [139].

More specifically, kinetic schemes are relying on a kinetic interpretation which exists for the shallow water system (see Chap. IV, Sect. 7.4.1); we refer to [949]. Together with the hydrostatic reconstruction, it leads to a scheme for which a fully discrete entropy inequality (but with an error term of order the square of the topography jumps) is obtained [64]. This approach may be extended to more complex models [65].

Note that if the exact Riemann solver is used in [301, 912], many works concern approximate Riemann solvers [788], in particular Roe, Roe-type, or VFRoe schemes: [214, 243, 503, 932] (which presents a nice state of the art), [609, 929]; these schemes have some drawback: they may produce negative water heights, which can be corrected via a relaxation approach [141]; similarly HLL-type schemes [249, 470] or relaxation schemes [163] are used. There are many links between the different approaches, and they are detailed in general in [779]; in the context of shallow water, see [141] and [933] which describes precisely the link with a specific relaxation system and a VFRoe scheme.

For the computation of general steady states (moving-water equilibrium) see [25, 910, 1197].

For extension to 2D of well-balanced schemes and experiments such as practical examples of a dam break, see [209, 247, 450, 471, 850], and references therein. For high-order WB numerical schemes [251, 908], among which WENO schemes [909, 1194, 1197]; see also [1196].

Let us also notice that comparison of different schemes [50] shows that in case of nonuniqueness for a Riemann problem, different schemes may compute different solutions.

The analogous of the low Mach regime for compressible flow is the low Froude regime, and the problem is addressed, for instance, in [930]; then one may take into account some physical aspects such as friction [142, 406] or porosity [483].

From the modeling point of view, let us mention some extensions to various geophysical applications: we first quote roll waves, for a theoretical analysis see [903, 973], and their computation [662]. Then, recall that the shallow water equations were obtained from the incompressible Euler equations assuming that the pressure is hydrostatic and neglecting dissipative effects. We can

then consider viscous models, the viscous Saint-Venant system [512] (see also [189]), and then models including non-hydrostatic pressure terms [208] and their simulation [206, 1002] for a kinetic interpretation and [207] for a recent energy consistent model, not forgetting the Green-Naghdi model [734] (and the references therein).

Then multi-layer models are developed in [66, 892], and their computation may be done by various schemes [11, 62, 174, 248, 723].

Let us also mention section-averaged shallow water equations for river hydraulics [398] and [558] for a kinetic interpretation; a model for mixed flows, coupling free surface, and pressurized flow in pipe is found in [181] and its kinetic formulation [182]. Models for gravity-driven flows are derived in [178], then models for two-phase (solid-fluid) flow models for avalanche and debris flows in [953], and the simulation of such models is described in [168, 932, 933].

The list is not exhaustive; it gives a quick overview of many interesting works, some of them in progress.

## 4 Simple Approximate Riemann Solvers

We now extend the HLL (Harten, Lax, and van Leer) approximate Riemann solver to systems with source term in a quite natural way. Indeed, this framework allows a straightforward extension.

### 4.1 Definition of Simple Approximate Riemann Solvers

Consider a system of balance laws (3.2) which we write again for the reader's convenience

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{s}(\mathbf{u}), \quad (4.1)$$

where  $\mathbf{s}(\mathbf{u})$  may or not depend on  $x$  in this subsection, and assume the system is endowed with an entropy which we now denote  $\eta$ , with entropy flux  $\theta$ , so that entropy solutions of (4.1) are required to satisfy the entropy inequality

$$\partial_t \eta(\mathbf{u}) + \partial_x \theta(\mathbf{u}) \leq \sigma(\mathbf{u}), \quad (4.2)$$

with  $\theta'(\mathbf{u}) = \eta'(\mathbf{u})\mathbf{f}'(\mathbf{u})$  and  $\sigma(\mathbf{u}) = \eta'(\mathbf{u})\mathbf{s}(\mathbf{u})$ . Given a Riemann data

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0, \end{cases} \quad (4.3)$$

with  $\mathbf{u}_L$  and  $\mathbf{u}_R$  in  $\Omega$ , we have already noticed that, unlike the homogeneous case ( $\mathbf{s}(\mathbf{u}) = 0$ ), the exact solution of the Riemann problem (4.1), (4.3) that

we denote  $\mathbf{w}(x, t; \mathbf{u}_L, \mathbf{u}_R)$  is no longer self-similar. However an approximate Riemann solver (ARS)  $\tilde{\mathbf{w}}(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$  associated with  $\mathbf{w}(x, t; \mathbf{u}_L, \mathbf{u}_R)$  may be self-similar as in the homogeneous case. We have already given examples of such situations, for instance, in Sect. 3.1. In this section, we will focus on self-similar solvers which we call *simple* approximate Riemann solvers, as in Chap. IV, Sect. 3.1.2, which have the following form

$$\tilde{\mathbf{w}}\left(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R\right) = \begin{cases} \mathbf{u}_1 = \mathbf{u}_L, & \frac{x}{t} < a_1, \\ \mathbf{u}_k, & a_{k-1} < \frac{x}{t} < a_k, \quad k = 2, \dots, l, \\ \mathbf{u}_{l+1} = \mathbf{u}_R, & \frac{x}{t} > a_l, \end{cases} \quad (4.4)$$

with  $\mathbf{u}_k = \mathbf{u}_k(\mathbf{u}_L, \mathbf{u}_R)$  and  $a_k = a_k(\mathbf{u}_L, \mathbf{u}_R)$ ,  $k = 1, \dots, l$ , to be defined. The influence of the source term  $\mathbf{s}(\mathbf{u})$  will appear in the definition of the intermediate states  $\mathbf{u}_k$ ,  $k = 2, \dots, l$ .

We extend in a rather natural way the notions of consistency with the integral form of the balance law seen in Chap. IV, Sect. 3.1 in the homogeneous case, which lead to the definition of Godunov-type schemes.

#### *Definition 4.1*

A simple approximate Riemann solver (4.4) is said to be consistent with the integral form of the balance law (4.1) if and only if there exists a function  $\tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \xi, \tau)$  such that for  $\Delta t, \Delta x$  satisfying

$$\max_{1 \leq k \leq l} |a_k(\mathbf{u}_L, \mathbf{u}_R)| \frac{\Delta t}{\Delta x} \leq \frac{1}{2} \quad (4.5)$$

we have

$$\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) - \Delta x \tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \Delta x, \Delta t) = \sum_{k=1}^l a_k(\mathbf{u}_{k+1} - \mathbf{u}_k) \quad (4.6)$$

with

$$\lim_{(\Delta x, \Delta t, \mathbf{u}_L, \mathbf{u}_R) \rightarrow (0, 0, \mathbf{u}, \mathbf{u})} \tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \Delta x, \Delta t) = \mathbf{s}(\mathbf{u}). \quad (4.7)$$

We will set  $\lambda = \frac{\Delta t}{\Delta x}$  as usual. Similarly, we introduce the notion of consistency with the entropy inequality (4.2).

#### *Definition 4.2*

A simple approximate Riemann solver (4.4) is said to be consistent with the integral form of the entropy inequality (4.2) if and only if there exists a function  $\tilde{\sigma}(\mathbf{u}_L, \mathbf{u}_R; \xi, \tau)$  such that, for  $\lambda = \Delta t / \Delta x$  satisfying (4.5), we have

$$\theta(\mathbf{u}_R) - \theta(\mathbf{u}_L) - \Delta x \tilde{\sigma}(\mathbf{u}_L, \mathbf{u}_R; \Delta x, \Delta t) \leq \sum_{k=1}^l a_k(\eta(\mathbf{u}_{k+1}) - \eta(\mathbf{u}_k)) \quad (4.8)$$

with

$$\lim_{(\Delta x, \Delta t, \mathbf{u}_L, \mathbf{u}_R) \rightarrow (0, 0, \mathbf{u}, \mathbf{u})} \tilde{\sigma}(\mathbf{u}_L, \mathbf{u}_R; \Delta x, \Delta t) = \sigma(\mathbf{u}). \quad (4.9)$$

Note that the definition extends to an arbitrary grid, replacing in an adequate way in the formulas  $\Delta x$  by  $\Delta x_j = x_{j+1/2} - x_{j-1/2}$ , and the time step  $\Delta t$  by  $\Delta t_n = t_{n+1} - t_n$ .

Given a simple approximate Riemann solver  $\tilde{\mathbf{w}}(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$ , we consider the two averages

$$\begin{cases} \tilde{\mathbf{u}}^-(\mathbf{u}_L, \mathbf{u}_R) = \frac{2}{\Delta x} \int_{-\frac{\Delta x}{\Delta t}}^0 \tilde{\mathbf{w}}\left(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R\right) dx, \\ \tilde{\mathbf{u}}^+(\mathbf{u}_L, \mathbf{u}_R) = \frac{2}{\Delta x} \int_0^{\frac{\Delta x}{\Delta t}} \tilde{\mathbf{w}}\left(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R\right) dx \end{cases} \quad (4.10)$$

and define the following update formula

$$\mathbf{u}_j^{n+1} = \frac{1}{2} (\mathbf{u}_{j-\frac{1}{2}}^{n,+} + \mathbf{u}_{j+\frac{1}{2}}^{n,-}) \quad \text{with} \quad \mathbf{u}_{j+\frac{1}{2}}^{n,\pm} = \tilde{\mathbf{u}}^\pm(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n) \quad (4.11)$$

and under the usual CFL condition (4.5)

$$\max_{1 \leq k \leq l} |a_k(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n)| \frac{\Delta t}{\Delta x} \leq \frac{1}{2}$$

for all  $j \in \mathbb{Z}$ . A *Godunov-type* scheme (4.11) is required to satisfy the consistency property (4.6)–(4.7); it is *entropy satisfying* if it obeys in addition (4.8)–(4.9).

Introducing the notation

$$\tilde{\mathbf{g}}(\mathbf{u}_L, \mathbf{u}_R) = \frac{1}{2} \left\{ \mathbf{f}(\mathbf{u}_L) + \mathbf{f}(\mathbf{u}_R) - \sum_{k=1}^l |a_k|(\mathbf{u}_{k+1} - \mathbf{u}_k) \right\}, \quad (4.12)$$

we can write the scheme (4.11) in the usual conservative form with numerical flux  $\tilde{\mathbf{g}}$ . Indeed, we can state

*Proposition 4.1*

Let  $\tilde{\mathbf{w}}(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$  be a simple approximate Riemann solver consistent with the integral form of (4.1) in the sense of Definition 4.1 above. Then, the numerical scheme defined by (4.11) can be written in the following form

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda(\mathbf{g}_{j+\frac{1}{2}}^n - \mathbf{g}_{j-\frac{1}{2}}^n) + \frac{\Delta t}{2}(\mathbf{s}_{j-\frac{1}{2}}^n + \mathbf{s}_{j+\frac{1}{2}}^n) \quad (4.13)$$

with  $\mathbf{g}_{j+\frac{1}{2}}^n = \tilde{\mathbf{g}}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n)$  and

$$\mathbf{s}_{j+\frac{1}{2}}^n = \tilde{\mathbf{s}}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n; \Delta x, \Delta t) \quad (4.14)$$

for all  $j \in \mathbb{Z}$ .

Thus, the usual form of a conservative scheme is recovered in the absence of source term ( $\mathbf{s}(\mathbf{u}) = 0$ ), while in the general setting the influence of  $\mathbf{s}(\mathbf{u})$  is taken into account by means of two interfacial contributions, namely,  $\mathbf{s}_{j-\frac{1}{2}}^n$  and  $\mathbf{s}_{j+\frac{1}{2}}^n$  (and possibly in the intermediate states of the simple solver). Concerning the entropy inequality, introducing the notation

$$\begin{aligned} \tilde{\theta}(\mathbf{u}_L, \mathbf{u}_R) &= \frac{1}{2}(\theta(\mathbf{u}_L) + \theta(\mathbf{u}_R)) \\ &- \frac{\Delta x}{4\Delta t} \left( (\eta(\mathbf{u}_R) - \eta(\tilde{\mathbf{u}}^+(\mathbf{u}_L, \mathbf{u}_R))) - (\eta(\mathbf{u}_L) - \eta(\tilde{\mathbf{u}}^-(\mathbf{u}_L, \mathbf{u}_R))) \right), \end{aligned} \quad (4.15)$$

we have

*Proposition 4.2*

Let  $\tilde{\mathbf{w}}(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R)$  be a simple approximate Riemann solver consistent with the integral form of (4.2) in the sense of Definition 4.2. Then, the numerical scheme defined by (4.11) satisfies the following discrete entropy inequality

$$\eta(\mathbf{u}_j^{n+1}) \leq \eta(\mathbf{u}_j^n) - \lambda(\theta_{j+\frac{1}{2}}^n - \theta_{j-\frac{1}{2}}^n) + \frac{\Delta t}{2}(\sigma_{j-\frac{1}{2}}^n + \sigma_{j+\frac{1}{2}}^n) \quad (4.16)$$

with  $\theta_{j+\frac{1}{2}}^n = \tilde{\theta}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n)$  and  $\sigma_{j+\frac{1}{2}}^n = \tilde{\sigma}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n; \Delta x, \Delta t)$  for all  $j \in \mathbb{Z}$ .

Here again, we recover the usual form of a discrete entropy inequality when  $\mathbf{s}(\mathbf{u}) = 0$ , while, in the presence of a source term, the two interfacial terms  $\sigma_{j \pm \frac{1}{2}}^n$  account for its influence.

## 4.2 Well-Balanced Simple Schemes

Let us study the well-balanced property associated with a *Godunov-type* scheme. We introduce some stronger notions which are easier to handle for schemes which can be written in the form (4.11). Indeed, by definition (4.11), a sufficient condition for (3.4), i.e.,

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n, \quad j \in \mathbb{Z},$$

is

$$\mathbf{u}_j^n = \mathbf{u}_{j-\frac{1}{2}}^{n,+} = \mathbf{u}_{j+\frac{1}{2}}^{n,-}, \quad (4.17)$$

which motivates the following definition.

*Definition 4.3*

The sequence  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  is said to be an equilibrium solution of the Godunov-type numerical scheme (4.11) if (4.17) is satisfied.

The main advantage of (4.17) is that this property satisfied by a stationary numerical solution can be easily connected to the discrete property (3.5) of a stationary solution of system (4.1). More precisely, the next proposition holds true.

*Proposition 4.3*

An equilibrium solution of the Godunov-type numerical scheme (4.11) in the sense of Definition 4.3 satisfies the relation (3.5), i.e.,

$$\frac{1}{\Delta x}(\mathbf{f}(\mathbf{u}_{j+1}^n) - \mathbf{f}(\mathbf{u}_j^n)) = \mathbf{s}_{j+\frac{1}{2}}^n, \quad j \in \mathbb{Z},$$

where  $\mathbf{s}_{j+\frac{1}{2}}^n$  is defined by (4.14).

*Proof.* Let  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  be an equilibrium solution of the Godunov-type numerical scheme (4.11). We first write by definitions (4.10) and (4.11)

$$\mathbf{u}_j^{n+1} = \frac{1}{\Delta x} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}\left(\frac{x}{\Delta t}; \mathbf{u}_j^n, \mathbf{u}_{j+1}^n\right) dx - \frac{1}{2}(\mathbf{u}_{j+\frac{1}{2}}^{n,+} - \mathbf{u}_{j-\frac{1}{2}}^{n,+}). \quad (4.18)$$

But under the CFL condition (4.5), we easily show that

$$\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}\left(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R\right) dx = \frac{\Delta x}{2}(\mathbf{u}_L + \mathbf{u}_R) - \Delta t \sum_{k=1}^l a_k (\mathbf{u}_{k+1} - \mathbf{u}_k),$$

so that thanks to the consistency condition (4.6)

$$\begin{aligned} \int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} \tilde{\mathbf{w}}\left(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R\right) dx &= \frac{\Delta x}{2}(\mathbf{u}_L + \mathbf{u}_R) - \Delta t(\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)) \\ &\quad + \Delta x \Delta t \tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \Delta x, \Delta t). \end{aligned}$$

Plugging this equality in (4.17) gives by definition of  $\tilde{\mathbf{s}}_{j+\frac{1}{2}}^n$  in (4.13)

$$\begin{cases} \mathbf{u}_j^{n+1} = \frac{1}{2}(\mathbf{u}_j^n + \mathbf{u}_{j+1}^n) - \lambda(\mathbf{f}(\mathbf{u}_{j+1}^n) - \mathbf{f}(\mathbf{u}_j^n)) \\ \quad + \Delta t \mathbf{s}_{j+\frac{1}{2}}^n - \frac{1}{2}(\mathbf{u}_{j+\frac{1}{2}}^{n,+} - \mathbf{u}_{j-\frac{1}{2}}^{n,+}). \end{cases} \quad (4.19)$$

From (4.19) we infer

$$\begin{aligned} \mathbf{u}_j^{n+1} &= \mathbf{u}_j^n - \lambda(\mathbf{f}(\mathbf{u}_{j+1}^n) - \mathbf{f}(\mathbf{u}_j^n)) + \Delta t \mathbf{s}_{j+\frac{1}{2}}^n \\ &\quad + \frac{1}{2}(\mathbf{u}_{j-\frac{1}{2}}^{n,+} - \mathbf{u}_{j+\frac{1}{2}}^{n,+} - \mathbf{u}_j^n + \mathbf{u}_{j+1}^n). \end{aligned}$$

The sequence  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  being an equilibrium solution we have

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n = \mathbf{u}_{j-\frac{1}{2}}^{n,+} \text{ and } \mathbf{u}_{j+1}^n = \mathbf{u}_{j+\frac{1}{2}}^{n,+},$$

and relation (3.5) follows.  $\square$

Let us now define the concept of a well-balanced simple scheme.

*Definition 4.4*

The Godunov-type scheme (4.11) is said to be well-balanced if and only if, for any sequence  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  satisfying (3.5), the sequence  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  is an equilibrium solution, i.e., satisfies (4.17).

It turns out that relations (4.17) may be difficult to verify in practice. We are thus led to introduce a stronger (but easier to manipulate) notion of equilibrium solution.

*Definition 4.5*

The sequence  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  is said to be a strong equilibrium solution of the Godunov-type numerical scheme (4.11) if for all  $j \in \mathbb{Z}$  and  $t > 0$

$$\tilde{\mathbf{w}}\left(\frac{x}{t}; \mathbf{u}_j^n, \mathbf{u}_{j+1}^n\right) = \begin{cases} \mathbf{u}_j^n & x < 0, \\ \mathbf{u}_{j+1}^n & x > 0. \end{cases} \quad (4.20)$$

It is clear that (4.20) implies (4.17) so a strong equilibrium solution is an equilibrium solution in the sense of Definition 4.3. To conclude this section, we define the corresponding notion of strongly well-balanced numerical scheme.

*Definition 4.6*

The Godunov-type numerical scheme (4.11) is said to be strongly well-balanced if and only if, for any sequence  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  satisfying (3.5), the sequence  $(\mathbf{u}_j^n)_{j \in \mathbb{Z}}$  is a strong equilibrium solution, i.e., satisfies (4.20).

*Remark 4.1.* Let us interpret (4.20) if  $\tilde{\mathbf{w}}$  involves an exact Riemann solver; it says that we have with (4.20) a stationary solution of the Riemann problem. For a system with a geometric source term, we may have an approximate Riemann solver which is associated to an exact Riemann solver on a larger nonconservative system. This characterization (4.20) is then natural in this case; a discrete equilibrium in the sense of Definition 3.1, corresponding to a stationary contact discontinuity, is indeed a strong equilibrium. It is also the case with some relaxation schemes designed for Euler system with gravity and friction which we describe below.

Then, in link with Remark 2.2, for a system with discontinuous flux, these stationary solutions play an important role and lead in the scalar case to the notion of *germs* for which we refer to [45].  $\square$

### 4.3 Simple Approximate Riemann Solvers in Lagrangian or Eulerian Coordinates

We start by extending the correspondence between simple schemes in Lagrangian and Eulerian frames already seen in Chap. IV, Sect. 3.1 (see (3.20) and (3.24)) for the case with source terms not depending on  $x$ . Let us begin by recalling some notations introduced in Chap. II, Sect. 2. Let first be given a system of partial differential equations in Eulerian coordinates of the following form

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho \Phi) + \partial_x(\varrho \Phi u + \mathbf{g}(\varrho, \Phi)) = \varrho \mathbf{s}(\varrho, \Phi), \end{cases} \quad (4.21)$$

where we have put aside the equation of conservation of mass, which has no source term, from the convection part for the vector  $\varrho \Phi$  of the other conservative variables. We write system (4.21) in the form

$$\partial_t \mathbf{U}^E + \partial_x \mathbf{F}^E(\mathbf{U}^E) = \mathbf{S}^E(\mathbf{U}^E) \quad (4.22)$$

with

$$\mathbf{U}^E = \begin{pmatrix} \varrho \\ \varrho \Phi \end{pmatrix}, \quad \mathbf{F}^E(\mathbf{U}^E) = \begin{pmatrix} \varrho u \\ \varrho \Phi u + \mathbf{g}(\varrho, \Phi) \end{pmatrix}, \quad \mathbf{S}^E(\mathbf{U}^E) = \begin{pmatrix} 0 \\ \varrho \mathbf{s}(\varrho, \Phi) \end{pmatrix}. \quad (4.23)$$

System (4.23) is supplemented with an entropy inequality

$$\partial_t \eta^E(\mathbf{U}^E) + \partial_x \theta^E(\mathbf{U}^E) \leq \sigma^E(\mathbf{U}^E) = (\eta^E)'(\mathbf{U}^E) \cdot \mathbf{S}^E(\mathbf{U}^E), \quad (4.24)$$

where  $\eta^E : \mathbb{R}^n \mapsto \mathbb{R}$  is the (convex) entropy with entropy flux  $\theta^E$ . In Lagrangian coordinates, this system writes, setting  $\tau = \frac{1}{\varrho}$ ,  $m = \int \varrho dx$ ,

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t \Phi + \partial_m \mathbf{g}\left(\frac{1}{\tau}, \Phi\right) = \mathbf{s}\left(\frac{1}{\tau}, \Phi\right), \end{cases} \quad (4.25)$$

or equivalently

$$\partial_t \mathbf{U}^L + \partial_m \mathbf{F}^L(\mathbf{U}^L) = \mathbf{S}^L(\mathbf{U}^L) \quad (4.26)$$

with

$$\mathbf{U}^L = \begin{pmatrix} \tau \\ \Phi \end{pmatrix}, \quad \mathbf{F}^L(\mathbf{U}^L) = \begin{pmatrix} -u \\ \mathbf{g}\left(\frac{1}{\tau}, \Phi\right) \end{pmatrix}, \quad \mathbf{S}^L(\mathbf{U}^L) = \begin{pmatrix} 0 \\ \mathbf{s}\left(\frac{1}{\tau}, \Phi\right) \end{pmatrix}, \quad (4.27)$$

and

$$\partial_t \eta^L(\mathbf{U}^L) + \partial_m \theta^L(\mathbf{U}^L) \leq \sigma^L(\mathbf{U}^L) \quad (4.28)$$

where

$$\eta^E = \varrho\eta^{\mathcal{L}}, \quad \theta^E = \eta^{\mathcal{L}} u + \theta^{\mathcal{L}} \quad \text{and} \quad \sigma^E = \varrho\sigma^{\mathcal{L}}.$$

In fact  $\theta^{\mathcal{L}} = 0$  for fluid systems.

We know that the mapping  $\mathbf{U}^{\mathcal{L}} \mapsto \mathbf{U}^E$  defines an admissible change of variables in  $\mathbb{R}^n$  under the natural assumption  $\varrho > 0$ . In addition, easy calculations (see Chap. II, Theorem 2.1) show that the eigenvalues  $\lambda_k(\mathbf{U}^E)$ ,  $1 \leq k \leq n$ , of the Jacobian matrix  $(\mathbf{F}^E)'(\mathbf{U}^E)$  are related to the eigenvalues  $\mu_k$ ,  $1 \leq k \leq n$ , of the Jacobian matrix  $(\mathbf{F}^{\mathcal{L}})'(\mathbf{U}^{\mathcal{L}})$  by the following relation:

$$\lambda_k(\mathbf{U}^E) = u + \tau\mu_k(\mathbf{U}^{\mathcal{L}}). \quad (4.29)$$

We now consider a simple approximate Riemann solver in Lagrangian coordinates associated with (4.26), that is,

$$\widetilde{\mathbf{W}}^{\mathcal{L}}\left(\frac{m}{t}; \mathbf{U}_L^{\mathcal{L}}, \mathbf{U}_R^{\mathcal{L}}\right) = \begin{cases} \mathbf{U}_1^{\mathcal{L}} = \mathbf{U}_L^{\mathcal{L}}, & \frac{m}{t} < c_1, \\ \mathbf{U}_k^{\mathcal{L}}, & c_{k-1} < \frac{m}{t} < c_k, \quad k = 2, \dots, l, \\ \mathbf{U}_{l+1}^{\mathcal{L}} = \mathbf{U}_R^{\mathcal{L}}, & \frac{m}{t} > c_l, \end{cases} \quad (4.30)$$

where  $l$ , the number of approximate waves, satisfies  $l \leq n$ . We assume that the Rankine-Hugoniot relation associated with the first equation in (4.25) is satisfied across each approximate wave, which implies

$$u_k + c_k \tau_k = u_{k+1} + c_k \tau_{k+1}, \quad \text{for all } k = 1, \dots, l. \quad (4.31)$$

We then define the following natural simple approximate Riemann solver in Eulerian coordinates, setting  $\mathbf{U}_{\alpha}^E = \mathbf{U}^E(\mathbf{U}_{\alpha}^{\mathcal{L}})$ ,  $\alpha = L, R$ ,

$$\widetilde{\mathbf{W}}^E\left(\frac{x}{t}; \mathbf{U}_L^E, \mathbf{U}_R^E\right) = \begin{cases} \mathbf{U}_1^E = \mathbf{U}_L^E, & \frac{x}{t} < a_1, \\ \mathbf{U}_k^E = \mathbf{U}^E(\mathbf{U}_k^{\mathcal{L}}), & a_{k-1} < \frac{x}{t} < a_k, \quad k = 2, \dots, l, \\ \mathbf{U}_{l+1}^E = \mathbf{U}_R^E, & \frac{x}{t} > a_l, \end{cases} \quad (4.32)$$

with

$$a_k = u_k + c_k \tau_k, \quad k = 1, \dots, l. \quad (4.33)$$

Assume now that (4.30) is consistent with the integral form of (4.26) in the sense of Definition 4.1, i.e., there exists a function  $\tilde{\mathbf{S}}^{\mathcal{L}}(\mathbf{U}_L^{\mathcal{L}}, \mathbf{U}_R^{\mathcal{L}}; \xi, \tau)$  such that for  $\Delta t / \Delta m$  satisfying the CFL condition (4.5) one has

$$\mathbf{F}^{\mathcal{L}}(\mathbf{U}_R^{\mathcal{L}}) - \mathbf{F}^{\mathcal{L}}(\mathbf{U}_L^{\mathcal{L}}) - \Delta m \tilde{\mathbf{S}}^{\mathcal{L}}(\mathbf{U}_L^{\mathcal{L}}, \mathbf{U}_R^{\mathcal{L}}; \Delta m, \Delta t) = \sum_{k=1}^l c_k (\mathbf{U}_{k+1}^{\mathcal{L}} - \mathbf{U}_k^{\mathcal{L}}) \quad (4.34)$$

with

$$\lim_{(\mathbf{U}_L^{\mathcal{L}}, \mathbf{U}_R^{\mathcal{L}}, \Delta m, \Delta t) \rightarrow (\mathbf{U}^{\mathcal{L}}, \mathbf{U}^{\mathcal{L}}, 0, 0)} \tilde{\mathbf{S}}^{\mathcal{L}}(\mathbf{U}_L^{\mathcal{L}}, \mathbf{U}_R^{\mathcal{L}}; \Delta m, \Delta t) = \mathbf{S}^{\mathcal{L}}(\mathbf{U}^{\mathcal{L}}). \quad (4.35)$$

Easy calculations show that the validity of (4.34) is equivalent to

$$\mathbf{F}^E(\mathbf{U}_R^E) - \mathbf{F}^E(\mathbf{U}_L^E) - \Delta x \tilde{\mathbf{S}}^E(\mathbf{U}_L^E, \mathbf{U}_R^E; \Delta x, \Delta t) = \sum_{k=1}^l a_k (\mathbf{U}_{k+1}^E - \mathbf{U}_k^E) \quad (4.36)$$

where we have set

$$\Delta x = \frac{\Delta m}{\varrho^*(\mathbf{U}_L^E, \mathbf{U}_R^E)}, \quad (4.37)$$

for some positive density  $\varrho^*$  to be prescribed, and with

$$\tilde{\mathbf{S}}^E(\mathbf{U}_L^E, \mathbf{U}_R^E; \Delta x, \Delta t) = \varrho^*(\mathbf{U}_L^E, \mathbf{U}_R^E) \tilde{\mathbf{S}}^L(\mathbf{U}_L^L, \mathbf{U}_R^L; \Delta m, \Delta t). \quad (4.38)$$

We require that  $\varrho^*(\mathbf{U}_L^E, \mathbf{U}_R^E)$  is such that

$$\lim_{\mathbf{U}_L^E, \mathbf{U}_R^E \rightarrow \mathbf{U}^E} \varrho^*(\mathbf{U}_L^E, \mathbf{U}_R^E) = \varrho(\mathbf{U}^E), \quad (4.39)$$

a possible choice is

$$\varrho^*(\mathbf{U}_L^E, \mathbf{U}_R^E) = \frac{1}{2}(\varrho_L^E + \varrho_R^E).$$

It is then clear that the simple approximate Riemann solver (4.32) in Eulerian coordinates is consistent with the integral form of (4.24) in the sense of Definition 4.1.

*Remark 4.2.* The CFL condition obviously changes when we switch from Lagrangian to Eulerian coordinates. However the validity of (4.36) for the proposed approximate Riemann solver is valid for any  $\Delta t$  and  $\Delta m$ . Then, the equivalence between (4.34) and (4.36) is actually sufficient to prove the consistency of the simple approximate Riemann solver (4.34) with the integral form of (4.22).  $\square$

Similarly, concerning the entropy inequality we easily prove that the following two properties are equivalent, namely,

$$\theta^L(\mathbf{U}_R^L) - \theta^L(\mathbf{U}_L^L) - \Delta m \tilde{\sigma}^L(\mathbf{U}_L^L, \mathbf{U}_R^L; \Delta m, \Delta t) \leq \sum_{k=1}^l c_k (\eta^L(\mathbf{U}_{k+1}^L) - \eta^L(\mathbf{U}_k^L)) \quad (4.40)$$

and

$$\theta^E(\mathbf{U}_R^E) - \theta^E(\mathbf{U}_L^E) - \Delta x \tilde{\sigma}^E(\mathbf{U}_L^E, \mathbf{U}_R^E; \Delta x, \Delta t) \leq \sum_{k=1}^l a_k (\eta^E(\mathbf{U}_{k+1}^E) - \eta^E(\mathbf{U}_k^E)) \quad (4.41)$$

where we have set again  $\Delta x = \frac{\Delta m}{\varrho^*(\mathbf{U}_L^E, \mathbf{U}_R^E)}$ , and

$$\tilde{\sigma}^E(\mathbf{U}_L, \mathbf{U}_R; \Delta x, \Delta t) = \varrho^*(\mathbf{U}_L^E, \mathbf{U}_R^E) \tilde{\sigma}^L(\mathbf{V}_L, \mathbf{V}_R; \Delta m, \Delta t). \quad (4.42)$$

In other words, we have proved the following expected result, assuming the consistency condition (4.39).

*Proposition 4.4*

*The simple approximate Riemann solver (4.32) in Eulerian coordinates is consistent with the integral form of the entropy inequality (4.24) if and only if the simple approximate Riemann solver (4.30) is consistent with the integral form of (4.28).*

Note that this result relies on the validity of the Rankine-Hugoniot relations (4.31).

#### 4.4 The Example of the Gas Dynamics Equations with Gravity and Friction

We focus in this section on a representative example to which we can apply the correspondence between the Lagrangian and Eulerian frames, namely, the gas dynamics equations with friction and gravity terms seen before, in Example 1.6, (1.13) for the Eulerian frame or (1.14) in Lagrangian coordinates. It is an interesting example of a system with a source term which is not geometric, and has both physical equilibria with a simple analytic expression, and a specific asymptotic behavior when the friction becomes large.

Since the computations are easier, we will first derive the simple scheme in the Lagrangian frame. For convenience, we recall the PDE model here, which in Lagrangian coordinates writes

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m p = g - \alpha \varphi(u), \\ \partial_t e + \partial_m (pu) = gu - \alpha \psi(u) \end{cases} \quad (4.43)$$

and which can be written in the form (4.1) with the choice

$$\mathbf{u} = \begin{pmatrix} \tau \\ u \\ e \end{pmatrix}, \quad \mathbf{f}(\mathbf{u}) = \begin{pmatrix} -u \\ p \\ pu \end{pmatrix}, \quad \mathbf{s}(\mathbf{u}) = \begin{pmatrix} 0 \\ g - \alpha \varphi(u) \\ gu - \alpha \psi(u) \end{pmatrix}. \quad (4.44)$$

We will mainly consider the commonly used friction terms

$$\begin{cases} \varphi(u) = |u|^\chi u, \\ \psi(u) = |u|^{\chi+2}, \quad 0 \leq a \leq 1, \end{cases}$$

with  $\chi \geq 0$ ,  $\chi = 0$  for a linear friction or  $\chi = 1$  for a quadratic friction term. Moreover, for simplicity, we take  $a = 1$  so that in both cases,  $\psi(u) = \varphi(u)u$ .

Denoting by  $\eta$  the mathematical specific entropy (satisfying the second law of thermodynamics (see Chap. III, Sect. 1.1; we usually note it  $-s$ )

$$-Td\eta = d\varepsilon + pd\tau$$

we check that smooth solutions of (4.43) satisfy

$$\partial_t \eta = -\frac{\alpha}{T}(u\varphi(u) - \psi(u)), \quad (4.45)$$

and the right-hand side vanishes under the above assumption  $\psi(u) = \varphi(u)u$ .

The eigenvalues of system (4.43) are those of Euler system in Lagrangian coordinates (Chap. II, Sect. 2.2):  $\mu_1 = -c/\tau, \mu_2 = 0, \mu_3 = c/\tau$ , where  $\frac{c}{\tau} = -\sqrt{\partial_\tau p(\tau, s)}$ .

#### 4.4.1 Derivation of a Simple Approximate Riemann Solver in Lagrangian Coordinates

We propose a simple approximate Riemann solver of the form (4.4), for system (4.43) (with friction terms satisfying  $\psi(u) = u\varphi(u)$ ). The solver is made of three waves ( $l = 3$ ) and two intermediate states  $\mathbf{u}_L^*$  and  $\mathbf{u}_R^*$ . Because of the specificities of the exact eigenvalues  $\mu_i$ , it is natural to take  $c_2 = 0$ , whereas the two extreme waves propagate with symmetric speeds  $c_1 = -C$  and  $c_3 = C$ , which gives

$$\tilde{\mathbf{w}}\left(\frac{m}{t}; \mathbf{u}_L, \mathbf{u}_R\right) = \begin{cases} \mathbf{u}_L, & \frac{m}{t} < -C, \\ \mathbf{u}_L^*, & -C < \frac{m}{t} < 0, \\ \mathbf{u}_R^*, & 0 < \frac{m}{t} < C, \\ \mathbf{u}_R, & \frac{m}{t} > C. \end{cases} \quad (4.46)$$

The parameter  $C$  is an approximation of the exact Lagrangian sound speed  $c/\tau$  associated with the acoustic waves. We will see further below that  $C$  has to be taken large enough with respect to the sound speed. In order to define the intermediate states  $\mathbf{u}_L^*$  and  $\mathbf{u}_R^*$ , we first write that (4.46) should be consistent with the integral form of (4.43). By Definition 4.1, we look for a function  $\tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \xi, \tau)$  such that

$$\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L) - \Delta x \tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \Delta m, \Delta t) = -C(\mathbf{u}_L^* - \mathbf{u}_L) + C(\mathbf{u}_R - \mathbf{u}_R^*) \quad (4.47)$$

with

$$(\mathbf{u}_L, \mathbf{u}_R, \Delta m, \Delta t) \rightarrow (\mathbf{u}, \mathbf{u}, 0, 0) \tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \Delta m, \Delta t;) = \mathbf{s}(\mathbf{u}).$$

We naturally choose  $\tilde{\mathbf{s}}$  in the form of the exact source term taken on some state velocity

$$\tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \Delta m, \Delta t) = \begin{pmatrix} 0 \\ g - \alpha\varphi(\tilde{u}) \\ g\tilde{u} - \alpha\psi(\tilde{u}) \end{pmatrix} \quad (4.48)$$

with  $\tilde{u} = \tilde{u}(\mathbf{u}_L, \mathbf{u}_R; \Delta m, \Delta t)$  to be precised later on. Then, the consistency relations (4.47) give the three relations

$$\begin{cases} -\Delta u = C(\tau_L - \tau_L^* + \tau_R - \tau_R^*), \\ \Delta p - \Delta m(g - \alpha\varphi(\tilde{u})) = C(u_L - u_L^* + u_R - u_R^*), \\ \Delta(pu) - \Delta m(g\tilde{u} - \alpha\psi(\tilde{u})) = C(e_L - e_L^* + e_R - e_R^*), \end{cases} \quad (4.49)$$

where we have used the classical notation  $\Delta\varphi = \varphi_R - \varphi_L$  for each quantity  $\varphi$  related to the fluid.

Let us now turn to the definition of the intermediate states  $\mathbf{u}_L^*$  and  $\mathbf{u}_R^*$ . First, we note that the mass conservation equation in (4.43) does not contain any source term. So it is natural to impose the usual Rankine-Hugoniot jump relation associated with this first equation across each of the three waves of (4.46). We easily get

$$\begin{cases} u_L^* - C\tau_L^* = u_L - C\tau_L, \\ u_R^* = u_L^*, \\ u_R + C\tau_R = u_R^* + C\tau_R^*. \end{cases} \quad (4.50)$$

Then, the first relation in (4.49) becomes automatically satisfied, whereas the second one writes

$$\Delta p - \Delta m(g - \alpha\varphi(\tilde{u})) = 2C\left(\frac{1}{2}(u_L + u_R) - u^*\right),$$

where we have set  $u^* = u_L^* = u_R^*$ . This relation, together with the first and third ones in (4.50), provides the following formulas for  $u^*$ ,  $\tau_L^*$ , and  $\tau_R^*$ :

$$u^* = \frac{1}{2}(u_L + u_R) - \frac{1}{2C}\Delta P, \quad \text{where } \Delta P = \Delta p - \Delta m(g - \alpha\varphi(\tilde{u})), \quad (4.51)$$

$$\begin{cases} \tau_L^* = \tau_L + \frac{1}{2C}\Delta u - \frac{1}{2C^2}\Delta P, \\ \tau_R^* = \tau_R + \frac{1}{2C}\Delta u + \frac{1}{2C^2}\Delta P. \end{cases} \quad (4.52)$$

At this stage, assuming  $\tilde{u}$  has been chosen,  $u^* = u_L^* = u_R^*$  and  $\tau_L^*$ ,  $\tau_R^*$  are uniquely defined and depend on  $\Delta P$  (which depends itself on  $\tilde{u}$ ), and the first two consistency relations in (4.49) are satisfied. In order to complete the definition of the intermediate states, it remains to specify  $e_L^*$  and  $e_R^*$ . These two quantities are related by a single compatibility relation, namely, the third equation in (4.49), so that one has one more degree of freedom in addition to  $\tilde{u}$ . To fix the other degree of freedom, a particular form for the numerical flux  $\tilde{\mathbf{g}}(\mathbf{u}_L, \mathbf{u}_R)$  is proposed. From (4.12) we have

$$\tilde{\mathbf{g}}(\mathbf{u}_L, \mathbf{u}_R) = \frac{1}{2} \{ \mathbf{f}(\mathbf{u}_L) + \mathbf{f}(\mathbf{u}_R) - C(\mathbf{u}_L^* - \mathbf{u}_L + \mathbf{u}_R - \mathbf{u}_R^*) \}. \quad (4.53)$$

Using (4.51) and (4.52), the first component of  $\tilde{\mathbf{g}}$  is given by

$$\frac{1}{2} \{-u_L - u_R - C(\tau_L^* - \tau_L + \tau_R - \tau_R^*)\} = \frac{1}{2}(-2\frac{1}{2}(u_L + u_R) + \frac{1}{C}\Delta P) = -u^*,$$

the second by

$$\frac{1}{2} \{p_L + p_R - C(u^* - u_L + u_R - u^*)\} = p^*$$

with

$$p^* = \frac{1}{2}(p_L + p_R) - \frac{C}{2}\Delta u, \quad (4.54)$$

and the third one writes

$$\frac{1}{2} \{p_L u_L + p_R u_R - C(e_L^* - e_L + e_R - e_R^*)\}.$$

Since the third component of the flux  $\mathbf{f}(\mathbf{u})$  and the source term  $\mathbf{s}(\mathbf{u})$  are both obtained by multiplying the second components by  $u$ , we impose a similar relation at the numerical level, namely,

$$\frac{1}{2} \{p_L u_L + p_R u_R - C(e_L^* - e_L + e_R - e_R^*)\} = p^* u^*.$$

Combining this relation with the third consistency relation in (4.49) easily gives  $e_L^* - e_L$  and  $e_R^* - e_R$ . More precisely

$$\begin{cases} e_L^* - e_L = \frac{1}{C} \left\{ p_L u_L - p^* u^* + \frac{\Delta m}{2} (g\tilde{u} - \alpha\psi(\tilde{u})) \right\}, \\ e_R^* - e_R = \frac{1}{C} \left\{ p^* u^* - p_R u_R + \frac{\Delta m}{2} (g\tilde{u} - \alpha\psi(\tilde{u})) \right\}, \end{cases} \quad (4.55)$$

which completes the determination of the intermediate states  $\mathbf{u}_L^*$  and  $\mathbf{u}_R^*$ , up to the choice of  $\tilde{u}$ . To conclude, note that the proposed approximate Riemann solver, with  $\tilde{\mathbf{s}}$  of the form (4.48), is consistent with the integral form of (4.44) provided that  $\tilde{u}$  is such that

$$\lim_{(\mathbf{u}_L, \mathbf{u}_R, \Delta m, \Delta t) \rightarrow (\mathbf{u}, \mathbf{u}, 0, 0)} \tilde{u}(\mathbf{u}_L, \mathbf{u}_R; \Delta m, \Delta t) = u, \quad (4.56)$$

if  $u$  denotes the velocity of the state  $\mathbf{u}$ .

Since we are interested in the long-time behavior of the solution, we may be inclined to take for  $\tilde{u}$  the value  $u^*$  of the velocity inside the fan  $-C < \frac{x}{t} < C$ . We will come back on this choice later on. Thus assume  $\tilde{u} = u^*$  satisfies (4.51), which means

$$u^* + \frac{\alpha \Delta m}{2C} \varphi(u^*) = \frac{1}{2}(u_L + u_R) - \frac{1}{2C}(\Delta p - g\Delta m). \quad (4.57)$$

Since  $u \mapsto u + \frac{\alpha \Delta m}{2C} \varphi(u)$  is a strictly increasing function from  $\mathbb{R}$  onto  $\mathbb{R}$ , (4.57) admits a unique solution  $u^* = u^*(\mathbf{u}_L, \mathbf{u}_R; \Delta m, \alpha)$ .

We note that, by continuity arguments,  $u^*$  does satisfy the consistency condition (4.56). When  $\varphi$  is linear, we have an explicit formula, noting

$$K_\alpha = (1 + \frac{\alpha \Delta m}{2C})^{-1}, \quad (4.58)$$

it writes

$$u^* = K_\alpha \left( \frac{1}{2}(u_L + u_R) - \frac{1}{2C}(\Delta p - g \Delta m) \right).$$

Let us now give the formulas of the resulting scheme. Setting

$$u_{j+\frac{1}{2}} = u^*(\mathbf{u}_j, \mathbf{u}_{j+1}; \Delta m), \quad p_{j+\frac{1}{2}} = p^*(\mathbf{u}_j, \mathbf{u}_{j+1}),$$

the numerical scheme writes

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \frac{\Delta t}{\Delta m} (u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n), \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta m} (p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n) + \Delta t \left( g - \frac{\alpha}{2} (\varphi(u_{j-\frac{1}{2}}^n) + \varphi(u_{j+\frac{1}{2}}^n)) \right), \\ e_j^{n+1} = e_j^n - \frac{\Delta t}{\Delta m} (p_{j+\frac{1}{2}}^n u_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n u_{j-\frac{1}{2}}^n) \\ \quad + \frac{\Delta t}{2} \left( g(u_{j-\frac{1}{2}}^n + u_{j+\frac{1}{2}}^n) - \alpha (\psi(u_{j-\frac{1}{2}}^n) + \psi(u_{j+\frac{1}{2}}^n)) \right) \end{cases} \quad (4.59)$$

with, skipping the time superscripts  $n$  from (4.54),

$$p_{j+\frac{1}{2}} = \frac{1}{2} (p_j + p_{j+1} - C(u_{j+1} - u_j)), \quad (4.60)$$

and in the linear friction case, using the notation (4.58),

$$u_{j+\frac{1}{2}} = K_\alpha \left( \frac{1}{2}(u_j + u_{j+1}) - \frac{1}{2C}(\Delta p - g \Delta m) \right). \quad (4.61)$$

Note that in the absence of source term,  $\alpha = g = 0$ , we recover the scheme of Chap. IV, Sect. 4.5.2 (see (4.87, 4.88)), which was proved to be positively conservative and entropy satisfying (Chap. IV, Theorem 4.3).

#### 4.4.2 Derivation of a Simple Approximate Riemann Solver in Eulerian Coordinates

Following Sect. 4.3, we immediately derive a simple approximate Riemann solver for the gas dynamics equations with friction and gravity terms in Eulerian coordinates which we recall writes

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + p) = \varrho(g - \alpha\varphi(u)), \\ \partial_t(\varrho e) + \partial_x((\varrho e + p)u) = \varrho(gu - \alpha\psi(u)). \end{cases} \quad (4.62)$$

Following (4.30), the corresponding simple solver derived from (4.46) writes

$$\widetilde{\mathbf{W}}\left(\frac{x}{t}; \mathbf{U}_L, \mathbf{U}_R\right) = \begin{cases} \mathbf{U}_L, & \frac{x}{t} < a_1, \\ \mathbf{U}_L^*, & \lambda_1 < \frac{x}{t} < a_2, \\ \mathbf{U}_R^*, & \lambda_2 < \frac{x}{t} < a_3, \\ \mathbf{U}_R, & \frac{x}{t} > a_3 \end{cases} \quad (4.63)$$

with the relations coming from (4.29, 4.31)

$$\begin{cases} a_1 = u_L - C\tau_L = u^* - C\tau_L^*, \\ a_2 = u^*, \\ a_3 = u_R + C\tau_R = u^* + C\tau_R^*. \end{cases}$$

We recall that  $u^*$ ,  $\tau_L^*$ , and  $\tau_R^*$  satisfy (4.50)–(4.51) with here  $\Delta m = \varrho^*(\mathbf{U}_L, \mathbf{U}_R)$   $\Delta x$ , while  $e_L^*$  and  $e_R^*$  are given by (4.55) with  $p^*$  defined by (4.54). Setting

$$\tilde{\mathbf{s}}(\mathbf{U}_L, \mathbf{U}_R; \Delta x, \Delta t) = \varrho^*(\mathbf{U}_L, \mathbf{U}_R) \begin{pmatrix} 0 \\ g - \alpha\varphi(\tilde{u}) \\ g\tilde{u} - \alpha\psi(\tilde{u}) \end{pmatrix},$$

and assuming (4.39), the above simple Riemann solver is consistent with the integral form of (4.62).

#### 4.4.3 Equilibria of the Continuous System and Well-Balanced Property

For the Euler system (4.62), stationary solutions are characterized by

$$\begin{cases} \frac{d}{dx}(\varrho u) = 0, \\ \frac{d}{dx}(\varrho u^2 + p) = \varrho(g - \alpha\varphi(u)), \\ \frac{d}{dx}((\varrho e + p)u) = \varrho(gu - \alpha\psi(u)). \end{cases}$$

Hence  $\varrho u$  is constant. If we focus on solutions at rest (with  $u = 0$ ), solving the system

$$\begin{cases} u = 0, \\ \frac{dp}{dx} = \varrho g, \end{cases}$$

provides a classical equilibrium stationary solution of (4.62) where the gravity term balances the pressure gradient (the friction term is no longer present since friction acts only for a moving flow).

One says that a scheme is *well-balanced* if it preserves the corresponding *discrete equilibria*, which are defined according to Definitions 4.3 or 4.5.

*Remark 4.3.* In particular, stationary solutions are interesting in the context of atmosphere modeling (the Euler system with gravity but neglecting friction) for which one can exhibit many such solutions. Indeed, we may compute some analytical expression if we assume that  $p$  is given by a  $\gamma$  law. For an air column at rest, assume that the data may be known at the ground level, for instance, given  $(\varrho_0, \varepsilon_0)$  ( $\varepsilon_0$  a given internal energy or equivalently a given temperature  $T_0$ ). Then among the stationary solutions, one finds  $\varepsilon = \varepsilon_0$ ,  $\varrho = \varrho_0 e^{-gx/(\gamma-1)\varepsilon_0}$ . In practice, the physical atmosphere equilibrium corresponds to a known (tabulated) pressure  $p(x)$ , and then one solves the relation for the density.

This stationary solution is to be preserved in a computation. As in Remark 1.2, the definition of a discrete equilibrium could be given independently of the scheme: it should satisfy naturally  $u_j = 0$  and some consistent discretization of the second equation, for instance,  $(\frac{\Delta p}{\Delta x})_j = \varrho_j g$ . Then a scheme is well-balanced if, starting from a discrete initial data which satisfies  $u_j^0 = 0$ ,  $(\frac{\Delta p}{\Delta x})_j^0 = \varrho_j^0 g$ , at each time  $t_n = n\Delta t$ , the same relation holds:  $u_j^n = 0$ ,  $(\frac{\Delta p}{\Delta x})_j^n = \varrho_j^n g$ .  $\square$

We now focus on stationary solutions satisfying moreover  $u = \bar{u}$  constant. If  $\bar{u} \neq 0$ , we must have  $\varrho = \bar{\varrho}$  constant and

$$\begin{cases} \frac{dp}{dx} = \bar{\varrho}(g - \alpha\varphi(\bar{u})), \\ \bar{\varrho}u \frac{d\varepsilon}{dx} + \bar{u} \frac{dp}{dx} = \bar{\varrho}(g\bar{u} - \alpha\psi(\bar{u})), \end{cases}$$

and thus

$$\begin{cases} \frac{dp}{dx} = \bar{\varrho}(g - \alpha\varphi(\bar{u})), \\ \bar{u} \frac{d\varepsilon}{dx} = \alpha(\bar{u}\varphi(\bar{u}) - \psi(\bar{u})). \end{cases}$$

Since  $\psi(u) = \varphi(u)u$ , we get  $\frac{d\varepsilon}{dx} = 0$  and  $\varepsilon = \bar{\varepsilon}$ ,

$$p = p(\bar{\varrho}, \bar{\varepsilon}) = \bar{p}.$$

We obtain a constant stationary solution of (4.62), where the gravity term balances the friction one, given by

$$\begin{cases} \varrho = \bar{\varrho}, \quad \varepsilon = \bar{\varepsilon}, \\ \bar{u} = \varphi^{-1}(g/\alpha). \end{cases}$$

*Remark 4.4.* Stationary solutions for the Lagrangian formulation (4.43) satisfy

$$\begin{cases} u = \bar{u}, \\ \frac{dp}{dm} = g - \alpha\varphi(\bar{u}). \end{cases}$$

They are in correspondence with transport wave solutions in the Euler formulation, i.e.,  $u = \bar{u}$  constant, and  $\varrho, \varepsilon$  functions of  $x - \bar{u}t$ . For such a transport

wave solution, all convection terms of the form  $D_t \cdot = \partial_t \cdot + \bar{u} \partial_x \cdot$  vanish. The only remaining term gives  $\partial_x p = \varrho(g - \alpha\varphi(\bar{u}))$ . When  $\bar{u} = 0$ , we recover the steady solution. The constant (Eulerian) state with  $\bar{u} \neq 0$  and  $\partial_x p = 0$  gives in Lagrangian formulation a special case with  $p = \bar{p}$  constant.  $\square$

We can now study the well-balanced property of the Godunov-type scheme associated with the simple Riemann solver (4.46) which we have derived above (in the Lagrangian frame), using the result of Proposition 4.3. We note two given neighboring cell states by  $\mathbf{u}_j^n = \mathbf{u}_L$ ,  $\mathbf{u}_{j+1}^n = \mathbf{u}_R$ .

*Lemma 4.1*

Consider a discrete equilibrium solution satisfying (3.5)–(4.14), i.e.,

$$\frac{1}{\Delta m}(\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)) = \tilde{\mathbf{s}}(\mathbf{u}_L, \mathbf{u}_R; \Delta m, \Delta t).$$

Then, for the approximate Riemann solver (4.63), the intermediate states satisfy

$$\mathbf{u}_L^* = \mathbf{u}_L, \quad \mathbf{u}_R^* = \mathbf{u}_R.$$

Thus the discrete equilibrium is a strong equilibrium in the sense of Definition 4.5. Moreover, the velocity is constant  $\tilde{u} = u^* = u_L = u_R$ .

*Proof.* With our present notations, the discrete equilibrium (3.5)–(4.14) writes

$$\begin{cases} \Delta u = 0, \\ \Delta p - \Delta m(g - \alpha\varphi(\tilde{u})) = 0, \\ \Delta(pu) - \Delta m(g\tilde{u} - \alpha\psi(\tilde{u})) = 0. \end{cases} \quad (4.64)$$

The first two equalities give

$$\begin{cases} u_L = u_R \equiv \bar{u}, \\ \tilde{u} = \varphi^{-1}(\frac{1}{\alpha}(-\frac{\Delta p}{\Delta m} + g)), \end{cases}$$

together with formulas (4.51), (4.52), and (4.54), and we get

$$\begin{cases} u^* = \frac{1}{2}(u_L + u_R) = \bar{u}, \\ \tau_L^* = \tau_L, \quad \tau_R^* = \tau_R, \\ p^* = \frac{1}{2}(p_L + p_R). \end{cases}$$

The third condition (4.64) then gives  $\bar{u}\Delta p - \Delta m(g\tilde{u} - \alpha\psi(\tilde{u})) = 0$  so that

$$\begin{aligned} p_L u_L - p^* u^* + \frac{\Delta m}{2}(g\tilde{u} - \alpha\psi(\tilde{u})) &= p^* u^* - p_R u_R + \frac{\Delta m}{2}(g\tilde{u} - \alpha\psi(\tilde{u})) \\ &= -\frac{1}{2}(\bar{u}\Delta p - \Delta m(g\tilde{u} - \alpha\psi(\tilde{u}))) = 0, \end{aligned}$$

and from (4.55), we get:  $e_L^* = e_L$ ,  $e_R^* = e_R$ . Then, the discrete equilibrium condition implies

$$\mathbf{u}_L^* = \mathbf{u}_L, \quad \mathbf{u}_R^* = \mathbf{u}_R,$$

which gives the result.  $\square$

We have thus also proved:

*Proposition 4.5*

The Godunov-type scheme associated with the consistent and simple approximate Riemann solver defined by (4.59), with (4.51), (4.52), (4.55), is a strongly well-balanced scheme for system (4.43), in the sense of Definition 4.6.

Let us precise the discrete stationary solutions computed by the scheme, and check the consistency with the above formulas in Remark 4.4. These discrete stationary solutions are naturally such that  $u = \bar{u}$  is constant. The second and the third equations of (4.64) successively give

$$\bar{u} \frac{\Delta p}{\Delta m} = \bar{u}(g - \alpha\varphi(\tilde{u})) = \tilde{u}(g - \alpha\varphi(\tilde{u})),$$

which yields either  $\tilde{u} = \bar{u}$  or  $\alpha\varphi(\tilde{u}) = g$ . Assume we are in the second case, it occurs for the particular choice  $\tilde{u} = \varphi^{-1}(\frac{g}{\alpha})$ , and it implies either  $\bar{u} = 0$  or  $\frac{\Delta p}{\Delta m} = 0$ , in which case  $p_L = p_R$  is constant. Then we can have  $\bar{u} = \varphi^{-1}(\frac{g}{\alpha})$  too and  $\tilde{u} = \bar{u}$ . Is it possible that  $\tilde{u} \neq \bar{u}$ ? The consistency condition (4.56) requires  $\tilde{u} \rightarrow \bar{u}$ , as  $\Delta m, \Delta t \rightarrow 0$ , so that  $\bar{u}$  should be equal to  $\varphi^{-1}(\frac{g}{\alpha})$ , and this is naturally coherent when the pressure is constant, so that  $\tilde{u} \neq \bar{u}$  is indeed inconsistent. In all other cases, we have  $\tilde{u} = \bar{u}$ , and the consistency condition (4.56) is automatically satisfied.

Hence, the discrete stationary solutions computed by the scheme are such that  $u = \bar{u}$  is constant and  $p$  is piecewise constant with  $\frac{\Delta p}{\Delta m} = g - \alpha\varphi(\bar{u})$ , which is consistent with formulas in Remark 4.4.

*Remark 4.5.* The results of Proposition 4.5 involve the definition of a discrete equilibrium, as given in Lemma 4.1, and this definition in turn is linked to the choice of the source term in the form  $\mathbf{s}(\tilde{u})$  as given by (4.48), which, in the present context, is natural. Let us just remark that in other situations, one might think that the source term is approximated up to some  $\mathcal{O}(\Delta x)$  term, for instance, setting  $\tilde{\mathbf{S}}(\mathbf{U}_L, \mathbf{U}_R; \Delta m, \Delta t) = \mathbf{s}(\tilde{u}) + \Delta x$  would still lead to a consistent scheme, but then the precise definition of a discrete equilibrium should also change and be defined taking into account this  $\mathcal{O}(\Delta x)$  term (if one wants it to be exactly preserved). This is to illustrate that the precise definition of discrete equilibria may be scheme dependent. Also, the definition of a well-balanced scheme may be extended to take into account some small (up to some order of  $\Delta x$ ) terms around a stable equilibrium. However, note that the choice of  $\mathcal{O}(\Delta x)$  terms depending on the parameter  $\alpha$  might reveal itself particularly inadequate for deriving asymptotic preserving properties.  $\square$

The well-balanced property is obtained as soon as  $\tilde{u}$  satisfies  $\tilde{u} = u$  once  $u_L = u_R \equiv u$ , which holds for  $\tilde{u} = \frac{1}{2}(u_L + u_R)$  or for any usual average of the two values. The asymptotic preserving property of the scheme (in the weak sense of consistency with the limit system) requires a specific choice for  $\tilde{u}$ , and we will take  $\tilde{u} = u^*$ , as already done in (4.57).

The well-balanced property of the scheme in Eulerian coordinates associated to (4.46) is studied in a very similar way.

#### *Proposition 4.6*

*The Godunov-type scheme associated with the consistent and simple approximate Riemann solver defined by (4.63), with (4.51), (4.52), (4.55), is a strongly well-balanced scheme for system (4.62), in the sense of Definition 4.6.*

Instead of detailing the computations needed for the proof, we turn to another interpretation of the scheme, as a relaxation scheme, thus involving a relaxation system, which explains its good properties with a different point of view.

## **4.5 Link with Relaxation Schemes**

In many occasions, a simple Riemann solver for a given system can be obtained as an exact Godunov solver for a larger, but simpler, relaxation system which approximates the former one considered as the *equilibrium* system, as seen in Chap. IV, Sect. 8. This may also be the case for a system with a specific source term, such as gravity and friction for Euler, once the source terms have been transformed.

This approach has some formal link with the treatment of geometric source terms, since the transformation aims at incorporating the source in the differential part, with nonconservative products, so as to perform a better treatment of the various features associated with the balance laws, such as the preservation of equilibria. It leads to write a larger system of PDE, which keeps the same eigenvalues of the original one, and adds a linearly degenerate field. The preservation of equilibria is obtained using Riemann invariants associated to the new characteristic field, the linear degeneracy being again an important property, both from theoretical and technical points of views. Indeed, on the one hand, one could not think of adding nonlinear waves, and on the other hand, the Riemann invariants of the new LD field are easily computed.

The fact that a simple scheme may be seen as an exact Godunov solver on an approximate system is interesting because, once the relaxation approximation is justified, it explains in a quite straightforward fashion the stability properties which are inherited from the continuous relaxation system.

Since it is difficult to treat this topic in a very general frame, we come back to the case of Euler system with gravity and friction. We emphasize the fact that the present approach is developed only for theoretical purposes, in order to understand and explain the good behavior of the scheme. At the algorithmic level, the resulting scheme writes with the variables involved in the original Euler system only.

#### 4.5.1 Euler System with Gravity and Friction: Homogeneous Formulation

We introduce another formulation of (4.62) in order to derive a well-balanced scheme which happens to coincide with the simple Godunov-type scheme introduced above. The idea is to transform the source term in a differential one. In the present case, there is no geometric term, and one cannot introduce an additional stationary wave; however, following [237], we may introduce a potential  $q$  defined by

$$\begin{cases} \partial_x q = \varrho \\ \partial_t q = -\varrho u. \end{cases} \quad (4.65)$$

The existence of  $q$  comes from Poincaré's theorem and the compatibility conditions given by the conservation of mass. The potential satisfies a new PDE  $\partial_t q + u \partial_x q = 0$ . Then (4.62) writes equivalently

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho u) + \partial_x((\varrho u^2 + p) - (g - \alpha \varphi(u)) \partial_x q) = 0, \\ \partial_t(\varrho e) + \partial_x((\varrho e + p)u) - (gu - \alpha \psi(u)) \partial_x q = 0, \\ \partial_t \varrho q + \partial_x(\varrho u q) = 0. \end{cases} \quad (4.66)$$

We have replaced the transport equation for  $q$  by a conservative one. It holds naturally for smooth solutions and will be justified below in case of a discontinuity.

*Remark 4.6.* Note that if we neglect friction and consider only the effect of gravity, i.e., we make  $\alpha = 0$ , the potential, say  $q_h$ , is rather defined as  $\partial_x q_h = \varrho g$  (which is the usual hydrostatic pressure), then  $\partial_t q_h = -\varrho g u$ . In that case, it is natural to define a new pressure  $p_h = p - q_h$  and a new total energy  $\varrho \mathcal{E} = \varrho e + q_h$  (see [237, 770]). The formulation is then conservative, while in the presence of friction, the nonconservative terms remain.  $\square$

The new system can be written, with the set of variables  $\mathbf{v} = (\varrho, \varrho u, \varrho e, q)^T$ , in nonconservative form (2.3)

$$\partial_t \mathbf{v} + A(\mathbf{v}) \partial_x \mathbf{v} = 0$$

with a  $4 \times 4$  matrix  $A(\mathbf{v})$  augmented from the Jacobian matrix for Euler system (see (4.67) below).

*Lemma 4.2*

System (4.66) has four real eigenvalues  $u - c, u, u, u + c$ , where  $c$  is the usual sound speed ( $c^2 = \frac{\partial p}{\partial \varrho}(\varrho, s)$ ), and it has a basis of eigenvectors if and only if

$$\psi(u) = u\varphi(u),$$

or for the state  $u = 0$ . The first and last characteristic fields associated with  $u \pm c$  are GNL, while the characteristic field associated with  $u$  is LD.

*Proof.* We develop (4.66) in quasilinear form, and with the choice of variable  $\mathbf{v} = (\varrho, \varrho u, \varrho e, q)^T$ , the Jacobian matrix writes

$$\mathbf{A}(\mathbf{v}) = \begin{pmatrix} A^{Euler} & B_q \\ 0 & u \end{pmatrix} \quad (4.67)$$

where we have denoted by  $A^{Euler}$  the usual Jacobian  $3 \times 3$  matrix of Euler system (in conservative variables  $\mathbf{u} = (\varrho, \varrho u, \varrho e)^T$ ) and  $B_q$  is the column vector  $B_q = (0, -(g - \alpha\varphi(u)), -(gu - \alpha\psi(u)))^T$ . Hence, the eigenvalues are real and coincide with those of the Euler system  $u - c, u, u + c$  (where  $c$  denotes the Eulerian sound speed) and the corresponding eigenvectors may be taken as  $\mathbf{R}(\mathbf{v}) = (\mathbf{r}(\mathbf{u}), 0)$ , where we have noted by  $\mathbf{r}$  an eigenvector for the Euler system. For the fourth eigenvector which is associated with the double eigenvalue  $u$ , in order to have a basis of  $\mathbb{R}^4$ , we need to take it in the form  $\mathbf{R} = (\mathbf{r}, 1)$ , and a simple computation shows that this is possible only if  $\psi(u) = u\varphi(u)$  or  $u = 0$ .  $\square$

Because of the source term, the Riemann invariant for the eigenvalue  $u$  is no longer  $p$  but the quantity  $p - q(g - \alpha u)$ , while  $q$  is a new Riemann invariant for the other GNL fields; moreover  $q$  is discontinuous only across the contact discontinuity. More precisely, we can state

*Lemma 4.3*

The quantity  $w(\mathbf{v}) = p - q(g - \alpha\varphi(u))$  is a Riemann invariant associated with  $u$ , i.e.,  $\nabla_{\mathbf{v}} w \cdot \mathbf{R} = 0$  for any eigenvector  $\mathbf{R}$  associated with  $u$ . The potential  $q$  is a Riemann invariant for the 1- and 4-characteristic fields.

*Proof.* For what concerns the potential  $q$ , the result is straightforward because the 1- and 4-eigenvectors are of the form  $\mathbf{R}_i = (\mathbf{r}_i, 0)$ . This results from the equation  $\partial_t q + u\partial_x q = 0$ , which says that  $q$  is transported at the velocity  $u$  given by the flow. This property of  $q$  is sometimes referred to as:  $q$  is a *strong Riemann invariant* for  $u$ .

For the characteristic field  $u$ , one of the two eigenvectors may be taken in the form  $\mathbf{R}^E = (\mathbf{r}^E, 0)$ , and since  $\nabla_{\mathbf{v}} p = (\nabla_{\mathbf{u}} p, 0)$ , we have  $\nabla_{\mathbf{v}} p \cdot \mathbf{R}^E = 0$  because  $p$  is a  $u$  Riemann invariant for the Euler system. Then we compute  $\nabla_{\mathbf{v}}(q(g - \alpha\varphi(u))) = (\alpha\varphi'(u)uq/\varrho, -\alpha\varphi'(u)q/\varrho, 0, g - \alpha\varphi(u))$ , and if  $\mathbf{r}^E = (r_1^E, r_2^E, r_3^E)^T$ , it yields

$$\nabla_{\mathbf{v}} w \cdot \mathbf{R}^E = \nabla_{\mathbf{v}} (-q(g - \alpha\varphi(u))) \cdot \mathbf{R}^E = -\alpha\varphi'(u)q(ur_1^E - r_2^E)/\varrho.$$

But any eigenvector of  $\mathbf{A}(\mathbf{v})$  associated with  $u$  satisfies  $r_2 = ur_1$  as results by identifying the first component of each side of the equality  $\mathbf{A}\mathbf{R} = u\mathbf{R}$ , from which we deduce that  $\nabla_{\mathbf{v}} w \cdot \mathbf{R}^E = 0$ .

Now, for the other eigenvector  $\mathbf{R} = (\mathbf{r}, 1)$ , writing that  $\mathbf{R}$  is an eigenvector of  $\mathbf{A}$ , i.e.,  $\mathbf{A}\mathbf{R} = u\mathbf{R}$ , and taking the second component of both sides of the equality, while expressing the components of  $\mathbf{A}$ , gives a relation

$$\nabla_{\mathbf{u}}(p + \varrho u^2) \cdot \mathbf{r} - (g - \alpha\varphi(u)) = ur_2,$$

if we note  $\mathbf{r} = (r_1, r_2, r_3)^T$ . First, we have  $\nabla_{\mathbf{u}}(\varrho u^2) = (-u^2, 2u, 0)^T$  and thus  $\nabla_{\mathbf{u}}(\varrho u^2) \cdot \mathbf{r} = -u^2 r_1 + 2ur_2$  which gives

$$\nabla_{\mathbf{v}} p \cdot \mathbf{R} = \nabla_{\mathbf{u}} p \cdot \mathbf{r} = -ur_2 + u^2 r_1 + g - \alpha\varphi(u).$$

Since again  $r_2 = ur_1$ , we deduce that  $\nabla_{\mathbf{v}} p \cdot \mathbf{R} = g - \alpha\varphi(u)$ . Then we compute  $\nabla_{\mathbf{v}}((g - \alpha\varphi(u))q) = (\alpha\varphi'(u)uq/\varrho, -\alpha\varphi'(u)q/\varrho, 0, g - \alpha\varphi(u))$ , and it gives

$$\nabla_{\mathbf{v}}(q(g - \alpha\varphi(u))) \cdot \mathbf{R} = \alpha\varphi'(u)(ur_1 - r_2)q/\varrho + g - \alpha\varphi(u) = g - \alpha\varphi(u),$$

which yields  $\nabla_{\mathbf{v}} w \cdot \mathbf{R} = 0$ . □

One consequence is that when  $q$  is discontinuous, which may happen only across a contact discontinuity of speed  $u$ , the nonconservative product  $(g - \alpha\varphi(u))\partial_x q$  is well defined; in that case the pressure  $p$  is no longer constant, as it is for the usual Euler system without source, and  $[p] = (g - \alpha\varphi(u))[q]$ .

Note now that if  $\eta$  is the mathematical specific entropy for Euler system (which we usefully note  $-s$ ), smooth solutions of (4.62) satisfy the following equality (see (4.45))

$$-T(\partial_t \varrho\eta + \partial_x(\varrho\eta u)) = \alpha\varrho(u\varphi(u) - \psi(u)),$$

and since we have assumed  $\psi = u\varphi$ , it becomes a conservation law

$$\partial_t(\varrho\eta) + \partial_x(\varrho\eta u) = 0. \quad (4.68)$$

Thus, system (4.66) is hyperbolic. Moreover, while smooth solutions of system (4.62) satisfy the system

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + p) = \varrho(g - \alpha\varphi(u)), \\ \partial_t(\varrho\eta) + \partial_x(\varrho\eta u) = 0, \end{cases} \quad (4.69)$$

(with  $p = \tilde{p}(\tau, \eta)$ ,  $\tau = 1/\varrho$ , which means we express the pressure as a function of the density or specific volume and of the specific entropy), smooth solutions of (4.66) satisfy

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + p) - (g - \alpha\varphi(u))\partial_x q = 0 \\ \partial_t(\varrho\eta) + \partial_x(\varrho\eta u) = 0 \\ \partial_t(\varrho q) + \partial_x(\varrho u q) = 0, \end{cases} \quad (4.70)$$

(with again  $p = \tilde{p}(\tau, \eta), \tau = 1/\varrho$ , which is equivalent to (4.66) (for smooth solutions)).

We can prove:

*Lemma 4.4*

Let  $(\varrho, u, \eta, q)(x, t)$  be a smooth solution of (4.70) associated with a given initial data  $(\varrho_0, u_0, \eta_0, q_0)(x)$  such that  $q_0$  satisfies  $\frac{dq_0}{dx}(x) = \varrho_0(x)$ . Then  $q(x, t)$  satisfies  $\partial_x q(x, t) = \varrho(x, t)$ , and  $(\varrho, u, \eta)(x, t)$  is a smooth solution of system (4.69) associated with  $(\varrho_0, u_0, \eta_0)(x)$ .

*Proof.* Let  $(\varrho, u, \eta)(x, t)$  be the smooth solution of (4.70) associated with the data  $(\varrho_0, u_0, \eta_0)(x)$  (this solution exists at least for  $t$  small enough). Define, associated with this solution, the function  $\bar{q}(x, t)$  by  $\partial_x \bar{q} = \varrho, \partial_t \bar{q} = -\varrho u$ . Then  $(\varrho, u, \eta, \bar{q})(x, t)$  is the smooth solution of (4.69) associated with  $(\varrho_0, u_0, \eta_0, q_0)(x)$ , hence  $q = \bar{q}$ .  $\square$

Hence, we can use the formulation with potential which is equivalent for smooth solutions. Now for discontinuous solutions, the Rankine-Hugoniot relations for (4.70) give that when a discontinuity propagates with speed  $\sigma$ :

- either  $\sigma \neq u$ , we have a shock and then  $[q] = 0$ ,  $q$  is continuous, and we have the same jump relations as for Euler (4.62),
- or  $\sigma = u$ , we have a contact discontinuity,  $[p] = (g - \alpha\varphi(u))[q]$  (equivalently,  $w = p - (g - \alpha\varphi(u))q$  is a Riemann invariant). If  $[q] = 0$ , we recover  $[p] = 0$  as for Euler.

Thus, the formulation (4.70) introduces possible discontinuities of  $q$  propagating with velocity  $u$ , but no new discontinuities for (4.62) and shocks propagate at the right speed. Given two constant states  $(\varrho_i, \varrho_i u_i, \varrho_i e_i, q_i), i = L, R$ , close enough, we can solve the Riemann problem for (4.70) following the same steps used to solve the Riemann problem for Euler.

Stationary states for (4.70) with null velocity satisfy  $\partial_x(p - gq) = 0$ . For what concerns the well-balanced property, the benefit of the formulation with potential is that the quantity  $p - q(g - \alpha u)$  being a Riemann invariant, it will be naturally preserved in the exact Riemann solver, which ensures the preservation of the corresponding equilibria at the discrete level.

Since the Godunov scheme for the nonlinear system is still expansive, one can think of introducing an approximate Riemann solver through a relaxation scheme. We will present the relaxation scheme in the next section.

*Remark 4.7.* In Lagrangian coordinates, the equation  $\partial_t(\varrho q) + \partial_x(\varrho u q) = 0$  writes, setting  $\tilde{q}(m, t) = q(x, t)$  where  $m$  is the mass variable,  $\partial_t \tilde{q} = 0$ , which is added to (4.43). The analog of (4.70) writes

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m p = (g - \alpha \varphi(u)), \\ \partial_t e + \partial_m (pu) = (gu - \alpha \psi(u)), \\ \partial_t \tilde{q} = 0. \end{cases}$$

Since  $\partial_x q = \varrho$ , we may take  $\tilde{q}(m) = m$ ,  $\partial_m \tilde{q} = 1$ . Thus, introducing  $q$  corresponds exactly to add a standing wave in Lagrangian coordinates, to take into account a source term which now writes, for the momentum equation,  $(g - \alpha \varphi(u))\partial_m q$ .

In the Godunov scheme,  $q$  is discretized by a piecewise constant. The exact Riemann solver involves two intermediate constant states, separated by a contact discontinuity at  $x = 0$ . The  $\lambda_0$ -Riemann invariants are  $u$  and  $w = p - (g - \alpha \varphi(u))$ . Then, across  $x = 0$ , we have  $u_- = u_+ \equiv u^*$ , and  $w_- = w_+$ , which reads  $p_- - (g - \alpha \varphi(u^*))q_- = p_+ - (g - \alpha \varphi(u^*))q_+$ . It gives  $(g - \alpha \varphi(u^*))\Delta q = p_+ - p_-$ . We will not explicit the exact Riemann solver, but in the relaxation scheme constructed below, where the exact Godunov solver is used for an approximating LD system, this relation is indeed satisfied (see (4.76) with  $p_+ = \Pi_R^*$ ,  $p_- = \Pi_L^*$ ). See also Remark 4.9.

This is very close to the approach seen in Sect. 3.2.2, an approach which was followed in order to design well-balanced schemes for a system with geometric source term.  $\square$

*Remark 4.8.* Following Remark 4.6, when  $\alpha = 0$ , the potential is defined by  $\partial_x q_h = \varrho g$ ,  $\partial_t q_h = -\varrho u g$ . The system (4.70) becomes conservative with a pressure term  $p_h = p - q_h$  and provided the total energy is taken as  $\varrho e_h = \varrho e + q_h$ , it writes

$$\begin{cases} \partial_t \varrho + \partial_x (\varrho u) = 0, \\ \partial_t (\varrho u) + \partial_x (\varrho u^2 + p_h) = 0, \\ \partial_t (\varrho e_h) + \partial_x ((\varrho e_h + p_h)u) = 0, \\ \partial_t \varrho q_h + \partial_x (\varrho u q_h) = 0. \end{cases} \quad (4.71)$$

Note that the energy flux is unchanged  $\varrho e_h + p_h = \varrho e + p$ . For this system (4.71),  $p_h$ , which plays the role of the classical pressure for the Euler system, is naturally continuous at a material discontinuity, which is coherent with what we found in the above case with friction if we set  $\alpha = 0$ .  $\square$

#### 4.5.2 Relaxation Scheme for Euler System with Gravity and Friction

A relaxation scheme for system (4.70) can be easily constructed, following the lines of Chap. IV, Sect. 8.3.1. Indeed, as previously noted in Remark 4.7,

in a Lagrangian frame, the transport equation for the potential  $q$ ,  $\partial_t(\varrho q) + \partial_x(\varrho u q) = 0$ , becomes simply  $\partial_t q = 0$ ; hence, it is straightforward to extend the relaxation system, and there is no need to introduce relaxation for this equation, and thus, coming back to the Eulerian frame, the relaxation system for (4.70) writes

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) &= 0, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + \Pi) - (g - \alpha\varphi(u))\partial_x q &= 0, \\ \partial_t \varrho \eta + \partial_x(\varrho \eta u) &= 0, \\ \partial_t(\varrho \mathcal{T}) + \partial_x(\varrho \mathcal{T} u) &= \nu \varrho(\tau - \mathcal{T}), \\ \partial_t(\varrho q) + \partial_x(\varrho u q) &= 0, \end{cases} \quad (4.72)$$

with

$$\Pi = \tilde{\Pi}(\tau, \eta, \mathcal{T}) \equiv \tilde{p}(\mathcal{T}, \eta) + \bar{C}^2(\mathcal{T} - \tau). \quad (4.73)$$

Here  $\bar{C}$  is a positive constant which plays the role of a frozen Lagrangian sound speed and is required to bound the exact sound speed for Euler system (if we note the latter one  $C^\mathcal{L}$ , then  $(C^\mathcal{L})^2 = -\partial_\tau \tilde{p}(\tau, \eta)$ ),

$$\bar{C}^2 > -\partial_\tau \tilde{p}(\tau, \eta), \quad (4.74)$$

for all the values  $\tau, \eta$  under consideration, (4.74) is known as the Whitham (or subcharacteristic) condition. The five characteristic fields are linearly degenerate. The solution of the Riemann problem is easily obtained.

*Proposition 4.7*

Given two constant states  $\mathbf{U}_L, \mathbf{U}_R$ , the solution  $W_q(x/t; \mathbf{U}_L, \mathbf{U}_R)$  of the Riemann problem for (4.72) consists of (at most) three contact discontinuities, each propagating with a characteristic speed (resp.  $u_L - \bar{C}\tau_L, u^*, u_R + \bar{C}\tau_R$ ), separating  $\mathbf{U}_L$ , two intermediate states  $\mathbf{U}_L^*, \mathbf{U}_R^*$  and  $\mathbf{U}_R$ . The states  $\mathbf{U}_L^*, \mathbf{U}_R^*$  are respectively characterized by  $(u^*, \Pi_L^*, \mathcal{T}_L, \tau_L^*, q_L)$  and  $(u^*, \Pi_R^*, \mathcal{T}_R, \tau_R^*, q_R)$  with  $u^*$  solving

$$u^* + \frac{\alpha}{2\bar{C}}\varphi(u^*)\Delta q = \frac{u_L + u_R}{2} - \frac{1}{2\bar{C}}(\Delta\Pi - g\Delta q) \quad (4.75)$$

and the other quantities satisfying

$$\begin{cases} \Pi_L^* = \frac{1}{2}(\Pi_L + \Pi_R) - \frac{\bar{C}}{2}\Delta u - \frac{1}{2}(g - \alpha\varphi(u^*))\Delta q \\ \Pi_R^* = \frac{1}{2}(\Pi_L + \Pi_R) - \frac{\bar{C}}{2}\Delta u + \frac{1}{2}(g - \alpha\varphi(u^*))\Delta q \\ \tau_L^* = \tau_L + \frac{1}{\bar{C}}(u^* - u_L), \\ \tau_R^* = \tau_R + \frac{1}{\bar{C}}(u_R - u^*). \end{cases} \quad (4.76)$$

The exact Godunov scheme for system (4.72) provides a scheme for Euler system, which is moreover entropy satisfying. The proof follows exactly that of Proposition 8.5 in Chap. IV.

The scheme is well-balanced in the sense that it preserves a (nonconstant) equilibrium for (4.69) which is discretized in a “natural” way. This is equivalent to the strongly well-balanced property of the scheme derived in Sect. 3.2 (in the sense of Definition 4.6).

*Proposition 4.8*

Let  $\mathbf{u}_0(x) = (\varrho_0, \varrho_0 u_0, \varrho_0 e_0)^T(x)$  be an equilibrium data for the Euler system (4.62), i.e., satisfying  $u_0 = 0$ ,  $d_x p_0 = \varrho_0 g$ . Assume that  $\mathbf{u}_j^0, j \in \mathbb{Z}$  is defined by

$$\mathbf{u}_j^0 = \frac{1}{\Delta x} \int_{C_j} \mathbf{u}_0(x) dx,$$

and define  $q^0 = (q_j^0)$  such that  $\forall j \in \mathbb{Z}$ ,  $\Delta q_{j+1/2}^0 = \frac{1}{g} \Delta p_{j+1/2}^0$ . Then  $\forall n > 0$ ,  $\forall j \in \mathbb{Z}$ ,  $\mathbf{u}_j^n = \mathbf{u}_j^0$ .

*Proof.* Let  $\mathbf{u}_0(x)$  be a (nonconstant) equilibrium data for the Euler system (see Sect. 4.4.3); then the “potential”  $q(x)$  satisfies  $d_x p_0 = g d_x q_0$ . This data  $\mathbf{u}_0(x)$  is discretized for the scheme by piecewise constant functions, thus  $u_i^0 = 0, \forall i \in \mathbb{Z}$ , and we have  $\frac{\Delta p_{j+1/2}^0}{\Delta x} = \frac{\varrho_j^0 + \varrho_{j+1}^0}{2} g + \mathcal{O}(\Delta x)$  (we might also start directly from a discrete equilibrium data satisfying  $u_j^0 = 0$  and  $\frac{\Delta p_{j+1/2}^0}{\Delta x} = \frac{\varrho_j^0 + \varrho_{j+1}^0}{2} g$  or any consistent discretization). We define  $\mathcal{T}_j^0 = \tau_j^0$ , thus  $\Pi_j^0 = p_j^0$ , and moreover we define  $q^0$  so as to satisfy  $\Delta q_{j+1/2}^0 = \frac{1}{g} \Delta p_{j+1/2}^0$ .

Then for all local Riemann problems involved in Godunov’s scheme, noting for simplicity by  $\mathbf{U}_L, \mathbf{U}_R$  the states  $\mathbf{U}_j^0, \mathbf{U}_{j+1}^0$  on each side of an interface  $x_{j+1/2}$  which can be located at the origin, from (4.75)(4.76) we get  $u^* = 0$ ,  $\tau_{L,R}^* = \tau_{L,R}$

$$\Pi_L^* = \frac{\Delta \Pi}{2} + \Pi_L - \frac{1}{2} g \Delta q = \Pi_L = p_L,$$

similarly  $\Pi_R^* = \Pi_R = p_R$  the solution of the Riemann problem is indeed stationary. This is valid for any cell  $C_j$ , and (using the notation of Proposition 4.7)

$$\mathbf{W}_q(x/t; \mathbf{U}_j^0, \mathbf{U}_{j+1}^0) = \begin{cases} \mathbf{U}_j^0 & x < 0, \\ \mathbf{U}_{j+1}^0 & x > 0, \end{cases} \quad (4.77)$$

thus  $\mathbf{U}_j^1 = \mathbf{U}_j^0$ , the state is at equilibrium (i.e.,  $\mathcal{T} = \tau$ ); it is a piecewise constant stationary solution for system (4.72), and this implies that  $\mathbf{u}_j^1 = \mathbf{u}_j^0$  is a discrete equilibrium for the resulting scheme, i.e., a stationary solution with  $u = 0$  and  $\frac{\Delta p}{\Delta x} = \varrho g + \mathcal{O}(\Delta x)$ .  $\square$

*Remark 4.9.* As already announced in Remark 4.7, this WB property is derived directly from the nature of the exact Godunov solver. Indeed, we

notice that (nonconstant) equilibria for system (4.72) with  $u = 0$ , thus  $\psi(0) = \varphi(0) = 0$ , satisfy  $\partial_x \Pi - g \partial_x q = 0$ . Then the (exact) solution at the evolution step must satisfy the associated Rankine-Hugoniot condition which writes  $[\Pi] = g[q]$ , which means that the solution remains stationary, and this property is still valid after projection.  $\square$

Similarly, we derive a scheme in Lagrangian coordinates which can be obtained from the change of frame Lagrange  $\leftrightarrow$  Euler performed directly on the relaxation system (4.72) or from the equivalence at the discrete level seen for simple schemes (Proposition 4.4).

We notice that (4.75) and (4.57) coincide provided we define  $\Delta q = \Delta m$ . This is natural in a Lagrangian frame, since in Eulerian coordinates we have  $\partial_x q = \varrho$ , then in Lagrangian coordinates  $\partial_m q = 1$  (as already seen in Remark 4.7).

Then it is easy to check that the scheme we obtain coincides with the simple Godunov-type scheme (4.59), derived in Sect. 4.4.1, provided we do choose the same value  $\bar{C} = C$  in (4.46, 4.73) and  $\tilde{u} = u^*$  in this last scheme. This means that, thanks to this relaxation approach with “potential” which has transformed the source term (a term that is now incorporated in the differential part), the definition of the velocity term  $\tilde{u}$  is natural: it is the velocity of the intermediate state in the solution of the Riemann problem.

We recall that the resulting scheme only writes with the variables involved in the original Euler system and the resulting algorithm does not need a definition of  $q$ .

We will see in the next section that the scheme is naturally asymptotic preserving, at least from a heuristic point of view, in the weak sense of consistency with the asymptotic system.

## 5 Stiff Source Terms, Asymptotic Preserving Numerical Schemes

### 5.1 Introduction

This section is devoted to the numerical treatment of systems with stiff source terms, for instance, of the form  $\frac{1}{\varepsilon}\mathbf{s}$ , i.e., depending on a parameter  $\varepsilon$  which may become very small. These questions belong to the wider category of singularly perturbed problems. The solution of a singularly perturbed problem, say  $P_\varepsilon$ , is supposed to converge, as the perturbation parameter  $\varepsilon$  tends to zero, toward a solution of the limit problem  $P_0$  which we assume is a well-posed problem. In practice, one is interested in the approximate computation of  $P_\varepsilon$  for a certain range of values of the parameter  $\varepsilon$ . When it occurs that the singular limit leads to a change in the type of the equation, for instance, from hyperbolic to parabolic, it may create difficulties when trying to solve

$P_\varepsilon$  for  $\varepsilon$  small, requiring very restrictive conditions on the time step, all the most if the stability constants depend on  $\varepsilon$ . This motivates the definition of asymptotic preserving (AP for short) schemes. If  $P_\varepsilon$  is approximated by some discrete scheme, say  $P_{\varepsilon,\Delta}$ , where  $\Delta$  represents some discretization parameter, an AP scheme is such that  $\lim P_{\varepsilon,\Delta}$  as  $\varepsilon \rightarrow 0$  gives a consistent approximation, say  $P_{0,\Delta}$  of  $P_0$ ; in other words, the diagram approximation ( $\Delta \rightarrow 0$ )—perturbation ( $\varepsilon \rightarrow 0$ ) commutes and

$$\lim_{\Delta \rightarrow 0} \lim_{\varepsilon \rightarrow 0} P_{\varepsilon,\Delta} = \lim_{\varepsilon \rightarrow 0} \lim_{\Delta \rightarrow 0} P_{\varepsilon,\Delta}.$$

This is rather schematic; the limit may involve some stability criterion; then one requires that this stability condition does not depend on the parameter  $\varepsilon$  (or at least the dependence is “tolerable”).

We have already considered such a situation when studying the low Mach regime for gas dynamics, at the end of Chap. V, even if the abstract notion of AP property is alike or if “all-Mach schemes” were needed for that situation; the approach was specific, and we will not come again over this example, focusing on stiff source terms.

Now, we have already encountered stiff source terms. Stiff-relaxation source terms are naturally involved in the context of relaxation approximation, for which we refer to Chap. IV, Sect. 8.1. Relaxation is linked to a small *relaxation parameter*, meant to tend to zero; the formal zero limit of the *relaxation system* is called the *equilibrium system*. In this case, both systems, relaxation and equilibrium, are assumed to be hyperbolic, and the waves propagate with a finite speed; moreover a stability condition requires that the speeds are interlaced (as seen in the above mentioned Sect. 8.1, Chap. IV). In the precise context of relaxation schemes, the relaxation is instantaneous, and there is no particular need of studying the AP property, and a splitting method with instantaneous relaxation is convenient.

From a more physical point of view, relaxation terms are essential to model phenomena involving distinct physical “time scales,” in particular in the modeling of two-phase flows, where the non-equilibrium may be related to a difference of pressure, velocity, or chemical potential between the phases (see, for example, [1217]). In some contexts, the relaxation is nearly instantaneous; in other situations, the difference must be taken into account, and then one may want to treat both situations, meaning find a numerical model (a scheme) which should be able to restore the adequate behavior in both cases of non-equilibrium ( $\varepsilon$  not too small) and equilibrium  $\varepsilon \rightarrow 0$ . In these situations, a splitting method, where the convection stage and the source stage are treated separately, may be inaccurate [935].

The design of schemes which handle accurately both stiff and nonstiff source terms continues to receive a lot of interest; early attempts are found in [114, 116, 133, 221].

Now, in some other cases, one is moreover interested in the large time behavior of the solution. Often, the scaling, wrt. a small parameter, affects not

only the source but also the independent time variable, and the nature of the limit model may change, becoming diffusive, with roughly speaking infinite speed of propagation (see Chap. IV, Remark 8.4). In this case a scheme, which is consistent with a given system involving a small parameter, say  $\varepsilon$ , is said to be asymptotic preserving if, performing the same scaling at the discrete level as at the continuous one, it is consistent with the limit model as  $\varepsilon \rightarrow 0$ , under relevant stability conditions (for a given mesh size,  $\Delta t$  should not go to 0 with  $\varepsilon$ ). A natural remark is that the CFL stability condition for an explicit scheme for the limit equation, for which the numerical speed is of a greater order of magnitude, does impose a more restrictive limitation on the time step; an implicit treatment is then needed, at least a partially implicit one, meaning concerning the waves whose velocity becomes infinite (many works try to split the continuous model in two parts, separating the fast and slow waves, see, for instance, [269, 270]).

On some examples of systems with source terms of simple relaxation type  $\frac{1}{\varepsilon}\mathbf{s}(\mathbf{u})$ , depending in a simple way on the parameter  $\varepsilon$ , we will be able to derive AP schemes. Again, this does not concern the low Mach regime.

## 5.2 Some Simple Examples

### 5.2.1 Stiffness, Multiscale Resolution

A central issue in numerical simulation of a stiff relaxation system is how well the discretization resolves the small relaxation scale. First, one must precise this notion of stiffness and relaxation scale which we illustrate on a representative example.

We come back to the simple example encountered in Chap. IV, Remark 8.4, of the  $p$ -system with a source term

$$\begin{cases} \partial_t v - \partial_x u = 0 \\ \partial_t u + \partial_x p(v) = -\alpha u, \end{cases} \quad (5.1)$$

where  $\alpha > 0$  represents a friction coefficient. In the context of relaxation,  $\varepsilon = \frac{1}{\alpha}$  corresponds to a relaxation time. A critical nondimensional parameter is the ratio of the distance that a typical sound wave of the homogeneous system travels (with speed  $a = \sqrt{-p'(v)}$ ) over a time  $\varepsilon$  to a typical reference length  $L$ :  $\epsilon = a\varepsilon/L$ . When this ratio is such that  $\epsilon \ll 1$ , the relaxation term is called *stiff*.

The spatial resolution of a given discretization is measured by the smallness of the nondimensional parameter  $\delta = \Delta x/L$ . Then one measures the spatial resolution of the relaxation scale by the ratio  $Pe = \delta/\epsilon = \Delta x/a\varepsilon$  often called the *cell Peclet number*. Note that in the study of transport phenomena in fluid flows, the Peclet number is defined to be the ratio of the rate of advection of

a physical quantity by the flow to the rate of diffusion of the same quantity driven by an appropriate gradient (if this number is much smaller than 1, diffusion dominates over advection).

A grid will be called coarse (relative to the relaxation scale) if  $Pe \gg 1$ , then the discretization is under resolved, and the computation cannot reproduce the small-scale phenomena; the grid is fine if  $Pe \ll 1$ .

An analogous issue is addressed in the context of kinetic transport equations, for which a diffusive scaling may lead to diffusion equations in the hydrodynamic limit. The parameter  $\varepsilon$  represents the mean free path; it varies from  $\mathcal{O}(1)$  values in the rarefied regime to very small values in the hydrodynamic (diffusive) limit.

A possible way of addressing the multiscale computational issue is a domain decomposition method; one solves the coarse model ( $\varepsilon$  small) and the fine model ( $\varepsilon = \mathcal{O}(1)$ ) in respective coarse and fine domains; the models are coupled at the interface (see Example 6.3 below); then one may think of a model adaptation procedure [228]. This is a very rough description of a method which is also receiving much attention. On the contrary, in the following subsections, one is interested in developing numerical schemes for approximating transport equations for which the diffusive limit (as  $\varepsilon \rightarrow 0$ ) is also important to solve numerically, but one uses only one scheme on the whole domain. The asymptotic regime requires to solve small scales, which may be very expensive if the scheme is not *asymptotic preserving* (AP): the AP property refers to an efficient behavior of a scheme when the relaxation time  $\varepsilon$  goes to 0. However, the situation is a little more complex since in the context of large friction, one is also interested in the large time behavior, so that we also have to precise the setting and the classical scaling factor which is often introduced.

### 5.2.2 A Linear Model

When the function  $p$  is moreover linear, (5.1) provides a simple model, often called the telegraph equations or the Goldstein-Taylor model, depending on the context. The system was also obtained as a discrete kinetic formulation (with two velocities) in Chap. IV, Sect. 7.4.2. The model has been studied by many authors ([666] and references therein) including Jin-Levermore [663], Hsiao-Liu [634], and Gosse-Toscani [556, 557]. Moreover, when one is interested in the long-time behavior, this model linear system usually writes directly with a scaling factor  $\varepsilon = \frac{1}{\alpha}$  in front of the time derivative terms

$$\begin{cases} \varepsilon \partial_t v + \partial_x u = 0 \\ \varepsilon \partial_t u + a^2 \partial_x v = -\frac{1}{\varepsilon} u, \end{cases} \quad (5.2)$$

where  $a$  is the characteristic wave speed. Setting  $u = \varepsilon w + \mathcal{O}(\varepsilon^2)$ , neglecting terms of second order in  $\varepsilon$ , the asymptotic behavior is then governed by the

heat equation

$$\partial_t v - a^2 \partial_{xx}^2 v = 0, \quad \text{and } w = -a^2 \partial_x v.$$

This example already illustrates some of the difficulties which a numerical scheme encounters if one wants to simulate both the low ( $\varepsilon \sim 1$ ) and high ( $\varepsilon \ll 1$ ) damping regimes. Since the limit system (as  $\varepsilon \rightarrow 0$ ) is parabolic, if one considers *explicit* schemes, one expects in the first case a CFL condition for the hyperbolic system ( $a\Delta t \leq \frac{1}{2}\Delta x$ ), while in the limit as  $\varepsilon \rightarrow 0$ , a parabolic kind stability condition  $a^2\Delta t \leq \frac{1}{2}\Delta x^2$  is to be required. Then, an asymptotic preserving scheme should involve a stability condition satisfying those requirements: it may depend on  $\varepsilon$  but  $\Delta t$  should not go to 0 as  $\varepsilon \rightarrow 0$ .

### 5.2.3 Some Related Simple Models

Let us give some more examples and references of physical situations where a very similar model can be found, since these situations led to the development of AP schemes.

System (5.2) is found as a two-velocity Boltzmann (one dimensional) kinetic model which describes the evolution of the velocity distribution of a gas composed of two kinds of particles, moving with constant and equal speeds  $c$  along an axis, either in the positive direction with a density  $f_1 \equiv u$  or in the negative direction with a density  $f_2 \equiv v$ , and a collision operator in the form  $k(u, v)(v - u)$ . Moreover, in Carleman's model of the Boltzmann equation, one assumes binary interactions between particles, then  $k = u + v$ , and the model writes in the diffusive scaling

$$\begin{cases} \varepsilon \partial_t u - c \partial_x u = \frac{1}{\varepsilon}(v^2 - u^2) \\ \varepsilon \partial_t v + c \partial_x v = -\frac{1}{\varepsilon}(v^2 - u^2), \end{cases}$$

where  $\varepsilon$  stands for the mean free path. Then setting  $\varrho = u + v$  for the (macroscopic) density,  $j = c(v - u)$  for the flux of particles, it gives

$$\begin{cases} \varepsilon \partial_t \varrho + c \partial_x j = 0, \\ \varepsilon \partial_t j + c^2 \partial_x \varrho = -\frac{2}{\varepsilon} \varrho j. \end{cases}$$

When  $\varepsilon$  goes to zero, the hydrodynamic limit in the diffusive scaling corresponds to  $\partial_t \varrho - \frac{1}{4} \partial_x (\frac{1}{\varrho} (\partial_x \varrho)) = 0$ , and  $j = j_\varepsilon \rightarrow 0$ . Then, it is possible to define instead  $j_1 = \frac{c(v-u)}{\varepsilon}$  and write

$$\begin{cases} \partial_t \varrho + c \partial_x j_1 = 0, \\ \varepsilon^2 \partial_t j_1 + c^2 \partial_x \varrho = -2 \varrho j_1, \end{cases}$$

and as  $\varepsilon \rightarrow 0$ , one gets  $j_1 = j_{1\varepsilon} \rightarrow -\frac{1}{2\varrho} \partial_x \varrho = -\frac{1}{2} \partial_x (\log \varrho)$  and of course the same equation for  $\varrho$ , also called the nonlinear porous medium equation.

Theoretical results can be found in [815]; see also [659] for other collision terms and diffusive limit and AP schemes and also [557, 666, 891].

The system is also linked to the  $M_1$  moment model of radiative transfer, and in this context, the simplest model writes

$$\begin{cases} \partial_t E_r + \partial_x F_r = 0, \\ \partial_t F_r + c^2 \partial_x P_r = -c\sigma F_r, \end{cases}$$

where  $E_r \geq 0$  is the energy of radiation,  $F_r$  the radiation flux,  $P_r$  the radiative transfer pressure,  $c$  the speed of light, and  $\sigma$  the absorption coefficient which may become large, in which case it is associated with a scaling parameter. The system is obtained by taking the first two moments of the full radiative transfer equation. The pressure is such that  $\frac{P_r}{E_r} = \chi(f)$ , where  $\chi$ , called the Eddington factor, is an explicit function of  $f = \frac{F_r}{cE_r}$ , namely,  $\chi(s) = \frac{3+4s^2}{5+2\sqrt{4-3s^2}}$ , and  $f$  measures the anisotropy of radiation; for physical reason it satisfies  $|f| \leq 1$ . The modeling of strong interaction (in opaque regions) leads to write a scaled system

$$\begin{cases} \varepsilon \partial_t E + \partial_x F = 0 \\ \varepsilon \partial_t F + c^2 \partial_x P = -\frac{c\sigma}{\varepsilon} F \end{cases}$$

and as  $\varepsilon \rightarrow 0$ , we obtain  $F_r \rightarrow 0$  and a diffusion limit  $\partial_t E_r - \partial_x (\frac{c}{3\sigma} \partial_x E_r) = 0$ ; see [134, 144, 210, 211, 556]. In this context it may be important to have a scheme able to simulate the radiative transfer both in the opaque and transparent regions.

Let us also mention some convergence result [1115] and for a splitting method approximating the same model (in an abstract framework) with more general stiff source term [482] after [481]; see also [261].

Different approaches have been followed in the above references in order to derive AP schemes. We cannot describe nor treat in detail all the specific schemes, so we will focus below on some simple facts, more or less far from the physical context, in order to understand a few basic ideas.

#### 5.2.4 A Nonlinear Model

Coming back to system (5.1), which models a compressible flow with friction (in Lagrangian coordinates), there are theoretical results concerning the asymptotic behavior of its solutions ([634, 635, 1045]), [637] and using the relative entropy [743, 1044, 1146] and references therein). It has been proved that a solution of (5.1) can be described time asymptotically by a solution of the problem

$$\begin{cases} \partial_t v - \frac{1}{\alpha} \partial_{xx}^2 p(v) = 0, \\ u = -\frac{1}{\alpha} \partial_x(p(v)), \end{cases} \quad (5.3)$$

(the porous media equation) in the sense that the difference (in  $\mathbf{L}^2$  and  $\mathbf{L}^\infty$  norm) tends to 0 with a rate  $t^{\frac{1}{2}}$  (we do not give the precise assumptions on initial data and refer to the above mentioned reference and to [637] and references therein).

Note that the formula for  $u$  corresponds exactly to the first term of the Chapman-Enskog expansion wrt.  $\frac{1}{\alpha}$  (neglecting terms of order 2) and the equation for  $v$  to the intermediate system with diffusive term seen in Chap. IV, Sect. 8.2.1 (see (8.8) and Remark 8.4).

It seems natural to perform the following scaling

$$t = \alpha s, \quad w = \alpha u$$

in (5.1); then let  $\alpha \rightarrow \infty$  (this is a formal analysis). The long-time behavior for large friction of solutions of the hyperbolic system (5.1) is described by the equation

$$\begin{cases} \partial_s v - \partial_{xx}^2(p(v)) = 0, \\ w = -\partial_x p(v), \end{cases}$$

which gives a parabolic-type asymptotic behavior, with the velocity given by the second equation which is known as Darcy's law.

### 5.3 Derivation of an AP Scheme for the Linear Model

The definition of the asymptotic preserving property is something rather delicate which needs some care. We must be cautious and precise whether we consider a given scheme as approximating the solution of (5.2) where some scaling has been directly included or the solution of a system before scaling, as (5.1) or in the linear case (5.4) below.

#### 5.3.1 Scheme for the System Before and After Scaling

First we focus on a system of the form (5.1) in the linear case and consider the simple well-balanced scheme for (5.1) which we have constructed in the above Sect. 4.4.1 for the full Euler system with energy. Passing to the Lagrangian frame (where we however note the mass variable  $x$ ), restricting to the simple barotropic case, with moreover a linear "pressure" function, we get the system

$$\begin{cases} \partial_t v + \partial_x u = 0 \\ \partial_t u + a^2 \partial_x v = -\alpha u. \end{cases} \quad (5.4)$$

We have seen that, after the scaling  $t \mapsto t/\alpha$ ,  $u \mapsto \alpha u$ , the expected asymptotic behavior as  $\alpha \rightarrow \infty$  is the heat equation for  $v$  (see (5.2)).

Let us be given initial conditions  $(v_0, u_0)(x)$  for (5.4) and define as usual  $v_j^0 = \frac{1}{\Delta x} \int_{C_j} v_0(x) dx$ , similarly for  $u_j^0$ . It is easy to see that (from (4.59)) the scheme writes for  $n \geq 0$

$$\begin{cases} v_j^{n+1} = v_j^n - \frac{\Delta t}{\Delta x} (u_{j+\frac{1}{2}}^n - u_{j-\frac{1}{2}}^n), \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} (p_{j+\frac{1}{2}}^n - p_{j-\frac{1}{2}}^n) - \Delta t \frac{\alpha}{2} (u_{j-\frac{1}{2}}^n + u_{j+\frac{1}{2}}^n), \end{cases} \quad (5.5)$$

with  $p(v) = a^2 v$  and the natural choice  $C = a$ . Relations (4.60)–(4.61) lead to the following formula for a scheme for (5.4)

$$p_{j+\frac{1}{2}} = \frac{a^2}{2} (v_j + v_{j+1}) - \frac{a}{2} (u_{j+1} - u_j) \quad (5.6)$$

and also

$$u_{j+\frac{1}{2}} = K_\alpha \left( \frac{1}{2} (u_j + u_{j+1}) - \frac{a}{2} (v_{j+1} - v_j) \right). \quad (5.7)$$

Thus (5.5) writes

$$\begin{cases} v_j^{n+1} = v_j^n - K_\alpha \frac{\Delta t}{2\Delta x} (u_{j+1}^n - u_{j-1}^n - a(v_{j+1}^n - 2v_j^n + v_{j-1}^n)), \\ u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left( \frac{a^2}{2} (v_{j+1}^n - v_{j-1}^n) - \frac{a}{2} (u_{j+1}^n - 2u_j^n + u_{j-1}^n) \right) \\ \quad - \frac{\alpha}{2} K_\alpha \Delta t \left( \frac{1}{2} (u_{j+1}^n + 2u_j^n + u_{j-1}^n) - \frac{a}{2} (v_{j+1}^n - v_{j-1}^n) \right), \end{cases} \quad (5.8)$$

where  $K_\alpha$  is defined by (4.58)

$$K_\alpha = (1 + \frac{\alpha \Delta x}{2a})^{-1}, \quad (5.9)$$

which yields

$$\alpha K_\alpha = \alpha (1 + \frac{\alpha \Delta x}{2a})^{-1} \rightarrow \frac{2a}{\Delta x}, \text{ as } \alpha \rightarrow \infty.$$

It is easy to check that scheme (5.8) writes equivalently

$$\begin{cases} v_j^{n+1} = v_j^n - K_\alpha \frac{\Delta t}{2\Delta x} (u_{j+1}^n - u_{j-1}^n - a(v_{j+1}^n - 2v_j^n + v_{j-1}^n)), \\ u_j^{n+1} = u_j^n - K_\alpha \frac{\Delta t}{\Delta x} (p_{j+1/2} - p_{j-1/2}) - \alpha K_\alpha \Delta t u_j^n, \end{cases} \quad (5.10)$$

with the definition (5.6) of  $p_{j+1/2}$ .

Let us perform the same scaling as for the continuous system and set

$$\Delta t = \alpha \Delta s$$

in (5.10), we get

$$\begin{cases} v_j^{n+1} = v_j^n - \alpha K_\alpha \frac{\Delta s}{2\Delta x} (u_{j+1}^n - u_{j-1}^n - a(v_{j+1}^n - 2v_j^n + v_{j-1}^n)), \\ u_j^{n+1} = u_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta x} (p_{j+1/2}^n - p_{j-1/2}^n), -\alpha^2 K_\alpha \Delta s u_j^n, \end{cases} \quad (5.11)$$

again with the definition (5.6) of  $p_{j+1/2}$ .

If we now set  $w = \alpha u$ , then (5.11) gives

$$\begin{cases} v_j^{n+1} = v_j^n - \alpha K_\alpha \frac{\Delta s}{2\Delta x} \left( \frac{1}{\alpha} (w_{j+1}^n - w_{j-1}^n) - a(v_{j+1}^n - 2v_j^n + v_{j-1}^n) \right), \\ \frac{1}{\alpha} w_j^{n+1} = \frac{1}{\alpha} w_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta x} (p_{j+1/2}^n - p_{j-1/2}^n), -\alpha K_\alpha \Delta s w_j^n, \end{cases} \quad (5.12)$$

with

$$p_{j+\frac{1}{2}} = \frac{a^2}{2} (v_j + v_{j+1}) - \frac{a}{2\alpha} (w_{j+1} - w_j).$$

As  $\alpha \rightarrow \infty$ , we obtain (formally, if  $\Delta s$  is kept constant wrt.  $\alpha$ ) that (5.12) gives the limit scheme

$$\begin{cases} v_j^{n+1} = v_j^n + a^2 \frac{\Delta s}{\Delta x^2} (v_{j+1}^n - 2v_j^n + v_{j-1}^n), \\ w_j^n = -\frac{1}{\Delta x} a^2 (v_{j+1}^n - v_{j-1}^n). \end{cases} \quad (5.13)$$

We need to precise the consistency and convergence properties hidden in the term *limit scheme*.

We first recall that in (5.8),  $v_j^n$  (resp.  $u_j^n$ ) has been defined for the relaxation scheme, or equivalently the simple Godunov-type scheme, as a consistent approximation of  $v(x_j, t_n)$  (resp.  $u(x_j, t_n)$ ), with  $t_n = n\Delta t$ , where  $(v, u)$  satisfy (5.4) (at least when the solution is smooth, otherwise it is a cell-average). In the formula (5.11),  $v_j^n$  has not been changed; thus, it still denotes an approximation of  $v(x_j, t_n) = v(x_j, n\Delta t) = v(x_j, n\alpha\Delta s)$ , similarly for the second component  $u_j^n$ .

Let us introduce some notations emphasizing the dependence of the different solutions, continuous and discrete, on  $\alpha$ . We first define for any function  $\varphi(x, t)$

$$\bar{\varphi}(x, s) = \varphi(x, \alpha s) = \varphi(x, t).$$

Let  $(v_\alpha, u_\alpha)(x, t)$  be the solution of (5.4), for the given initial data  $(v_0, u_0)(x)$ , and define  $(\bar{v}_\alpha, \bar{u}_\alpha)(x, s) = (v_\alpha, u_\alpha)(x, t)$ .

Note  $v_{\alpha,j}^n(\Delta t)$  (respectively  $u_{\alpha,j}^n(\Delta t)$ ) the value  $v_j^n$  (resp.  $u_j^n$ ) computed by scheme (5.8). The above remark yields that  $v_{\alpha,j}^n(\Delta t)$  denotes an approximation of  $v(x_j, n\alpha\Delta s) = \bar{v}(x_j, n\Delta s)$ ; thus, we may write  $v_{\alpha,j}^n(\Delta t) = \bar{v}_{\alpha,j}^n(\Delta s)$  which defines  $\bar{v}_{\alpha,j}^n(\Delta s)$ . This notation is coherent with the fact that it approximates  $\bar{v}(x_j, n\Delta s)$ ; and we note similarly  $u_{\alpha,j}^n(\Delta t) = \bar{u}_{\alpha,j}^n(\Delta s)$ .

We then write (5.10) using these notations

$$\begin{cases} v_{\alpha,j}^{n+1}(\Delta t) = v_{\alpha,j}^n(\Delta t) - K_\alpha \frac{\Delta t}{2\Delta x} \left( u_{j+1}^n - u_{j-1}^n - a(v_{j+1}^n - 2v_j + v_{j-1}^n) \right), \\ u_{j,\alpha}^{n+1}(\Delta t) = u_{\alpha,j}^n - K_\alpha \frac{\Delta t}{\Delta x} \left( p_{\alpha,j+1/2}^n - p_{\alpha,j-1/2}^n \right) - \alpha K_\alpha \Delta t u_j^n, \end{cases}$$

where we have only emphasized some of the notations (all in the left-hand side), but the terms in the right-hand side should all bear the same index  $\alpha$  and dependence on the time step  $\Delta t$ . These formulas can thus be written using the notations  $\bar{v}, \bar{u}$

$$\begin{cases} \bar{v}_{\alpha,j}^{n+1}(\Delta s) = \bar{v}_{\alpha,j}^n(\Delta s) - \alpha K_\alpha \frac{\Delta s}{2\Delta x} \left( \bar{u}_{j+1}^n - \bar{u}_{j-1}^n - a(\bar{v}_{j+1}^n - 2\bar{v}_j + \bar{v}_{j-1}^n) \right), \\ \bar{u}_{\alpha,j}^{n+1}(\Delta s) = \bar{u}_{\alpha,j}^n - \alpha K_\alpha \frac{\Delta s}{\Delta x} \left( \bar{p}_{\alpha,j+1/2}^n - \bar{p}_{\alpha,j-1/2}^n \right) - \alpha^2 K_\alpha \Delta s \bar{u}_j^n, \end{cases} \quad (5.14)$$

where the notations in the right-hand side should now bear the same index  $\alpha$  and dependence on the time step  $\Delta s$  as in the left-hand side, and

$$\bar{p}_{\alpha,j+\frac{1}{2}} = \frac{a^2}{2} (\bar{v}_{\alpha,j} + \bar{v}_{\alpha,j+1}) - \frac{a}{2} (\bar{u}_{\alpha,j+1} - \bar{u}_{\alpha,j}).$$

Then the limit scheme as  $\alpha \rightarrow \infty$  writes (formally, if  $\Delta s$  is kept constant)

$$\begin{cases} \bar{v}_j^{n+1}(\Delta s) = \bar{v}_j^n + a^2 \frac{\Delta s}{2\Delta x^2} \left( \bar{v}_{j+1}^n - 2\bar{v}_j^n + \bar{v}_{j-1}^n \right) \\ \bar{w}_j^n(\Delta s) = -\frac{1}{\Delta x} a^2 \left( \bar{v}_{j+1}^n - \bar{v}_{j-1}^n \right). \end{cases} \quad (5.15)$$

This means, at least in a heuristic way, that for fixed  $\Delta x, \Delta s$ , assuming that the “sequence” in  $\alpha$  (we skip the rigorous notation of a sequence with integer index  $\alpha_k, k \in \mathbb{N}, \alpha_k \rightarrow \infty$  as  $k \rightarrow \infty$ ):  $(\bar{v}_{\alpha,j}^n(\Delta s))_\alpha$  has a limit as  $\alpha \rightarrow \infty$  which we note  $\bar{v}_j^n(\Delta s)$ , the limit satisfies the limit scheme (5.15).

We check easily that the scheme (5.15) is consistent with the equation  $\partial_s \bar{v} - a^2 \partial_{xx}^2 \bar{v} = 0, w = -a^2 \partial_x \bar{v}$ , if  $\bar{v}_j^n(\Delta s)$  denotes an approximation of  $\bar{v}(x_j, s_n)$ , and  $s_n = n\Delta s$ . We can thus conclude that scheme (5.8) is AP (asymptotic preserving) in the weak sense, meaning that once we have performed the same scaling, the scheme has a limit which is consistent with the limit system.

The above statement has passed over the stability condition in silence; it is now time to precise this notion.

Since we are in the linear case, we can perform an  $L^2$ -stability analysis (we follow the approach in [180]). In the following assertion, we assume  $a = 1$  for simplicity.

*Proposition 5.1*

*Scheme (5.8) is  $L^2$ -stable under the CFL condition  $\lambda = \frac{\Delta t}{\Delta x} \leq 1$ .*

*Proof.* As seen in Chap. IV, Sect. 1.2, it is easier to use the  $\mathbf{L}^2$  norm of the piecewise constant functions  $(v_{\alpha,\Delta}(x, t), u_{\alpha,\Delta}(x, t))$  associated with  $(v_j^n, u_j^n) =$

$(v_{\alpha,j}^n, u_{\alpha,j}^n)(\Delta t)$ , solution of (5.8):  $v_{\alpha,\Delta}(x, t) = v_{\alpha,j}^n(\Delta t)$  if  $x \in C_j, t \in [t_n, t_{n+1}]$ , similarly for  $u$ . Then, for  $t \in (t_n, t_{n+1})$

$$\|\mathbf{u}_\Delta(x, t)\|_{\mathbf{L}^2(\mathbb{R})}^2 = \Delta x \sum_{j \in \mathbb{Z}} (u_j^n)^2.$$

We can write (5.10) in the form: for any  $x \in \mathbb{R}, t > 0$

$$\begin{cases} v_{\alpha,\Delta}(x, t + \Delta t) = v_{\alpha,\Delta}(x, t) - K_\alpha \frac{\Delta t}{2\Delta x} \left( u_{\alpha,\Delta}(x + \Delta x, t) - u_{\alpha,\Delta}(x - \Delta x, t) \right. \\ \quad \left. - a(v_{\alpha,\Delta}(x + \Delta x, t) - 2v_{\alpha,\Delta}(x, t) + v_{\alpha,\Delta}(x - \Delta x, t)) \right), \\ u_{\alpha,\Delta}(x, t + \Delta t) = u_{\alpha,\Delta}(x, t) - K_\alpha \frac{\Delta t}{\Delta x} \left( \frac{a^2}{2} (v_{\alpha,\Delta}(x + \Delta x, t) - v_{\alpha,\Delta}(x - \Delta x, t)) \right. \\ \quad \left. - \frac{a}{2} (u_{\alpha,\Delta}(x + \Delta x, t) - 2u_{\alpha,\Delta}(x, t) + u_{\alpha,\Delta}(x - \Delta x, t)) \right) \\ \quad - \alpha K_\alpha \Delta t u_{\alpha,\Delta}(x, t). \end{cases}$$

Let us perform a Fourier transform wrt. the space variable. Setting  $\mathbf{U} = (v, u)^T$ , and noting  $\hat{\mathbf{U}}(\xi, t) = (\hat{v}(\xi, t), \hat{u}(\xi, t))^T$  the Fourier transform, defined for a function  $\varphi$  by

$$\hat{\varphi}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-ix\xi} \varphi(x) dx,$$

we get a linear relation between  $\hat{\mathbf{U}}_{\alpha,\Delta}(\xi, t + \Delta t)$  and  $\hat{\mathbf{U}}_{\alpha,\Delta}(\xi, t)$ . This relation writes  $\hat{\mathbf{U}}_{\alpha,\Delta}(\xi, t + \Delta t) = \mathbf{H} \hat{\mathbf{U}}_{\alpha,\Delta}(\xi, t)$  with an amplification matrix given by

$$\mathbf{H} = \begin{pmatrix} 1 - 2\lambda a K_\alpha \sin^2(\frac{1}{2}\xi \Delta x) & -i\lambda K_\alpha \sin(\xi \Delta x) \\ -i\lambda a^2 K_\alpha \sin(\xi \Delta x) & 1 - 2\lambda a K_\alpha \sin^2(\frac{1}{2}\xi \Delta x) - \alpha K_\alpha \Delta t \end{pmatrix} \quad (5.16)$$

(with  $\lambda = \frac{\Delta t}{\Delta x}$ ). Setting  $y = \sin^2(\frac{1}{2}\xi \Delta x)$ , this amplification matrix can be written in the form  $\mathbf{H} = \mathbf{I} - 2\lambda K_\alpha \tilde{\mathbf{M}}$ , with

$$\mathbf{M} = \begin{pmatrix} ay & i\sqrt{y(1-y)} \\ ia^2 \sqrt{y(1-y)} & ay + \frac{1}{2}\alpha \Delta x \end{pmatrix}. \quad (5.17)$$

Consider for simplicity the case  $a = 1$  so that the matrix  $\mathbf{H}$  is symmetric. Then its  $\mathbf{L}^2$  norm is equal to the spectral radius which is easily computed from that of  $\mathbf{M}$ . The eigenvalues  $X_\pm$  of  $\mathbf{M}$  are solution of

$$X^2 - 2X(y + \frac{1}{4}\alpha \Delta x) + y(1 + \frac{1}{2}\alpha \Delta x) = 0,$$

and the discriminant is  $\delta = (y - \frac{1}{2})^2 + (\frac{1}{4}\alpha \Delta x)^2 - \frac{1}{4}$ .

When the discriminant  $\delta = (y - \frac{1}{2})^2 + (\frac{1}{4}\alpha \Delta x)^2 - \frac{1}{4}$  is negative, the eigenvalues of  $\mathbf{H}$  are complex conjugate with squared modulus  $(1 - 2\lambda K_\alpha X_+)(1 - 2\lambda K_\alpha X_-) = 1 - 2\lambda K_\alpha (2y + \frac{1}{2}\alpha \Delta x) + 4\lambda^2 K_\alpha y$ . When we develop, we see that the modulus is bounded by 1 iff  $y(\lambda - 1) \leq \frac{1}{2}\alpha \Delta x$ , which is satisfied

whatever  $\alpha$  (in the present case such that  $\delta \leq 0$ , in particular for small  $\alpha$ ) and  $y \in [0, 1]$  iff  $\lambda \leq 1$ .

When the discriminant is positive, in particular for large  $\alpha$ ,  $\alpha\Delta x \geq 2$ , the eigenvalues of  $\mathbf{M}$  belong to  $[-1, 1]$  iff  $0 \leq \lambda K_\alpha X_- \leq \lambda K_\alpha X_+ \leq 1$ . The lower bound is satisfied if  $\delta \leq (y + \frac{1}{4}\alpha\Delta x)^2$ , which gives  $y(1 + \frac{1}{2}\alpha\Delta x) \geq 0$ . This relation is indeed satisfied  $\forall y \in [0, 1]$ . For the upper bound, which writes  $\lambda K_\alpha(y + \frac{1}{4}\alpha\Delta x + \sqrt{\delta}) \leq 1$  or  $\sqrt{\delta} \leq \frac{1}{\lambda K_\alpha} - y - \frac{1}{4}\alpha\Delta x$ , when we develop, it becomes an inequality for a linear function on  $y$ . If we have a condition valid for  $y = 0$  and  $y = 1$ , it will hold in the interval  $y \in (0, 1)$ . Now, for  $y = 0$ , we can compute directly the eigenvalues of  $\mathbf{H}$ , which are 1 and  $1 - \alpha K_\alpha \Delta t$ , the condition resumes to  $\lambda \leq \frac{1}{2} + \frac{1}{\alpha\Delta x}$ , which is satisfied if  $\lambda \leq 1$  in the case under consideration where  $\alpha\Delta x \geq 2$ . While for  $y = 1$ , the eigenvalues are  $1 - 2\lambda K_\alpha$  and  $1 - 2\lambda K_\alpha - \alpha K_\alpha \Delta t$ , and the condition writes  $\lambda + \frac{1}{2}\alpha\Delta t \leq 1 + \frac{1}{2}\alpha\Delta x$  which is also satisfied if  $\lambda \leq 1$ .

Then we conclude that  $\forall n \geq 0$

$$\|\mathbf{U}_{\alpha,\Delta}(\cdot, t_n)\|_{\mathbf{L}^2(\mathbb{R})} \leq \|\mathbf{U}_{\alpha,\Delta}(\cdot, 0)\|_{\mathbf{L}^2(\mathbb{R})}$$

and the right-hand side depends only on the initial condition  $(v_0, u_0)$ ; it is independent of  $\alpha$ .  $\square$

Now, the stability condition for the limit scheme (5.15) is (when  $a = 1$ )  $\frac{\Delta s}{\Delta x^2} \leq \frac{1}{2}$ . If we write

$$\frac{\Delta s}{\Delta x^2} = \frac{\Delta t}{\Delta x} \frac{1}{\alpha\Delta x} = \lambda \frac{1}{\alpha\Delta x}$$

we see that the (limit) stability condition is ensured under the assumption we have made before  $\alpha\Delta x \geq 2$  and  $\lambda \leq 1$ .

Thus, at least in a heuristic way, we may say that the scheme is AP in the sense that its stability condition does not depend on  $\alpha$ , and once we have performed the same scaling as for the continuous equation, it provides a scheme consistent with the limit equation.

Indeed, the proposition shows that the sequences (in  $\alpha$ )  $v_{\alpha,j}^n(\Delta t)$  and  $u_{\alpha,j}^n(\Delta t)$  are bounded (uniformly with respect to  $j, n$  and  $\Delta t$ ). We can extract (diagonal) subsequences which converge respectively as  $\alpha \rightarrow \infty$  toward some  $v_{\infty,j}^n$  and  $u_{\infty,j}^n$ . Now since  $v_{\alpha,j}^n(\Delta t) = \bar{v}_{\alpha,j}^n(\Delta s)$  and  $u_{\alpha,j}^n(\Delta t) = \bar{u}_{\alpha,j}^n(\Delta s)$ , we conclude that the sequences (in  $\alpha$ )  $\bar{v}_{\alpha,j}^n(\Delta s), \bar{u}_{\alpha,j}^n(\Delta s)$  converge as  $\alpha \rightarrow \infty$  toward  $v_{\infty,j}^n$  and  $u_{\infty,j}^n$ .

Formally, i.e., when  $\Delta s$  is constant, passing to the limit in (5.14), we obtain that  $\bar{v}_{\infty,j}^n$  and  $\bar{u}_{\infty,j}^n$  satisfy

$$\begin{cases} \bar{v}_{\infty,j}^{n+1}(\Delta s) = \bar{v}_{\infty,j}^n + a^2 \frac{\Delta s}{2\Delta x^2} (\bar{v}_{\infty,j+1}^n - 2\bar{v}_{\infty,j}^n + \bar{v}_{\infty,j-1}^n), \\ \bar{w}_{\infty,j}^n(\Delta s) = -\frac{1}{\Delta x} a^2 (\bar{v}_{\infty,j+1}^n - \bar{v}_{\infty,j-1}^n). \end{cases}$$

Clearly the problem lies in the assumption  $\Delta s$  is constant, since  $\Delta s$  has been defined as  $\Delta t = \alpha \Delta s$ . We have introduced a scaling wrt. time to take into account the fact that we are interested by the large time behavior of the solution. However, it is not possible to write a unique scheme which takes into account both scaled and unscaled variable. If one considers (5.4), one is authorized to use only one scheme such as (5.8) (or another scheme with unscaled time step); for fixed  $\alpha$ , and given  $\Delta t$ , the time asymptotic behavior amounts to take  $n$  large; letting  $\alpha \rightarrow \infty$  does not bring more information than taking  $\alpha \rightarrow \infty$  in (5.4). If one wants to compute a solution for both systems, before scaling and after scaling, i.e., for any  $\alpha$  and for two different time scales, we cannot write one unique scheme, with the same time step. A possibility is to consider (5.14) for large  $\alpha$  in place of the original scheme (5.8), but then one needs to take a new time step  $\Delta s$  independent of  $\alpha$ . Thus an idea could be to choose according to the size of  $\alpha$ , the resulting scheme writes:

– for  $\alpha \Delta x < 2$ , compute  $(v_{\alpha,j}^n, u_{\alpha,j}^n) = (v, u)_{\alpha,j}^n(\Delta t)$ ,  $j \in \mathbb{Z}, n > 0$ , solution of (5.8),  $\Delta t$  should satisfy the hyperbolic CFL condition  $\lambda a = \frac{\Delta t}{\Delta x} a \leq 1$ ;  $(v_{\alpha,j}^n, u_{\alpha,j}^n)$  is an approximation of  $(v_\alpha, u_\alpha)(x, n\Delta t)$ ,  $(v, u)_\alpha$  solution of (5.4): the scheme is stable and consistent with (5.4).

– for  $\alpha \Delta x > 2$ , compute  $(v_{\alpha,j}^n, u_{\alpha,j}^n) = (\bar{v}, \bar{u})_{\alpha,j}^n(\Delta s)$ ,  $j \in \mathbb{Z}, n > 0$ , solution of (5.14),  $\Delta s$  should satisfy the parabolic CFL condition  $a \frac{\Delta s}{\Delta x^2} \leq \frac{1}{2}$ . Now  $(v_{\alpha,j}^n, u_{\alpha,j}^n)$  is an approximation of  $(\bar{v}_\alpha, \bar{u}_\alpha)(x, n\Delta s)$ , solution of the scaled system

$$\begin{cases} \frac{1}{\alpha} \partial_s \bar{v} + \partial_x \bar{u} = 0 \\ \frac{1}{\alpha} \partial_s \bar{u} + a^2 \partial_x \bar{v} = -\alpha \bar{u}; \end{cases} \quad (5.18)$$

the scheme is stable and consistent with (5.18), and as  $\alpha \rightarrow \infty$ , the limit scheme is consistent with the limit system (see (5.19) below).

For the threshold value  $\alpha \Delta x = 2$ , the two coincide, provided  $\Delta s = \frac{1}{2} \Delta t \Delta x$ .

The main conclusion is that one must define clearly the objective when considering the AP property!

### 5.3.2 AP Scheme for the Scaled System

We can more easily study the AP property for a scheme for system (5.2), since there is only one time scale; a natural idea is to require that the scheme be consistent and stable with system (5.2) whatever the value of  $\varepsilon = \frac{1}{\alpha}$ .

In order to be coherent with the above analysis, it will be easier to note the time variable by  $s$  and to overline the functions in the continuous system (5.18)

$$\begin{cases} \frac{1}{\alpha} \partial_s \bar{v} + \partial_x \bar{u} = 0 \\ \frac{1}{\alpha} \partial_s \bar{u} + a^2 \partial_x \bar{v} = -\alpha \bar{u}. \end{cases}$$

If we set again  $\bar{v}(x, s) = v(x, \alpha s)$ ,  $\partial_s \bar{v} = \alpha \partial_t v$ , similarly for  $\bar{u}$ , if  $(v_\alpha, u_\alpha)$  is solution of (5.4),  $(\bar{v}_\alpha, \bar{u}_\alpha)(x, s)$  is solution of (5.18). As  $\alpha \rightarrow \infty$ , at least formally, a solution  $(\bar{v}_\alpha, \bar{u}_\alpha)$  of (5.18) converges to  $(\bar{v}, 0)$  where  $\bar{v}$  satisfies the

limit equation

$$\partial_x \bar{v} - a^2 \partial_{xx}^2 \bar{v} = 0. \quad (5.19)$$

Let us be given initial conditions  $(\bar{v}_0, \bar{u}_0)(x)$  for (5.18) and define as usual  $\bar{v}_j^0 = \frac{1}{\Delta x} \int_{C_j} \bar{v}(x) dx$ , similarly for  $\bar{u}_j^0$ . We consider a given time step  $\Delta s$ . The scheme for (5.18) writes from the scheme for (5.4) with the terms with superscripts  $n$  and  $n+1$  divided by  $\alpha = \frac{1}{\varepsilon}$ . Thus it gives scheme (5.14), and this scheme is consistent with (5.18), i.e.,  $(\bar{v}_{\alpha,j}^n, \bar{u}_{\alpha,j}^n)$  denote an approximation of  $(\bar{v}_\alpha, \bar{u}_\alpha)(x_j, n\Delta s)$ , where  $(\bar{v}_\alpha, \bar{u}_\alpha)$  is solution of (5.18).

We can also conclude that scheme (5.14) is AP (asymptotic preserving) in the weak sense, meaning the (formal) limit scheme (5.15) is consistent with the limit Eq. (5.19).

For the stability analysis, let  $(\bar{v}_{\alpha,\Delta}(x, t), \bar{u}_\Delta(x, t))$  be the piecewise constant functions associated with  $(\bar{v}_{\alpha,j}^n, \bar{u}_{\alpha,j}^n)$  in (5.14). We thus write

$$\left\{ \begin{array}{l} \bar{v}_\Delta(x, s + \Delta s) = \bar{v}_\Delta(x, s) - \alpha K_\alpha \frac{\Delta s}{2\Delta x} \left( \bar{u}_\Delta(x + \Delta x, s) - \bar{u}_\Delta(x - \Delta x, s) \right. \\ \quad \left. - a(\bar{v}_\Delta(x + \Delta x, s) - 2\bar{v}_\Delta(x, s) + \bar{v}_\Delta(x - \Delta x, s)) \right), \\ \bar{u}_\Delta(x, s + \Delta s) = \bar{u}_\Delta(x, s) - \alpha K_\alpha \frac{\Delta s}{\Delta x} \left( \frac{a^2}{2} (\bar{v}_\Delta(x + \Delta x, s) - \bar{v}_\Delta(x - \Delta x, s)) \right. \\ \quad \left. - \frac{a}{2} (\bar{u}_\Delta(x + \Delta x, s) - 2\bar{u}_\Delta(x, s) + \bar{u}_\Delta(x - \Delta x, s)) \right) \\ \quad - \alpha^2 K_\alpha \Delta s \bar{u}_\Delta(x, s). \end{array} \right.$$

After Fourier transform, this relation writes  $\hat{\bar{\mathbf{U}}}_{\alpha,\Delta}(\xi, s + \Delta s) = \mathbf{H}_\alpha \hat{\bar{\mathbf{U}}}_{\alpha,\Delta}(\xi, s)$  with an amplification matrix given by

$$\mathbf{H}_\alpha = \begin{pmatrix} 1 - 2\lambda_\alpha a K_\alpha \sin^2(\frac{1}{2}\xi \Delta x) & -i\lambda_\alpha K_\alpha \sin(\xi \Delta x) \\ -i\lambda_\alpha a^2 K_\alpha \sin(\xi \Delta x) & 1 - 2\lambda_\alpha a K_\alpha \sin^2(\frac{1}{2}\xi \Delta x) - \alpha^2 K_\alpha \Delta s \end{pmatrix} \quad (5.20)$$

where now  $\lambda_\alpha = \alpha \frac{\Delta s}{\Delta x}$ . The amplification matrix  $\mathbf{H}_\alpha$  can be written in the form  $\mathbf{H}_\alpha = \mathbf{I} - 2\lambda_\alpha K_\alpha \mathbf{M}$ , with the same matrix (5.17) as before

$$\mathbf{M} = \begin{pmatrix} ay & i\sqrt{y(1-y)} \\ ia^2 \sqrt{y(1-y)} & ay + \frac{1}{2}\alpha \Delta x \end{pmatrix}.$$

Consider for simplicity the case  $a = 1$  so that the matrix  $\mathbf{H}_\alpha$  is symmetric. Then its  $\mathbf{L}^2$  norm is equal to the spectral radius which is easily computed from that of  $\mathbf{M}$ . And the above analysis shows that the stability condition is  $\lambda_\alpha = \alpha \frac{\Delta s}{\Delta x} \leq 1$  which obviously depends on  $\alpha$  and the scheme is not strongly AP.

Following [556] (see again [180]), we define a numerical scheme for (5.2) which is asymptotic preserving by introducing some implicit treatment which makes the stability condition independent of  $\varepsilon = \frac{1}{\alpha}$ , or at least the dependence does not enforce the time step to go to 0 with  $\varepsilon$ . However, for a linear

system, it amounts to add a multiplicative coefficient, and thus the scheme is in fact explicit.

So let us consider an implicit modification of (5.14). For that purpose, it is better to start from the form (5.11), which writes (we drop the notations emphasizing the dependence on  $\alpha$  when there is no ambiguity)

$$\begin{cases} v_j^{n+1} = v_j^n - \alpha K_\alpha \frac{\Delta s}{2\Delta x} (u_{j+1}^n - u_{j-1}^n - a(v_{j+1}^n - 2v_j^n + v_{j-1}^n)) \\ u_j^{n+1} = u_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta x} (p_{j+1/2} - p_{j-1/2}) - \alpha^2 K_\alpha \Delta s u_j^n, \end{cases} \quad (5.21)$$

with (5.6)

$$p_{j+\frac{1}{2}} = \frac{a^2}{2}(v_j + v_{j+1}) - \frac{a}{2}(u_{j+1} - u_j),$$

the new scheme writes

$$\begin{cases} v_j^{n+1} = v_j^n - \alpha K_\alpha \frac{\Delta s}{2\Delta x} (u_{j+1}^n - u_{j-1}^n - a(v_{j+1}^n - 2v_j^n + v_{j-1}^n)) \\ u_j^{n+1} = u_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta x} (p_{j+1/2} - p_{j-1/2}) - \alpha^2 K_\alpha \Delta s u_j^{n+1}. \end{cases} \quad (5.22)$$

For this new scheme, we can prove the stability, assuming again for simplicity that  $a = 1$ .

*Proposition 5.2*

*Under the CFL condition*

$$\Delta s \leq \frac{\Delta x}{\alpha} + \frac{\Delta x^2}{2}, \quad (5.23)$$

*the scheme (5.22) is  $L^2$ -stable.*

*Proof.* The scheme (5.22) can be written in two steps for the equation defining  $u_j^{n+1}$

$$\begin{cases} v_j^{n+1} = v_j^n - \alpha K_\alpha \frac{\Delta s}{2\Delta x} (u_{j+1}^n - u_{j-1}^n - a(v_{j+1}^n - 2v_j^n + v_{j-1}^n)) \\ u_j^{n+1/2} = u_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta x} (p_{j+1/2} - p_{j-1/2}) \\ u_j^{n+1} = (1 + \alpha^2 K_\alpha \Delta s)^{-1} u_j^{n+1/2}. \end{cases} \quad (5.24)$$

The first step defining  $v_j^{n+1}, u_j^{n+1/2}$  corresponds to an amplification matrix  $\mathbf{H}_{\alpha,0}$  which can be written in the form  $\mathbf{H}_{\alpha,0} = \mathbf{I} - 2\lambda_\alpha K_\alpha \mathbf{M}_0$ , with a matrix

$$\mathbf{M}_0 = \begin{pmatrix} ay & i\sqrt{y(1-y)} \\ ia^2 \sqrt{y(1-y)} & ay \end{pmatrix}.$$

The discriminant for  $\mathbf{M}_0$  is negative. Still considering the simple case  $a = 1$ , the eigenvalues are complex conjugate with modulus bounded by one under the condition  $\lambda_\alpha K_\alpha = \alpha K_\alpha \frac{\Delta s}{\Delta x} \leq 1$ . It is thus  $L^2$  stable under this CFL condition. The second step clearly makes the  $\mathbf{L}^2$  norm decrease.  $\square$

*Remark 5.1.* The above split scheme corresponds at the continuous level to an operator splitting of (5.18) between the convection and the source, i.e.,

$$\begin{cases} \frac{1}{\alpha} \partial_s \bar{v} + \partial_x \bar{u} = 0, \\ \frac{1}{\alpha} \partial_s \bar{u} + a^2 \partial_x \bar{v} = 0, \end{cases}$$

and

$$\begin{cases} \frac{1}{\alpha} \partial_s \bar{v} = 0, \\ \frac{1}{\alpha} \partial_s \bar{u} = -\alpha \bar{u}. \end{cases}$$

For the first step, from (5.24), we solve the system numerically using the relaxation (or simple) scheme with time step say  $\Delta_\alpha s \equiv K_\alpha \Delta s$ . Since  $K_\alpha = (1 + \frac{\alpha \Delta x}{2a})^{-1}$ , for any fixed  $\alpha$  and  $\Delta x$  small enough,  $\Delta_\alpha s = \Delta s(1 + \mathcal{O}(\Delta x))$ , and the scheme is indeed consistent with the first step. Then, recall that the scheme coincides with the upwind scheme in characteristic variable  $w = u + av, z = u - av$ , which is  $L^\infty$  stable under CFL  $a \frac{\Delta_\alpha s}{\Delta x} = a \alpha K_\alpha \frac{\Delta s}{\Delta x} \leq 1$ .

The second step is an ODE in  $u$  to take the source into account; it is treated with the Euler implicit scheme in time and again with time step  $\Delta_\alpha s$

$$u^{n+1} = u^{n+1/2} - \alpha^2 \Delta_\alpha s u^{n+1},$$

and this implicit Euler scheme is unconditionally stable.  $\square$

For  $a = 1$ , the CFL condition  $\lambda_\alpha K_\alpha \leq 1$  writes  $\Delta s \leq \frac{\Delta x}{\alpha} + \frac{\Delta x^2}{2}$ . For small  $\alpha$ , the term  $\frac{\Delta x}{\alpha}$  dominates, and it is a classical hyperbolic CFL condition; for  $\alpha$  large, the term  $\frac{\Delta x^2}{2}$  dominates and gives a parabolic-type condition. A sufficient condition is  $\Delta s \leq \frac{\Delta x^2}{2}$ .

*Corollary 5.1*

*Under the condition (5.23), the scheme (5.22) is strongly asymptotic preserving for (5.18): it is  $L^2$ -stable, consistent with (5.18), and there exists a subsequence  $((v, u)_{\alpha,j}^n)_\alpha$  which converges, as  $\alpha \rightarrow \infty$ , to  $(v, u)_\infty^n$  which satisfies a scheme that is consistent with the limit equation.*

*Proof.* Under the CFL condition, for fixed  $n$ , because their  $\ell^2$  norm is bounded, the sequences (in  $j$ )  $(v_{\alpha,j}^n, u_{\alpha,j}^{n+1/2})$  are uniformly bounded, since the bound involves only the initial condition and does not depend on  $n$ , nor on  $\alpha$ . Thus, for any  $j, n$  we can extract a subsequence which converges as  $\alpha \rightarrow \infty$  to some  $v_{\infty,j}^n, u_{\infty,j}^{n+1/2}$  (and in fact, we can extract a diagonal subsequence). Then, because of the second step, the sequence  $u_{\alpha,j}^n$  converges to  $u_{\infty,j}^n = 0$ . We check easily that  $v_{\infty,j}^n$  satisfies the limit scheme (5.15), which as we have already seen is consistent with the limit equation.  $\square$

We conclude that under the CFL condition  $\Delta s \leq \frac{\Delta x^2}{2}$ , the scheme (5.22) is indeed asymptotic preserving.

## 5.4 Euler System with Gravity and Friction

We come back to the more physical example of Sect. 4.4, in Eulerian or Lagrangian coordinates.

### 5.4.1 Asymptotic Behavior for Large Friction

Let us consider again system (4.62), assuming a linear friction for simplicity

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho u) = 0, \\ \partial_t(\varrho u) + \partial_x(\varrho u^2 + p) = \varrho(g - \alpha u), \\ \partial_t(\varrho e) + \partial_x((\varrho e + p)u) = \varrho u(g - \alpha u). \end{cases} \quad (5.25)$$

From a physical point of view, it is interesting to study the asymptotic behavior of the (smooth) solutions as the friction coefficient becomes large ( $\alpha \rightarrow \infty$ ). Dividing by  $\alpha$  the second equation, we get immediately that (at least formally)  $u \rightarrow 0$  as  $\alpha \rightarrow \infty$ . We can precise the asymptotic behavior wrt.  $\alpha$  of all variables by performing some expansion for each variable:  $\varphi = \varphi^0 + \frac{1}{\alpha} \varphi^1 + \dots$ , by plugging these expansions in (5.25) and considering all the terms of the same order wrt.  $\frac{1}{\alpha}$ . Then, since we are interested in the long-time behavior of the solution for large friction, we introduce as previously a scaling and a notation

$$t = \alpha s, v = \alpha u, \bar{\varphi}(x, s) = \varphi(x, t) \quad (5.26)$$

(if the friction were quadratic, the scaling would be done with  $\alpha$  replaced by  $\beta = \sqrt{\alpha}$ ). System (5.25) becomes

$$\begin{cases} \partial_s \bar{\varrho} + \partial_x(\bar{\varrho} \bar{v}) = 0, \\ \frac{1}{\alpha^2} (\partial_s(\bar{\varrho} \bar{v}) + \partial_x(\bar{\varrho} \bar{v}^2)) + \partial_x \bar{p} = \bar{\varrho}(g - \bar{v}), \\ \frac{1}{2\alpha^2} (\partial_s(\bar{\varrho} \bar{v}^2) + \partial_x(\bar{\varrho} \bar{v}^3)) + \partial_s(\bar{\varrho} \bar{\varepsilon}) + \partial_x((\bar{\varrho} \bar{\varepsilon} + \bar{p}) \bar{v}) = \bar{\varrho} \bar{v}(g - \bar{v}). \end{cases}$$

Letting  $\alpha \rightarrow \infty$ , we obtain formally the following system (dropping the bars for simplicity)

$$\begin{cases} \partial_s \varrho + \partial_x(\varrho v) = 0, \\ \partial_x p = \varrho(g - v), \\ \partial_s(\varrho \varepsilon) + \partial_x((\varrho \varepsilon + p)v) = \varrho v(g - v); \end{cases} \quad (5.27)$$

thus,  $v$  can be given in terms of  $(\varrho, \varepsilon)$

$$v = g - \frac{1}{\varrho} \partial_x p, \quad p = p(\varrho, \varepsilon),$$

the energy equation can be simplified, and we get that  $(\varrho, \varepsilon)$  satisfies

$$\begin{cases} \partial_s \varrho + \partial_x(\varrho v) = 0, \\ \partial_s(\varrho \varepsilon) + \partial_x(\varrho \varepsilon v) - p \partial_x(\frac{1}{\varrho} \partial_x p) = 0, \\ v = g - \frac{1}{\varrho} \partial_x p, \quad p = p(\varrho, \varepsilon); \end{cases} \quad (5.28)$$

one speaks of a *diffusive regime*.

We can study similarly the asymptotic behavior in the Lagrangian frame

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m p = g - \alpha u, \\ \partial_t e + \partial_m(pu) = u(g - \alpha u). \end{cases} \quad (5.29)$$

We perform a change of variables  $t = \alpha s$ ,  $v(m, s) = \alpha u(m, t)$ ,  $\bar{\varphi}(m, s) = \varphi(m, t)$ . With this scaling, (4.43) becomes

$$\begin{cases} \partial_s \bar{\tau} - \partial_m \bar{v} = 0, \\ \frac{1}{\alpha^2} \partial_s \bar{v} + \partial_m \bar{p} = g - \bar{v} \\ \partial_s \bar{\varepsilon} + \frac{1}{2\alpha^2} \partial_s \bar{v}^2 + \partial_m(\bar{p}\bar{v}) = \bar{v}(g - \bar{v}) \end{cases}$$

and the limit system (dropping the bars for simplicity)

$$\begin{cases} \partial_s \tau - \partial_m v = 0, \\ \partial_m p = g - v, \\ \partial_s \varepsilon + \partial_m(pv) = v(g - v). \end{cases} \quad (5.30)$$

Note that (5.30) can be obtained directly from (5.27) by a change of frame from Euler to Lagrange. Now, (5.30) writes equivalently

$$\begin{cases} \partial_s \tau + \partial_{mm}^2 p = 0, \\ \partial_s \varepsilon - p \partial_{mm}^2 p = 0, \\ v = g - \partial_m p, \quad p = \tilde{p}(\tau, \varepsilon), \end{cases}$$

which models the asymptotic behavior as  $\alpha \rightarrow \infty$  of the solutions of (4.43) (for a linear friction).

*Remark 5.2.* In the barotropic case, the Euler system with gravity and friction writes

$$\begin{cases} \partial_t \varrho + \partial_x(\varrho v) = 0, \\ \partial_t \varrho u + \partial_x(\varrho u^2 + p) = \varrho(g - \alpha u), \quad p = p(\varrho), \end{cases}$$

and we get (after scaling) a nonlinear parabolic equation for the limit system in  $\varrho$

$$\begin{cases} \partial_s \varrho + \partial_x(\varrho v) = 0, \\ v = g - \frac{1}{\varrho} \partial_x p, \end{cases}$$

with now  $p = p(\varrho)$ ; hence we find a Darcy-type model

$$\partial_s \varrho + g \partial_x \varrho - \partial_{xx}^2 p(\varrho) = 0.$$

And in the Lagrangian frame, the (limit) barotropic system writes

$$\begin{cases} \partial_s \tau + \partial_{mm}^2 p = 0, \\ v = g - \partial_m p, \quad p = \check{p}(\tau), \end{cases} \quad (5.31)$$

where we have noted  $p = \check{p}(\tau) = p(\varrho)$ . As said for the example of Sect. 5.2.3, in this barotropic case which involves a  $2 \times 2$  system, there are several existing theoretical results justifying the above formal analysis [634, 637, 849].  $\square$

### 5.4.2 Asymptotic Preserving Property

Let us consider the simple well-balanced scheme (4.59) for (5.29) which we have constructed in Sect. 4.5.2 for the full Euler system with energy and study the asymptotic behavior of the scheme when  $\alpha$  goes to infinity. As for the linear case in the above section, we begin by some formal analysis.

The asymptotic preserving property of the scheme requires a specific choice for  $\tilde{u}$ , and we have chosen

$$\tilde{u} = u^*,$$

the value which was naturally given in the relaxation approach with potential. This choice of  $\tilde{u}$  together with (4.51) makes the computation of  $u^*$  easy for a linear friction  $\varphi(u) = u$ , with (4.58)

$$K_\alpha = (1 + \frac{\alpha \Delta m}{2C})^{-1},$$

we get

$$u^* = u^*(\mathbf{u}_L, \mathbf{u}_R) = K_\alpha \left( \frac{1}{2}(u_L + u_R) - \frac{1}{2C}(\Delta p - g \Delta m) \right), \quad (5.32)$$

we note that (4.56) holds true by continuity arguments. Setting

$$u_{j+\frac{1}{2}} = u^*(\mathbf{u}_j, \mathbf{u}_{j+1}), \quad p_{j+\frac{1}{2}} = p^*(\mathbf{u}_j, \mathbf{u}_{j+1}),$$

and skipping the time superscripts in the right-hand side to lighten the notations, the numerical scheme writes

$$\left\{ \begin{array}{l} \tau_j^{n+1} = \tau_j + \frac{\Delta t}{\Delta m} (u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}), \\ u_j^{n+1} = u_j - \frac{\Delta t}{\Delta m} (p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) + \Delta t \left( g - \frac{\alpha}{2} (u_{j-\frac{1}{2}} + u_{j+\frac{1}{2}}) \right), \\ e_j^{n+1} = e_j - \frac{\Delta t}{\Delta m} (p_{j+\frac{1}{2}} u_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} u_{j-\frac{1}{2}}) \\ \quad + \frac{\Delta t}{2} \left( g(u_{j-\frac{1}{2}} + u_{j+\frac{1}{2}}) - \alpha(u_{j-\frac{1}{2}}^2 + u_{j+\frac{1}{2}}^2) \right) \end{array} \right. \quad (5.33)$$

with, using (5.32),

$$\left\{ \begin{array}{l} u_{j+\frac{1}{2}} = K_\alpha \left( \frac{1}{2} (u_j + u_{j+1}) - \frac{1}{C} (p_{j+1} - p_j - \Delta m g) \right), \\ p_{j+\frac{1}{2}} = \frac{1}{2} \{ p_j + p_{j+1} - C(u_{j+1} - u_j) \}. \end{array} \right. \quad (5.34)$$

Now we make the change of variable induced by (5.26)

$$\Delta t = \alpha \Delta s, \quad v = \alpha u \quad (5.35)$$

(in case of a quadratic friction term,  $\alpha$  should be replaced by  $\beta = \sqrt{\alpha}$ ), so that  $v_L = \alpha u_L$  and  $v_R = \alpha u_R$ ; it is also relevant to set

$$v^* = \alpha u^*; \quad (5.36)$$

in fact, it is natural to require that the change of variables  $v = \alpha u$  be relevant for all velocities under consideration; then (5.32) becomes

$$v^*(\mathbf{u}_L, \mathbf{u}_R) = \alpha K_\alpha \left( \frac{1}{2\alpha} (v_L + v_R) - \frac{1}{2C} (\Delta p - \Delta m g) \right).$$

Recall that as  $\alpha$  goes to  $+\infty$ , one has

$$\alpha K_\alpha = \alpha \left( 1 + \frac{\alpha \Delta m}{2C} \right)^{-1} \rightarrow \frac{2C}{\Delta m},$$

and thus

$$v^* \rightarrow -\frac{\Delta p}{\Delta m} + g. \quad (5.37)$$

With (5.35), (5.36), the scheme becomes

$$\left\{ \begin{array}{l} \tau_j^{n+1} = \tau_j + \frac{\Delta s}{\Delta m} (v_{j+\frac{1}{2}} - v_{j-\frac{1}{2}}), \\ \frac{1}{\alpha} v_j^{n+1} = \frac{1}{\alpha} v_j - \frac{\alpha \Delta s}{\Delta m} (p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) + \alpha \Delta s \left( g - \frac{1}{2} (v_{j-\frac{1}{2}} + v_{j+\frac{1}{2}}) \right), \\ \varepsilon_j^{n+1} + \frac{1}{2\alpha^2} (v_j^{n+1})^2 = \varepsilon_j + \frac{1}{2\alpha^2} v_j^2 - \frac{\Delta s}{\Delta m} (p_{j+\frac{1}{2}} v_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} v_{j-\frac{1}{2}}) \\ \quad + \frac{\Delta s}{2} \left( g(v_{j-\frac{1}{2}} + v_{j+\frac{1}{2}}) - (v_{j-\frac{1}{2}}^2 + v_{j+\frac{1}{2}}^2) \right), \end{array} \right. \quad (5.38)$$

with

$$\left\{ \begin{array}{l} v_{j+\frac{1}{2}} = \alpha K_\alpha \left( \frac{1}{2\alpha} (v_j + v_{j+1}) - \frac{1}{2C} (p_{j+1} - p_j - \Delta m g) \right), \\ p_{j+\frac{1}{2}} = \frac{1}{2} \left( p_j + p_{j+1} - \frac{C}{\alpha} (v_{j+1} - v_j) \right). \end{array} \right. \quad (5.39)$$

In the limit  $\alpha \rightarrow \infty$ , the numerical scheme tends (formally, if  $\Delta s$  is constant) to

$$\left\{ \begin{array}{l} \tau_j^{n+1} = \tau_j + \frac{\Delta s}{\Delta m} (v_{j+\frac{1}{2}} - v_{j-\frac{1}{2}}), \\ \frac{1}{\Delta m} (p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) = g - \frac{1}{2} (v_{j-\frac{1}{2}} + v_{j+\frac{1}{2}}), \\ \varepsilon_j^{n+1} = \varepsilon_j - \frac{\Delta s}{\Delta m} (p_{j+\frac{1}{2}} v_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} v_{j-\frac{1}{2}}) \\ \quad + \frac{\Delta s}{2} \left( g(v_{j-\frac{1}{2}} + v_{j+\frac{1}{2}}) - (v_{j-\frac{1}{2}}^2 + v_{j+\frac{1}{2}}^2) \right), \end{array} \right.$$

with

$$\left\{ \begin{array}{l} v_{j+\frac{1}{2}} = -\frac{p_{j+1} - p_j}{\Delta m} + g, \\ p_{j+\frac{1}{2}} = \frac{1}{2} (p_j + p_{j+1}). \end{array} \right. \quad (5.40)$$

In particular, the pair  $(\tau, \varepsilon)$  evolves according to

$$\left\{ \begin{array}{l} \tau_j^{n+1} = \tau_j + \frac{\Delta s}{\Delta m} (v_{j+\frac{1}{2}} - v_{j-\frac{1}{2}}), \\ \varepsilon_j^{n+1} = \varepsilon_j - \frac{\Delta s}{\Delta m} (p_{j+\frac{1}{2}} v_{j+\frac{1}{2}} - p_{j-\frac{1}{2}} v_{j-\frac{1}{2}}) \\ \quad + \frac{\Delta s}{2} \left( g(v_{j-\frac{1}{2}} + v_{j+\frac{1}{2}}) - (v_{j-\frac{1}{2}}^2 + v_{j+\frac{1}{2}}^2) \right). \end{array} \right. \quad (5.41)$$

We have not used the same heavy system of notations as in (5.14), (5.15); a quantity  $\varphi_j^n$  that would be noted  $\bar{\varphi}_{\alpha,j}^n(\Delta s)$  represents an approximation of  $\bar{\varphi}(x_j, s_n)$  (with the notations of (5.26)); then we have dropped the notation with a bar in (5.29). Thus scheme (5.40), together with the relations (5.40) and  $p_j = p(\tau_j, \varepsilon_j)$ , is indeed a consistent explicit numerical scheme for the asymptotic system (5.30)

$$\begin{cases} \partial_s \tau - \partial_m v = 0, \\ \partial_s \varepsilon + \partial_m (pv) = gv - v^2, \\ v = g - \partial_m p, \quad p = p(\tau, \varepsilon). \end{cases}$$

To sum up, we have shown:

*Proposition 5.3.*

The Godunov-type scheme (5.33) is asymptotic preserving for system (5.29) in a weak sense, meaning that, when performing the same scaling (5.35) as the one done to obtain the limit system (5.30), it provides as  $\alpha \rightarrow \infty$  a scheme which is consistent with the limit system if  $\Delta s$  is constant.

We have the same result for the Eulerian system.

*Remark 5.3.* This (weak) AP property is derived directly from the nature of the exact Godunov solver, at least from a heuristic point of view. When considering only the simple Godunov-type solver approach, the choice  $\tilde{u} = u^*$  for the velocity of the source term  $\tilde{u}$  seems quite “natural”:  $u^*$  is the common speed of the intermediate states of the Riemann solver (in  $-C < m/t < C$ ), and we are considering the large time behavior  $t = \alpha s$ , so that considering the value inside the fan for the asymptotic velocity was a priori judicious. But it does not “prove” the AP property.

If we consider the relaxation point of view, introducing a “potential,” we note that the scaling is only active in the evolution step, not in the projection (instantaneous relaxation) step, and the evolution step uses an exact solver for a differential (relaxation) system which mimics the original system (4.72) (the Euler system with friction and a potential), which preserves the asymptotic behavior of the original system. Indeed, in Eulerian coordinates, the advection equations are invariant:  $\partial_t \theta + u \partial_x \theta = 0$ , where  $\theta$  is any quantity advected by the flow, becomes after scaling (5.26)  $\partial_s \tilde{\theta} + \tilde{v} \partial_x \tilde{\theta} = 0$ . The momentum equation becomes (at the order 0 in  $\frac{1}{\alpha}$ )  $\partial_x \Pi = (g - v) \partial_x q$ . Thus, mimicking what is done for the Euler system, we obtain (from a formal continuity argument) that the solutions of the homogeneous relaxation system with potential

$$\begin{cases} \partial_t \varrho + \partial_x (\varrho u) &= 0, \\ \partial_t (\varrho u) + \partial_x (\varrho u^2 + \Pi) - (g - \alpha u) \partial_x q &= 0, \\ \partial_t \varrho \eta + \partial_x (\varrho \eta u) &= 0, \\ \partial_t (\varrho T) + \partial_x (\varrho T u) &= 0, \\ \partial_t (\varrho q) + \partial_x (\varrho u q) &= 0, \end{cases}$$

tend (after scaling (5.26), and as  $\alpha \rightarrow \infty$ ) to those of the following system (dropping the notation tilde)

$$\begin{cases} \partial_s \varrho + \partial_x(\varrho v) &= 0, \\ \partial_x \Pi - (g - v) \partial_x q &= 0, \\ \partial_s(\varrho \eta) + \partial_x(\varrho \eta v) &= 0, \\ \partial_s(\varrho \mathcal{T}) + \partial_x(\varrho \mathcal{T} v) &= 0, \\ \partial_s(\varrho q) + \partial_x(\varrho v q) &= 0, \end{cases}$$

or for the energy equation

$$\partial_s \varrho \varepsilon + \partial_x((\varrho \tilde{\varepsilon} + \Pi)v) - v(g - v) \partial_x q = 0.$$

For the relaxation process, we also scale the relaxation parameter  $\mu = \alpha\nu$  so that the relaxation term writes  $\mu\varrho(\tau - \mathcal{T})$ . Thus for the global relaxation solver, after first reconstruction by piecewise constant for Godunov's scheme, the exact evolution step and the projection-relaxation on the equilibrium manifold (set of states satisfying  $\mathcal{T} = \tau$ ), we get (after scaling) as  $\alpha \rightarrow \infty$  a scheme which is by construction consistent with the limit system (this is formal in the sense that we assume that all these strong limits exist; we do not prove convergence).

It is even easier to write the above lines in Lagrangian coordinates for system (5.29), since in a Lagrangian frame, the advection of a quantity  $\theta$  writes  $\partial_t \theta = 0$ ; hence, the solutions of

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m \Pi - (g - \alpha u) \partial_m q = 0, \\ \partial_t \eta = 0, \\ \partial_t \mathcal{T} = 0, \\ \partial_t q = 0 \end{cases}$$

tend (after scaling, and as  $\alpha \rightarrow \infty$ ) to those of the following system

$$\begin{cases} \partial_s \tau - \partial_m v = 0, \\ \partial_m \Pi - (g - v) \partial_m q, \\ \partial_s \eta = 0, \\ \partial_s \mathcal{T} = 0, \\ \partial_s q = 0, \end{cases}$$

or for the energy equation

$$\partial_s \varepsilon + \partial_m(\Pi v) - v(g - v) \partial_m q = 0,$$

and we conclude similarly that the limit (as  $\alpha \rightarrow \infty$ ) relaxation scheme is consistent with the limit system (5.30).  $\square$

### 5.4.3 An AP Modification of the Scheme

As remarked in Sect. 5.3, the result of Proposition 5.3 is very formal since it assumes  $\Delta s$  constant. And we know from Sect. 5.3.2 that in the linear (barotropic) case, the stability condition for the scaled system is not acceptable, since it writes  $\alpha \Delta s \leq \frac{1}{2} \Delta m$ . As in Sect. 5.3.2, we now add an implicit treatment, so as to derive an AP scheme for the scaled system.

Let us focus on the barotropic case for simplicity. In Lagrangian coordinates, we are left with a system of the form (5.1); once scaled by setting  $t = \alpha s$ , this system writes

$$\begin{cases} \frac{1}{\alpha} \partial_s \tau - \partial_m u = 0, \\ \frac{\alpha}{\Delta m} \partial_s u + \partial_m p(\tau) = g - \alpha u. \end{cases} \quad (5.42)$$

Then the scheme resulting from (5.33) can be written before scaling

$$\begin{cases} \tau_j^{n+1} = \tau_j + \frac{\Delta t}{\Delta m} (u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}), \\ u_j^{n+1} = u_j - \frac{\Delta t}{\Delta m} (p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) + \Delta t \left( g - \frac{\alpha}{2} (u_{j-\frac{1}{2}} + u_{j+\frac{1}{2}}) \right), \end{cases} \quad (5.43)$$

and after scaling

$$\begin{cases} \tau_j^{n+1} = \tau_j + \alpha \frac{\Delta s}{\Delta m} (u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}), \\ u_j^{n+1} = u_j - \alpha \frac{\Delta s}{\Delta m} (p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) + \alpha \Delta s \left( g - \frac{\alpha}{2} (u_{j-\frac{1}{2}} + u_{j+\frac{1}{2}}) \right), \end{cases} \quad (5.44)$$

with, again using (5.34),

$$\begin{cases} u_{j+\frac{1}{2}} = K_\alpha \left( \frac{1}{2} (u_j + u_{j+1}) - \frac{1}{2C} (p_{j+1} - p_j - \Delta m g) \right), \\ p_{j+\frac{1}{2}} = \frac{1}{2} (p_j + p_{j+1} - C(u_{j+1} - u_j)). \end{cases}$$

It can be written equivalently (see (5.11))

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \alpha K_\alpha \frac{\Delta s}{2 \Delta m} \left( u_{j+1}^n - u_{j-1}^n - \frac{1}{C} (p_{j+1}^n - 2p_j^n + p_{j-1}^n) \right), \\ u_j^{n+1} = u_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta m} \left( p_{j+1/2}^n - p_{j-1/2}^n \right) + \alpha K_\alpha \Delta s g - \alpha^2 K_\alpha \Delta s u_j^n. \end{cases}$$

The implicit modification writes

$$\begin{cases} \tau_j^{n+1} = \tau_j^n + \alpha K_\alpha \frac{\Delta s}{2\Delta m} \left( u_{j+1}^n - u_{j-1}^n - \frac{1}{C} (p_{j+1}^n - 2p_j^n + p_{j-1}^n) \right), \\ u_j^{n+1} = u_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta m} \left( p_{j+1/2}^n - p_{j-1/2}^n \right) + \alpha K_\alpha \Delta s g - \alpha^2 K_\alpha \Delta s u_j^{n+1}. \end{cases} \quad (5.45)$$

*Proposition 5.4*

Let  $K_\alpha$  be defined by (4.58). The scheme (5.45) is consistent with system (5.42). Under Whitham's condition  $C > \max \sqrt{-p'(\tau)}$  (the maximum over the range of values taken by the  $\tau_j^n$ ), and a CFL condition  $\alpha K_\alpha C \Delta s \leq \frac{1}{2} \Delta m$ , it satisfies a discrete entropy inequality

$$e_j^{n+1} \leq e_j^n + \frac{\Delta t}{\Delta m} ((pu)_{j+1/2}^n - (pu)_{j-1/2})^n + \alpha K_\alpha \Delta s g u_j^n, \quad (5.46)$$

where  $e = e(\tau, u) = \epsilon(\tau) + \frac{1}{2} u^2$  and  $\epsilon'(\tau) = -p$ . Moreover the scheme is asymptotic preserving for system (5.42): it gives as  $\alpha \rightarrow \infty$  a scheme which is consistent with the limit system (5.31).

*Proof.* Scheme (5.45) can be decomposed in two steps (as for (5.24))

$$\begin{cases} \tau_j^{n+1/2} = \tau_j^n + \alpha K_\alpha \frac{\Delta s}{2\Delta m} \left( u_{j+1}^n - u_{j-1}^n - \frac{1}{C} (p_{j+1}^n - 2p_j^n + p_{j-1}^n) \right), \\ u_j^{n+1/2} = u_j^n - \alpha K_\alpha \frac{\Delta s}{\Delta m} \left( p_{j+1/2}^n - p_{j-1/2}^n \right) + \alpha K_\alpha \Delta s g, \\ \tau_j^{n+1} = \tau_j^{n+1/2}, \\ u_j^{n+1} = (1 + \alpha^2 K_\alpha \Delta s)^{-1} u_j^{n+1/2}. \end{cases} \quad (5.47)$$

As in Remark 5.1, we see that the above split scheme corresponds at the continuous level to an operator splitting of (5.42) between the scaled  $p$ -system with only the gravity source term, and the friction source, i.e.,

$$\begin{cases} \frac{1}{\alpha} \partial_s \tau - \partial_m u = 0, \\ \frac{1}{\alpha} \partial_s u + \partial_m p(\tau) = g, \end{cases}$$

and

$$\begin{cases} \frac{1}{\alpha} \partial_s v = 0, \\ \frac{1}{\alpha} \partial_s u = -\alpha u. \end{cases}$$

For the first step, (5.47) shows that we solve the scaled  $p$ -system with only the gravity source term numerically using the relaxation (or simple) scheme with time step say  $\Delta_\alpha s \equiv K_\alpha \Delta s$ . As already written in Remark 5.1, since (see (4.58))  $K_\alpha = (1 + \frac{\alpha \Delta m}{2C})^{-1}$ , for any fixed  $\alpha$  and  $\Delta x$  small enough,  $\Delta_\alpha s = \Delta s (1 + \mathcal{O}(\Delta x))$ , and the scheme is indeed consistent with the first step. The last step takes the friction source term into account; it is treated with the Euler implicit scheme in time and with time step  $\Delta_\alpha s$ .

$$u_j^{n+1} = u_j^{n+1/2} - \alpha^2 \Delta_\alpha s u_j^{n+1};$$

this step is again consistent for any fixed  $\alpha$ .

For the stability we notice that for the first step defining  $\tau_j^{n+1/2}, u_j^{n+1/2}$ , the numerical scheme corresponds to a relaxation (or HLL) scheme for the (unscaled)  $p$ -system with gravity

$$\begin{cases} \partial_t \tau - \partial_m u = 0, \\ \partial_t u + \partial_m p(\tau) = g, \end{cases}$$

provided the time step is set as  $\Delta t = \alpha K_\alpha \Delta s$ . This relaxation scheme satisfies an entropy inequality, where the entropy is the usual energy  $e(\tau, u) = \epsilon(\tau) + \frac{u^2}{2}$ , under Whitham's condition  $C > \sqrt{-p'(\tau)}$  (the maximum over the range of values taken by the  $\tau_j^n$ ), and a CFL condition  $C\Delta t \leq \frac{1}{2}\Delta m$ . Indeed, smooth solutions of the  $p$ -system with  $g$  for only source term satisfy  $\partial_t e + \partial_m(pu) = gu$ , where  $e = \epsilon + \frac{1}{2}u^2$  and  $\epsilon'(\tau) = -p$ ; it becomes an inequality for discontinuous solutions. For the relaxation system, the energy writes (see in Chap. IV, Sect. 8.3.2, (8.32)), only we drop the notation tilde which emphasizes the change of frame)  $\Sigma = \epsilon(\mathcal{T}) + \frac{1}{2}u^2 + \frac{1}{2C^2}(I^2 - p^2(\mathcal{T}))$ . Since all the fields are LD, this equation is satisfied by any weak solution. When we integrate the energy equation over a (Lagrangian) cell  $C_j \times (t_n, t_{n+1})$ , we get

$$\frac{1}{\Delta m} \int_{C_j} \Sigma(x, t_{n+1/2}) dx \leq e_j^n + \frac{\Delta t}{\Delta m} ((pu)_{j+1/2}^n - (pu)_{j-1/2})^n + \Delta t g u_j^n.$$

We have proved that  $\Sigma$  is minimum on the equilibrium manifold; hence,  $e_j^{n+1/2} = e(\tau_j^{n+1/2}, u_j^{n+1/2}) \leq \frac{1}{\Delta m} \int_{C_j} \Sigma(x, t_{n+1}) dx$ .

Thus  $e_j^{n+1/2} \leq e_j^n + \frac{\Delta t}{\Delta m} + (pu)_{j+1/2} - (pu)_{j-1/2} + \alpha K_\alpha \Delta s g u_j$ .

Then, in the last step the implicit Euler scheme is unconditionally stable, and we check easily that

$$e_j^{n+1} = e(\tau_j^{n+1}, u_j^{n+1}) = e(\tau_j^{n+1/2}, u_j^{n+1}) \leq e(\tau_j^{n+1/2}, u_j^{n+1/2}),$$

which yields the entropy inequality.  $\square$

*Remark 5.4.* The CFL condition  $\alpha K_\alpha \Delta s C \leq \frac{1}{2}$  writes  $C\Delta s \leq \frac{\Delta m}{2\alpha} + \frac{\Delta m^2}{4C}$ . For “small”  $\alpha$ , the first term  $\frac{\Delta m}{2\alpha}$  dominates, and it is a classical hyperbolic CFL condition; for  $\alpha$  large, the term  $\frac{\Delta m^2}{4C}$  dominates and gives a parabolic-type condition relevant for the limit system. A sufficient condition, valid whatever  $\alpha$ , is  $\Delta s \leq \frac{\Delta m^2}{4C^2}$ .  $\square$

We have only considered a particular example and a numerical scheme which we think are a good illustration of the topic; for this example, more details and many references are found in [267], a more recent nearby approach [299]. There are many other interesting works, among which [131, 140, 420,

[448, 482], and references therein, or [495] (in the context of ideal MHD with gravity), [118, 675, 1195]; it is impossible to cite them all.

We did not present a deeper description of IMEX (implicit/explicit) methods for systems of hyperbolic conservation laws with stiff diffusive relaxation, involving Runge-Kutta discretization [156, 157] and references therein.

## 6 Interface Coupling

We now address the new question of coupling two systems at an interface, in a limited framework which we will describe just below (in particular, we do not consider general heterogeneous domain decomposition methods with alternating Schwarz algorithm). According to the condition one chooses to impose at the interface, the coupling may generate a measure source term as we will see with (6.10). It thus provides a different category of source terms; we can however make a link with some of the notions introduced above (in particular the well-balanced property).

### 6.1 Introduction to Interface Coupling

The question of coupling was already mentioned above as possibly involved in the treatment of multiscale problems. If we have a fine (relaxation) and a coarse (equilibrium) model, assuming we know the regions where each one is respectively valid, one can imagine a dedicated scheme in each domain and a numerical interface coupling procedure. Works in this direction then involve moreover the definition of an indicator of the modeling error, together with an adaptive procedure in order to construct the fine and coarse regions and let the interface evolves with time. The method would then allow the simulation of a complex flow, where different models are taken into account, in general to save CPU time by using the coarse model whenever possible. In this strategy, some precise coupling model is necessary, and we restrict ourselves to the topic of coupling (a dynamic adaptation procedure is proposed in [856] and its justification on a toy model in [228]).

A very simple model of such a situation is the coupling of a Jin-Xin relaxation  $2 \times 2$  system [668]

$$\begin{cases} \partial_t u - \partial_x v = 0 \\ \partial_t v + a^2 \partial_x u = -\alpha(v - f(u)), \end{cases}$$

with the equilibrium equation

$$\partial_t u + \partial_x f(u) = 0,$$

an example on which we will come again below. A more complex example concerns the coupling of two homogeneous models used for describing non-isothermal compressible two-phase flows, namely, the HRM (relaxation)/HEM (equilibrium) models [37, 121]. We can also mention the coupling of Euler equations with friction with the limit model (hyperbolic/parabolic coupling [176]), or a kinetic/parabolic coupling (with a smooth transition) in [402], then coupling a compressible fluid and the incompressible limit [339, 900].

One may think of coupling fluid models of the same size, slightly differing by their equation of state only. A simple representative model of such a situation is the coupling of two Euler systems with slightly different pressure laws; assuming ideal gases, we will have two different values of  $\gamma$ . For this situation, the corresponding fluxes depend on a parameter (for instance  $\gamma$ ) which depends on  $x$ , but does not vary with time. The coupling at a fixed interface is different in Lagrangian and Eulerian coordinates. In the former frame, the interface is a material interface across which the pressure is continuous. There is a natural conservative model in this case; we may speak of conservative coupling. From the numerical point of view, the discontinuity of  $\gamma$  needs some special treatment, since the pressure is not a conservative variable. The numerical coupling model can correspond to a scheme for a multicomponent flow, which has received a particular interest in [10] (see also [81]).

In an Eulerian frame, a fixed interface is artificial; it may correspond either to the coupling of two codes which do not take the same equation of state into account or to some device which renders the flux discontinuous, for instance, the flow on one side of the interface is governed by standard Euler equations in a free medium, whereas it enters a porous medium on the other side.

The coupling condition is not so straightforward (see [274]); the choice is not unique. An illustration is given in Sect. 6.2.3 below. In the example where the two fluid systems differ by their equation of state, it is easy to see that starting from a uniform profile, keeping a uniform pressure and velocity profile is incompatible with energy conservation. One has to appeal to some physical intuition such as preservation of some quantity (or definition of some invariant). The choice of what is transmitted at the interface is expressed in the *interface coupling condition*. Mathematically, this coupling condition must lead to a well-posed problem.

The issue of coupling models naturally leads to the question of discontinuous flux which can occur in many applications; one of the first examples concerns flow in porous media with discontinuous permeability [1028], then continuous sedimentation [219] and references therein, traffic flow [220, 611], etc. There has been many theoretical works devoted to the subject in the last few years, and we only cite a few of them [22, 67, 70]; see [129] for a BGK approximation [42] and for a general frame [45] and many other references therein; let us also mention the case of linear equations with discontinuous coefficients [170]. This question was already mentioned in Remark 2.2, when

we considered solving the Riemann problem for a conservation law with a discontinuous flux function  $f(u, x)$  with

$$f(u, x) = \begin{cases} f_L, & x < 0, \\ f_R, & x > 0. \end{cases}$$

Given a Riemann data  $u_l, u_r$ , one solves

$$\begin{cases} \partial_t u + \partial_x f_L(u) = 0, & x < 0, \\ \partial_t u + \partial_x f_R(u) = 0, & x > 0, \end{cases}$$

with

$$u(x, 0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0. \end{cases}$$

If the problem can be written in the conservative form

$$\partial_t u + \partial_x f(u, x) = 0, \quad x \in \mathbb{R}, t > 0,$$

assuming there exists a weak solution which has well-defined traces at the interface  $x = 0$ , the solution will satisfy Rankine-Hugoniot condition which writes

$$f_l(u(0-, t)) = f_r(u(0+, t)).$$

One can speak of conservative coupling or *flux coupling*.

Such a situation arises naturally for the junction of pipes [341, 342] then particularly in traffic flow context in the modeling of junctions and traffic flow on networks (see, for example, [35, 610, 699]; see also [338, 507, 508]), [340, 509] and the recent survey [199, 971].

Otherwise, if some other interface condition is prescribed, we obtain

$$\partial_t u + \partial_x f(u, x) = \mathcal{M}\delta_0, \quad x \in \mathbb{R}, t > 0,$$

where  $\delta_0$  denotes the Dirac measure with support  $x = 0$  and the weight  $\mathcal{M} = f(u(0+, t), 0+) - f(u(0-, t), 0-)$ ; this equation takes into account the lack of conservation  $f_R(u(0+, t)) - f_L(u(0-, t))$ . For instance, one may want to ensure the continuity of the state,  $u(0+, t) = u(0-, t)$ . One then speaks of nonconservative coupling or *state coupling*.

Finally, the coupling condition can be modeled thanks to a bounded Dirac measure concentrated at the coupling interface. The coupling condition is then prescribed from the definition of the mass of the measure.

The interface is (infinitely) thin. We do not consider here the possible thickening of the interface [188, 402], neither do we develop a regularization procedure as in [353], nor for instance using Dafermos regularization [187].

The question can be put in a general setting for coupling systems at a fixed interface.

## 6.2 The Interface Coupling Condition

For simplicity, we present this framework for systems of the same dimension. Otherwise, one needs to add some projection and lift operators, which map one set of states to the other (see Sect. 6.2.4 below).

### 6.2.1 General Interface Coupling

Let  $\Omega \subset \mathbb{R}^p$  be the set of states, and let  $\mathbf{f}_\alpha$ ,  $\alpha = L, R$ , be two smooth functions from  $\Omega$  into  $\mathbb{R}^p$ , such that the corresponding systems are hyperbolic, i.e., for  $\alpha = L, R$ , the Jacobian matrix  $A_\alpha(\mathbf{u}) \equiv \mathbf{f}'_\alpha(\mathbf{u})$  of  $\mathbf{f}_\alpha(\mathbf{u})$  is diagonalizable with real eigenvalues  $\lambda_{\alpha,k}(\mathbf{u})$  and corresponding eigenvectors  $\mathbf{r}_{\alpha,k}(\mathbf{u})$ ,  $1 \leq k \leq p$ . Given a function  $\mathbf{u}_0 : x \in \mathbb{R} \mapsto \mathbf{u}_0(x)$ , we want to find a function  $\mathbf{u} : (x, t) \in \mathbb{R} \times \mathbb{R}_+ \mapsto \mathbf{u}(x, t) \in \Omega$  solution of

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}_L(\mathbf{u}) = \mathbf{0}, \quad x < 0, t > 0, \quad (6.1)$$

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}_R(\mathbf{u}) = \mathbf{0}, \quad x > 0, t > 0, \quad (6.2)$$

satisfying the initial condition

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x), \quad x \in \mathbb{R},$$

and, at the interface  $x = 0$ , a *coupling condition*.

The most frequent coupling condition is associated with a conservative modeling: one assumes that the problem comes from a system of conservation law with discontinuous flux

$$\begin{cases} \partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, x) = \mathbf{0}, x \in \mathbb{R}, t > 0 \\ \mathbf{f}(\mathbf{u}, x) = \begin{cases} \mathbf{f}_L, & x < 0 \\ \mathbf{f}_R, & x > 0 \end{cases} \end{cases} \quad (6.3)$$

which implies the continuity of the flux

$$\mathbf{f}_L(\mathbf{u}(0-, t)) = \mathbf{f}_R(\mathbf{u}(0+, t)), \quad (6.4)$$

assuming that these traces exist.

If we do not restrict ourselves to this conservative approach, there are several ways to define an interface condition. From a theoretical point of view, we want to define the coupling condition in order to obtain two well-posed initial boundary-value problems in  $x > 0$ , and in  $x < 0$ , for  $t \geq 0$ . This means that the trace  $\mathbf{u}(0-, t)$  (resp.  $\mathbf{u}(0+, t)$ ) should be an admissible boundary condition at  $x = 0$  for the system in  $x > 0$ ,  $t \geq 0$  (resp.  $\mathbf{u}(0+, t)$  is an admissible boundary condition at  $x = 0$  for the system in  $x < 0$ ,  $t \geq 0$ ). From the results of Chap. VI, we know that the most practical way to express the problem involves the traces of the solution of a Riemann problem. Thus,

we introduce the self-similar solution

$$\mathbf{u}(x, t) = \mathbf{W}_\alpha(x/t; \mathbf{u}_L, \mathbf{u}_R)$$

of the Riemann problem for the system associated with the flux  $\mathbf{f}_\alpha$ , i.e., the Cauchy problem with initial condition

$$\mathbf{u}(x, 0) = \begin{cases} \mathbf{u}_L, & x < 0, \\ \mathbf{u}_R, & x > 0. \end{cases} \quad (6.5)$$

We set for all  $\mathbf{b} \in \Omega$ ,

$$\begin{aligned} \mathcal{O}_L(\mathbf{b}) &= \{\mathbf{w} = \mathbf{W}_L(0-; \mathbf{u}_\ell, \mathbf{b}); \mathbf{u}_\ell \in \Omega\}, \\ \mathcal{O}_R(\mathbf{b}) &= \{\mathbf{w} = \mathbf{W}_R(0+; \mathbf{b}, \mathbf{u}_r); \mathbf{u}_r \in \Omega\} \end{aligned}$$

and following Chap. VI, Definition 2.2 we define admissible boundary conditions of the form

$$\mathbf{u}(0-, t) \in \mathcal{O}_L(\mathbf{b}(t)), \quad t > 0$$

and

$$\mathbf{u}(0+, t) \in \mathcal{O}_R(\mathbf{b}(t)), \quad t > 0,$$

for (6.1) and (6.2) respectively. Hence, it is natural to assume that the solution has left and right traces on the interface and to require as coupling conditions for problem (6.1)–(6.2)

$$\begin{cases} \mathbf{u}(0-, t) \in \mathcal{O}_L(\mathbf{u}(0+, t)), \\ \mathbf{u}(0+, t) \in \mathcal{O}_R(\mathbf{u}(0-, t)). \end{cases} \quad (6.6)$$

This means that at each time  $t > 0$ , there exists some states  $\mathbf{u}_-(t), \mathbf{u}_+(t) \in \Omega$  such that  $\mathbf{u}(0-, t) = \mathbf{W}_L(0-; \mathbf{u}_-(t), \mathbf{u}(0+, t))$ , and similarly we want  $\mathbf{u}(0+, t) = \mathbf{W}_R(0+; \mathbf{u}(0-, t), \mathbf{u}_+(t))$ . Then, if these conditions hold, one also has  $\mathbf{u}(0-, t) = \mathbf{W}_L(0-; \mathbf{u}(0-, t), \mathbf{u}(0+, t))$  and similarly for the right trace  $\mathbf{u}(0+, t) = \mathbf{W}_R(0+; \mathbf{u}(0-, t), \mathbf{u}(0+, t))$ . Using the formulation with Riemann problems to express admissible boundary conditions is very practical and suitable for the numerical approximation of the coupled problem. It is thoroughly justified in the scalar case and for linear systems.

From a heuristic point of view, (6.6) gives a number of relations between the traces which is linked to the number of positive eigenvalues on each side. If all the eigenvalues are positive  $> 0$  (resp. negative  $< 0$ ), it provides  $p$  relations linking the left and right traces, and the condition completely determines  $\mathbf{u}(0+, t)$  from  $\mathbf{u}(0-, t)$  (resp.  $\mathbf{u}(0-, t)$  from  $\mathbf{u}(0+, t)$ ). Thus, condition (6.6) resumes in a number of cases to the continuity of the solution at the interface

$$\mathbf{u}(0-, t) = \mathbf{u}(0+, t). \quad (6.7)$$

We may interpret the coupling condition as a way of ensuring in a weak sense the continuity or the *transmission* of the conservative variables  $\mathbf{u}$ . When eigenvalues of the  $\mathbf{f}'_\alpha(\mathbf{u})$  may change sign at the interface, the interface is characteristic; this situation leads to nonuniqueness, in particular if the corresponding characteristic field is nonlinear. A criterion has to be added to select the solution considered as admissible.

*Remark 6.1.* In [540] (see also [186], devoted to the scalar case), it is shown that (6.6) is indeed a reasonable way of coupling two scalar conservation laws in the sense that, in meaningful situations, the coupled problem has a unique solution and the associated interface numerical upwind scheme (the so-called two-flux scheme which we introduce below) converges to this solution.

For what concerns systems, it is much easier to study the coupling conditions (6.6) in the case of linear systems, for which we refer to [537].

If general theoretical existence and uniqueness results may be obtained in the scalar case, in particular for the existence of traces [726, 1162], it is beyond the scope of the present work to give a rigorous notion of solution and well-posedness for the general coupling Cauchy problem in the case of systems. We will rather focus on the solution of Riemann problems with coupling which are naturally involved in the numerical approach and for which explicit solutions may be constructed in many cases.  $\square$

When dealing with physical systems, we may prefer to transmit *not* the conservative variables but other *physical* variables. Indeed, define two distinct changes of variables

$$\mathbf{v} \rightarrow \mathbf{u} = \varphi_\alpha(\mathbf{v}), \quad \alpha = L, R$$

from some set  $\Omega_{\mathbf{v}} \subset \mathbb{R}^p$  onto  $\Omega$  such that  $\varphi'_\alpha(\mathbf{v})$  is an isomorphism of  $\mathbb{R}^p$ . Then if  $\mathbf{c}$  is a given boundary *physical* data, setting  $\mathbf{b}_\alpha = \varphi_\alpha(\mathbf{c})$ , we define

$$\begin{aligned} \mathcal{O}_L(\mathbf{b}_L) &= \{\mathbf{w} = \mathbf{W}_L(0-; \mathbf{u}_\ell, \mathbf{b}_L); \mathbf{u}_\ell \in \Omega\}, \\ \mathcal{O}_R(\mathbf{b}_R) &= \{\mathbf{w} = \mathbf{W}_R(0+; \mathbf{b}_R, \mathbf{u}_r); \mathbf{u}_r \in \Omega\} \end{aligned}$$

which are admissible boundary conditions for the systems (6.1) and (6.2), respectively. Thus we now require

$$\begin{cases} \mathbf{u}(0-, t) \in \mathcal{O}_L(\varphi_L(\mathbf{v}(0+, t))), \\ \mathbf{u}(0+, t) \in \mathcal{O}_R(\varphi_R(\mathbf{v}(0-, t))). \end{cases} \quad (6.8)$$

Since  $\varphi_L(\mathbf{v}(0+, t)) \neq \varphi_R(\mathbf{v}(0+, t)) = \mathbf{u}(0+, t)$ , the boundary sets in (6.8) and (6.6) are distinct. Conditions (6.8) will ensure the transmission of *physical* variables and whenever possible their continuity instead of (6.7)

$$\mathbf{v}(0-, t) = \mathbf{v}(0+, t). \quad (6.9)$$

In some noncharacteristic cases, when the Jacobian matrices  $\mathbf{f}'_\alpha(\mathbf{u})$  are invertible, the flux can be chosen as transmitted variable. Then, the model

can be written in a conservative way (6.3), and the coupling condition writes as (6.4). This is the case for the  $p$ -system which we study below.

Last, we may consider another kind of coupling condition with a singular source term

$$\begin{cases} \partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}, x) = \mathcal{M} \delta_0, & x \in \mathbb{R}, t > 0, \\ \mathbf{f}(\mathbf{u}, x) = \begin{cases} \mathbf{f}_L, & x < 0, \\ \mathbf{f}_R, & x > 0, \end{cases} \end{cases} \quad (6.10)$$

for some weight  $\mathcal{M}$  of the vector valued Dirac measure  $\delta_0$ . Then a weak solution satisfies

$$\mathbf{f}_R(\mathbf{u}(0+, t)) - \mathbf{f}_L(\mathbf{u}(0-, t)) = \mathcal{M}(t). \quad (6.11)$$

This weight results from the chosen coupling condition. For instance, for the conservative coupling  $\mathcal{M} = 0$ , if the coupling condition is (6.6) and if (6.7) holds, then  $\mathcal{M} = \mathbf{f}_R(\mathbf{u}(0+, t)) - \mathbf{f}_L(\mathbf{u}(0+, t)) = \mathbf{f}_R(\mathbf{u}(0-, t)) - \mathbf{f}_L(\mathbf{u}(0-, t))$ . In some situations,  $\mathcal{M}$  may be given in order to impose some particular relation at the interface (see [347]). It also appears in a simple model of coupling a fluid with a pointwise particle ([49], [23] and references therein).

Note that (6.10) can be written equivalently

$$\begin{cases} \partial_t \mathbf{u} + \partial_x \mathbf{F}(\mathbf{u}, H) - \mathcal{M}(t) \partial_x H = \mathbf{0}, & x \in \mathbb{R}, t > 0, \\ \partial_t H = 0, \end{cases} \quad (6.12)$$

where  $H(x, t) = H(x)$  is the Heaviside function and  $\mathbf{F}(\mathbf{u}, \varphi) = (1 - \varphi)\mathbf{f}_L(\mathbf{u}) + \varphi\mathbf{f}_R(\mathbf{u})$ . Thus, we get as in Sect. 2 (see (2.2)) a nonconservative system in variables  $(\mathbf{u}, H)$ , for which the initial data is  $\mathbf{u}_0$  for  $\mathbf{u}$ , augmented by  $H(x, 0) = H(x)$ . System (6.12) writes in quasilinear form with matrix

$$M(\mathbf{u}, H) = \begin{pmatrix} \partial_{\mathbf{u}} \mathbf{F} & \mathbf{f}_R - \mathbf{f}_L - \mathcal{M}(t) \\ 0 & 0 \end{pmatrix}$$

with  $\partial_{\mathbf{u}} \mathbf{F}(\mathbf{u}, H) = (1 - H)\mathbf{f}'_L(\mathbf{u}) + H\mathbf{f}'_R(\mathbf{u})$ . This shows clearly that the presence of the measure source term adds a stationary discontinuity at  $x = 0$  to the initial waves which are associated with the eigenvalues of  $\mathbf{f}'_{L/R}(\mathbf{u})$ . The system is resonant if a genuinely nonlinear field associated with the waves of each model  $\mathbf{f}_\alpha$  superimposes this discontinuity. This resonance phenomenon brings nonuniqueness. As already written, in the scalar case, a general framework exists [45]. For systems, a general theory is still out of scope. However, in practical situations, the coupling is performed so as to avoid resonance (for instance, assuming the flow is subsonic in the applications to gas dynamics [347]).

### 6.2.2 Coupled Riemann Problem (CRP)

As for a conservative system with continuous flux, we are interested in some particular solutions, associated with a Riemann data.

*Definition 6.1*

Given a coupling condition, either (6.6) or (6.8) for some specified change of variables  $\varphi_\alpha$ , the coupled Riemann problem (CRP) is the Cauchy problem for (6.1)(6.2) associated with the Riemann data (6.5) and this coupling condition.

For system (6.10), the stationary waves satisfy

$$\partial_x \mathbf{f}(\mathbf{u}, x) = \mathcal{M} \delta_0$$

or

$$\mathbf{f}_R(\mathbf{u}(0+)) - \mathbf{f}_L(\mathbf{u}(0-)) = \mathcal{M}. \quad (6.13)$$

Since we have emphasized the presence of a stationary discontinuity, it is natural that we select among the solutions of the CRP (coupled Riemann problem), the stationary solutions which deserve indeed a particular interest.

*Definition 6.2*

Let  $\mathbf{u}_L, \mathbf{u}_R$  be two constant states belonging to the space of states  $\Omega$  such that the function defined for  $t > 0$  by

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

is a stationary solution of the CRP. Then it is called an equilibrium or elementary solution for the coupling problem (6.10)(6.11).

Note that  $\mathcal{M}$  in (6.13) is then a constant weight such that

$$\mathbf{f}_R(\mathbf{u}_R) - \mathbf{f}_L(\mathbf{u}_L) = \mathcal{M}. \quad (6.14)$$

The stationary solutions of the coupled Riemann problem are an important tool in the study of solutions to (scalar) conservation laws with discontinuous flux. Indeed, they play the role which the constants play in the classical situation of a smooth flux (see [45] and references therein) and can thus be used in Kruzhkov-type entropies (see also [928]). These *elementary solutions* and the set of such states  $u_L, u_R$  satisfying  $f_L(u_L) = f_R(u_R)$  give birth to the notion of *germs*. The theory for systems with discontinuous flux is not as complete. However, the above definition is a natural extension to systems with a measure source term (hence to the nonconservative problem (6.12)). In the approach of numerical interface coupling developed below, we will

require our scheme to be *well-balanced* in the sense that it should preserve these equilibria.

### 6.2.3 Interface Coupling of Two $p$ -Systems

Let us illustrate two different coupling conditions on a simple example. We consider the coupling of two  $p$ -systems, i.e., (6.1)(6.2) with

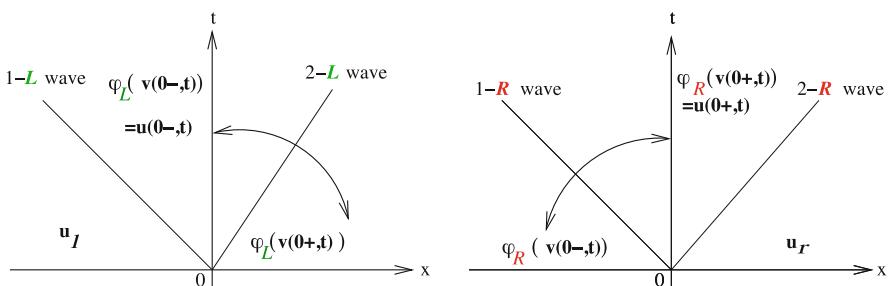
$$\begin{cases} \mathbf{u} = (\tau, u)^T, \tau > 0 \\ \mathbf{f}_\alpha(\mathbf{u}) = (-u, p)^T, p = p_\alpha(\tau), \alpha = L, R \end{cases} \quad (6.15)$$

We assume that  $p'_\alpha < 0, p''_\alpha > 0, \alpha = L, R$ . The two systems differ by the pressure law  $p$ . An important feature is that the signs of the two eigenvalues do not depend on  $\mathbf{u}$ :  $\lambda_1(\mathbf{u}) = -c < 0 < \lambda_2(\mathbf{u}) = c, c = \sqrt{-p'(\tau)}$ . Hence, in the left (resp. right) half plane, there can be only 1-waves (resp. 2-waves).

We first transmit the conservative variables  $(\tau, u)$ . We denote by  $\mathcal{C}_\alpha^k(\mathbf{u}_-)$ ,  $k = 1, 2, \alpha = L, R$  the  $k\alpha$ -wave curve, i.e., the set of states that can be connected to a given state  $\mathbf{u}_-$  by a  $k$ -wave,  $k = 1, 2$  (either rarefaction or admissible shock) relative to the  $p$ -system with flux  $\mathbf{f}_\alpha, \alpha = L, R$ . Expressing the coupling conditions (6.6) gives (for the left condition  $\mathbf{u}(0-, t) \in \mathcal{O}_L(\mathbf{u}(0+, t))$ ) that  $\mathbf{u}(0-, t)$  is connected to  $\mathbf{u}(0+, t)$  by a  $2L$ -wave which means  $\mathbf{u}(0+, t) \in \mathcal{C}_L^2(\mathbf{u}(0-))$  (we use shortened notation for  $\mathcal{C}_L^2(\mathbf{u}(0-, t))$ ) and similarly (for the right condition) by a  $1R$ -wave (see Fig. 6.1). Thus,  $\mathbf{u}(0+) \in \mathcal{C}_L^2(\mathbf{u}(0-)) \cap \mathcal{C}_R^1(\mathbf{u}(0-))$ , and since we have seen that the two wave curves intersect at only one point in the plane  $(\tau, u)$ , at least away from vacuum (see Chap. II, Sect. 7.2), it yields (6.7)

$$\mathbf{u}(0+, t) = \mathbf{u}(0-, t).$$

Let us consider another natural way of coupling the two systems, with transmission of the set of variables  $(u, p)$ . Indeed, the I.B.V.P.'s in both quarter planes ( $x < 0, t > 0$ ) and ( $x > 0, t > 0$ ) are also well posed if one wishes



**Fig. 6.1** The coupling conditions (6.6) for the  $p$ -system

to prescribe a given velocity and pressure  $(u, p)$  on  $x = 0$  in a weak sense, according to (6.8). Indeed, by assumption  $p'_\alpha < 0$ ; hence, we can define its inverse mapping  $\tau_\alpha(p)$  for  $\alpha = L, R$ . Setting  $\mathbf{v} = (v, p)^T$ , we have an admissible change of variables:  $\mathbf{u} = \varphi_\alpha(\mathbf{v})$  where

$$(u, p) \rightarrow \varphi_\alpha(u, p) \equiv (\tau, u) \quad (6.16)$$

is simply defined by  $\tau = \tau_\alpha(p)$  (for instance, if  $p_\alpha(\tau) = \tau^{-\gamma_\alpha}$ , then  $\tau_\alpha(p) = p^{-1/\gamma_\alpha}$ ).

*Proposition 6.1*

For the systems (6.15), the coupling conditions (6.8) with  $\mathbf{v} = (u, p)$  are equivalent to

$$\begin{cases} u(0-, t) = u(0+, t), \\ p(0-, t) = p(0+, t). \end{cases}$$

Moreover, the solution of the coupled Riemann problem (6.15), (6.5), (6.8) exists and is unique.

*Proof.* Let us express the coupling condition (6.8):  $\mathbf{u}(0+) \in \mathcal{O}_R(\varphi_R(\mathbf{v}(0-)))$  and  $\mathbf{u}(0-) \in \mathcal{O}_L(\varphi_L(\mathbf{v}(0+)))$ , with precisely  $\mathbf{v}(0\pm, t) = (u(0\pm), p(0\pm, t))^T$  and  $\mathbf{u}(0-, t) = \varphi_L(\mathbf{v}(0-, t))$ ,  $\mathbf{u}(0+, t) = \varphi_R(\mathbf{v}(0+, t))$ .

First  $\mathbf{u}(0+) \in \mathcal{O}_R(\varphi_R(\mathbf{v}(0-)))$  yields that  $\varphi_R(\mathbf{v}(0-))$  is connected to  $\mathbf{u}(0+) = \varphi_R(\mathbf{v}(0+))$  by a 1R-wave. The idea is that we can parametrize the wave curves by  $p$  and represent them in the  $(u, p)$ -plane. If the 1R-wave curve is  $\mathcal{C}_R^1(\mathbf{u}(0-)) = \{(\tau, u); u = \Psi_{1,R}(\tau)\}$ , let

$$\begin{aligned} \tilde{\mathcal{C}}_R^1(\mathbf{v}(0-)) &= \{(v, p); v = \Psi_{1,R}(\tau_R(p))\} = \{(v, p); \varphi_R(v, p) \in \mathcal{C}_R^1(\mathbf{u}(0-))\} \\ &= \varphi_R^{-1}(\mathcal{C}_R^1(\mathbf{u}(0-))) \end{aligned}$$

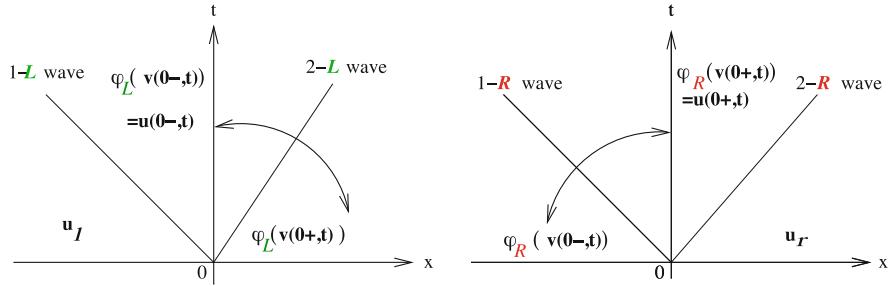
be its representation in the  $(u, p)$ -coordinates; we then have  $\mathbf{v}(0+, t) \in \tilde{\mathcal{C}}_R^1(\mathbf{v}(0-))$  (see Fig. 6.2).

Similarly,  $\mathbf{u}(0-) \in \mathcal{O}_L(\varphi_L(\mathbf{v}(0+)))$  yields that  $\mathbf{u}(0-) = \varphi_L(\mathbf{v}(0-))$  is connected to  $\varphi_L(\mathbf{v}(0+))$  by a 2L-wave. We parametrize the wave curves by  $p$  and represent them onto the  $(u, p)$ -plane. If the 2L-wave curve is  $\mathcal{C}_L^2(\mathbf{u}(0-)) = \{(\tau, u); v = \Psi_{2,L}(\tau)\}$ , let

$$\begin{aligned} \tilde{\mathcal{C}}_L^2(\mathbf{v}(0-)) &= \{(v, p); v = \Psi_{2,L}(\tau_L(p))\} = \{(v, p); \varphi_L(v, p) \in \mathcal{C}_L^2(\mathbf{u}(0-))\} \\ &= \varphi_L^{-1}(\mathcal{C}_L^2(\mathbf{u}(0-))) \end{aligned}$$

be its representation in the  $(u, p)$ -coordinates; we then have  $\mathbf{v}(0+, t) \in \tilde{\mathcal{C}}_L^2(\mathbf{v}(0-))$ .

We have  $\mathbf{v}(0+) \in \tilde{\mathcal{C}}_R^1(\mathbf{v}(0-)) \cap \tilde{\mathcal{C}}_L^2(\mathbf{v}(0-))$  thus  $\mathbf{v}(0+, t) = \mathbf{v}(0-, t)$  because it is easy to prove that the two curves intersect at only one point in the plane  $(u, p)$ . Hence, we do have continuity of  $u, p$ , not of  $\tau$  since  $\tau(0+) = p(0+)^{-1/\gamma_R} \neq p(0-)^{-1/\gamma_L} = \tau(0-)$ .



**Fig. 6.2** The coupling conditions (6.8) for the  $p$ -system

Under the assumptions made on the  $p_\alpha$ 's, existence and uniqueness of the solution of the coupled Riemann problem, away from vacuum, follow as in the usual noncoupled case.  $\square$

In this simple case, the state coupling condition with transmission of  $\mathbf{v} = (u, p)$  is exactly equivalent to conservative coupling. It happens that the flux can be chosen as a transmitted variable. This is possible in the noncharacteristic case when the interface is noncharacteristic: no eigenvalue of the Jacobian matrices  $\mathbf{f}'_\alpha(\mathbf{u})$  can vanish; the matrices are invertible and induce an admissible change of variables on each side.

*Remark 6.2.* Assume that the pressure laws are given as function of a parameter  $\lambda$ :  $p_\alpha(\tau) = p(\tau, \lambda_\alpha)$ . Then, system (6.15) can be written as a system with discontinuous flux

$$\begin{cases} \partial_t \tau - \partial_x u = 0 \\ \partial_t u + \partial_x p(\tau, \lambda) = 0 \\ \partial_t \lambda = 0 \end{cases} \quad (6.17)$$

with a Riemann data for  $\lambda$

$$\lambda_0(x) = \begin{cases} \lambda_L, & x < 0 \\ \lambda_R, & x > 0. \end{cases} \quad (6.18)$$

The parameter  $\lambda$  can be an adiabatic exponent  $\gamma$ , and (6.17) models the flow of a fluid mixture with two components; it can be a mass fraction  $Y$  of one component or a mass fraction of one phase (say vapor) in a fluid for a two-phase flow, then  $\lambda \in [0, 1]$ , and  $\lambda = 0$  or  $\lambda = 1$  characterizes the pure phase. For an isothermal pressure law,  $p(\tau) = \frac{a^2(\lambda)}{\tau}$  the speed of sound  $a(\lambda)$  depends on the composition of the fluid. There are existence results for a (weak) entropy solution to the Cauchy problem for (6.17) (6.18); see [30, 31] and references therein.  $\square$

*Remark 6.3.* For the full Euler system in Lagrangian coordinates, the interface is characteristic; however, it corresponds to a linearly degenerate field

(one eigenvalue is 0); thus, the resonance phenomena does not occur. For general data, the coupling does not yield the continuity (6.7) nor (6.9). However, since we still have the property that on both sides one eigenvalue is negative ( $< 0$ ), one is positive ( $> 0$ ), it yields the continuity of a subset of two variables out of three. If one chooses the transmission of the primitive variable  $(\tau, u, p)$ , it yields the continuity of  $u, p$ . Indeed, the solution of the Riemann problem is obtained by projection of the wave curves on the plane  $u, p$ ; these curves intersect at a unique point, which corresponds to a unique velocity and pressure, but two densities since they are reconstructed by two different laws; we have a material contact discontinuity. We refer to [38] for details and extension to general fluid systems in Lagrangian coordinates.

The case of Euler system in Eulerian coordinates is much more complicated since the eigenvalues corresponding to GNL fields may change sign, which means that the interface may become characteristic and the system (6.12) may be resonant. We refer to [274].

In this context of the full Euler system with energy, if we write the problem in a muticompoment way similar to (6.17) (6.18), numerical schemes have been derived by [7, 680, 1010, 1011]; see also [240]; we refer to [10] for a full description; see also [559] and [81]. These schemes introduced the idea of a nonconservative treatment to reduce pressure oscillations at a contact discontinuity, by transporting the quantity  $\frac{1}{\gamma-1}$ ,  $\gamma = \frac{YC_{p1}+(1-Y)C_{p2}}{YC_{v1}+(1-Y)C_{v2}}$  at the contact velocity which is supposed to be constant, instead of computing this quantity as an average value. This kind of scheme can be understood in the framework of interface coupling by transmission of a well-chosen set of variables (recall that in Lagrangian coordinates, the contact discontinuity is stationary and  $\gamma$  is discontinuous only across this material interface).  $\square$

### 6.2.4 Coupling Systems of Different Sizes

We can think of coupling systems of different dimensions provided there is some kind of hierarchy between them. Let  $\Omega_\alpha \subset \mathbb{R}^{p_\alpha}$ , be the sets of states, and let  $\mathbf{f}_\alpha, \alpha = L, R$ , be two smooth functions from  $\Omega_\alpha$  into  $\mathbb{R}^{p_\alpha}$  such that the corresponding systems are hyperbolic (the notations may be misleading: in the following lines,  $p_\alpha$  denotes a positive integer, not a pressure). There is no loss in generality in assuming  $p_L < p_R$ . We then note  $\mathbf{u} \in \Omega_L$  and  $\mathbf{U} \in \Omega_R$  the respective conservative variables. We also assume that there exists some *lift* operator  $\Pi_{L \rightarrow R}$  from  $\mathbb{R}^{p_L}$  to  $\mathbb{R}^{p_R}$  and a projection operator  $\Pi_{R \rightarrow L}$  from  $\mathbb{R}^{p_R}$  to  $\mathbb{R}^{p_L}$ . There are three significative examples.

*Example 6.1.* Coupling the  $p$ -system with  $p_L = 2$  and the full Euler system with  $p_R = 3$ . If we consider that the first one corresponds to the isentropic case, it is possible to define  $\mathbf{U} = \Pi_{L \rightarrow R}(\mathbf{u}) = (\mathbf{u}, s_0)$ , for some given constant

entropy  $s_0$ . This example illustrates a situation where one model may be considered as coarser than the other, with the knowledge that the approximation may be valid for some flow regimes (for instance, one assumes a reversible adiabatic process for isentropic flow); see [38] for details.  $\square$

*Example 6.2.* Coupling the 1D Euler system with  $p_L = 3$  and the 2D Euler system  $p_R = 4$ , for this example, we refer to the work [607]. This example illustrates a situation where some (geometrical) symmetry assumption may be taken into account in a part of the computational domain. It may be useful to perform this kind of simplification for the simulation of a complex flow ([607] cites the coolant circuit of a pressurized water reactor, in a nuclear power plant). A full 3D solver may be irrelevant due to high computational costs, when simulating the flow by a 1D model may be sufficiently accurate in some parts, for instance, in pipes. The approach is also developed in [404] for the simulation of cavitating fluid flow (liquid and vapor) in complex hydraulic systems (e.g., diesel injectors). The coupling of 1D and 2D shallow-water models is studied in [471].  $\square$

*Example 6.3.* Coupling a relaxation system and its equilibrium limit. In physical examples, for instance, two-phase flow modeling, the relaxation system takes into account some differences between quantities (pressure, temperature, velocity, Gibbs potential) for a system which is not at equilibrium (there may be different kinds of equilibria, thermodynamical, thermal, mechanical) (see [167, 487, 602]).

If there is some link with the previous example, which can indeed be put in a relaxation setting, the relation between the two systems (relaxation and equilibrium) does not reduce in general to a simple assumption of setting one quantity constant (see [39] and [36] for an example of a drift-flux asymptotic limit model obtained from a barotropic two-phase two-pressure model).

The theoretical background of relaxation systems is well developed, and many results are available (in particular since the pioneer work [291]). A simple example of this situation is the coupling of a relaxation law, left (in  $x < 0$ )

$$\partial_t u + \partial_x f(u) = 0, \quad (6.19)$$

with right (in  $x > 0$ ) the linear Jin-Xin relaxation system ( $p_R = 2$ ) (see Chap. IV, Sect. 8.2.2, system (8.13))

$$\begin{cases} \partial_t u + \partial_x v = 0 \\ \partial_t v + a^2 \partial_x u = \frac{1}{\varepsilon}(f(u) - v) \end{cases} \quad (6.20)$$

with  $\varepsilon > 0$ . In this context, theoretical results are available in the noncharacteristic case, when either  $f' > 0$  or  $f' < 0$ . We refer to the work in [222],

where one can find the existence and uniqueness of the solution of the coupled problem. More recently, in [353], the relaxation parameter  $\varepsilon(x)$  depends on  $x$  and is defined by

$$\varepsilon(x) = \begin{cases} 1, & x < 0 \\ \frac{1}{\varepsilon}, & x > 0 \end{cases} \quad (6.21)$$

with  $\varepsilon \ll 1$ .  $\square$

The problem is proved to be well-posed for any given  $\varepsilon > 0$  and the existence and uniqueness (in classical functional spaces) of a solution  $u, v$  to the limit problem as  $\varepsilon \rightarrow 0$  which exactly writes as the coupled problem in the sense which has been defined above (see also [665]).

*Remark 6.4.* Let us also mention a classical coupling which we will not consider, the kinetic/fluid coupling [401]; the numerical coupling of Boltzmann and Navier-Stokes equations is studied in [750]; see also [698]. A rigorous derivation of the coupling of a kinetic equation and Burgers' equation is given in [1164].

Then we mention in the context of traffic flow the coupling of microscopic and macroscopic models [740].  $\square$

### 6.3 Numerical Coupling

For the numerical procedure, we use a finite volume method. Let  $\Delta x$ ,  $\Delta t$ , denote the uniform space and time steps; we set  $\mu = \Delta t / \Delta x$ ,  $t_n = n \Delta t$ ,  $n \in \mathbb{N}$ . In order that  $x = 0$  be an interface between two cells, we change a little our notations and consider the cells  $C_{j+1/2} = (x_j, x_{j+1})$ , with center  $x_{j+1/2} = (j + 1/2) \Delta x$ ,  $j \in \mathbb{Z}$ . The initial condition is discretized as usual by

$$\mathbf{u}_{j+1/2}^0 = \frac{1}{\Delta x} \int_{C_{j+1/2}} \mathbf{u}_0(x) dx, \quad j \in \mathbb{Z}.$$

For the numerical coupling, we are given two numerical fluxes  $\mathbf{g}_L$ ,  $\mathbf{g}_R$  ( $\mathbf{g}_\alpha$  is consistent with  $\mathbf{f}_\alpha$ ) corresponding to 3-point schemes. We assume these schemes are monotone in the scalar case under some CFL condition. Let us define the scheme for the coupling model by

$$\mathbf{u}_{j-1/2}^{n+1} = \mathbf{u}_{j-1/2}^n - \mu (\mathbf{g}_{L,j}^n - \mathbf{g}_{L,j-1}^n), \quad j \leq 0, n \geq 0, \quad (6.22)$$

$$\mathbf{u}_{j+1/2}^{n+1} = \mathbf{u}_{j+1/2}^n - \mu (\mathbf{g}_{R,j+1}^n - \mathbf{g}_{R,j}^n), \quad j \geq 0, n \geq 0. \quad (6.23)$$

We have  $\mathbf{g}_{\alpha,j}^n = \mathbf{g}_\alpha(\mathbf{u}_{j-1/2}^n, \mathbf{u}_{j+1/2}^n)$ ,  $\alpha = L, j < 0$ ,  $\alpha = R, j > 0$ , and at the interface  $x = 0$ , for  $j = 0$ , we have two fluxes  $\mathbf{g}_{\alpha,0}^n$ , and for the fluxes at this interface, we choose  $\mathbf{g}_{\alpha,0}^n$  according to the coupling procedure. The choice

$$\mathbf{g}_{\alpha,0}^n = \mathbf{g}_\alpha(\mathbf{u}_{-1/2}^n, \mathbf{u}_{1/2}^n), \quad \alpha = L, R$$

corresponds to transmit the conservative variables  $\mathbf{u}$ . Namely, if  $j \geq 0$ , the scheme (6.22) with flux  $\mathbf{g}_R$  consistent with  $\mathbf{f}_R$  approximates the I.B.V.P. (6.2) with initial condition  $u(x, 0) = u_0(x), x > 0$ , and for boundary condition at  $x = 0$ , the scheme takes  $\mathbf{u}_{-1/2}^n$ . Since  $\mathbf{g}_{L,0}^n \neq \mathbf{g}_{R,0}^n$ , it is a nonconservative numerical approach, as for the continuous problem. For example, the flux at the boundary with Godunov's scheme is  $\mathbf{g}_{R,0}^n = \mathbf{f}_R(\mathbf{W}_R(0+; \mathbf{u}_{-1/2}^n, \mathbf{u}_{1/2}^n))$ .

*Remark 6.5.* The boundary data is taken into account by the interface numerical flux. For a monotone scheme, there are convergence results in the scalar case for an I.B.V.P. [770]. Then, it is proved (cf. [540]) that, in a number of practical situations, scheme (6.18)–(6.22) converges toward a solution of (6.1)–(6.2) satisfying (6.6).  $\square$

We can also transmit another set of variables  $\mathbf{v}$  by choosing

$$\mathbf{g}_{L,0}^n = \mathbf{g}_L(\mathbf{u}_{-1/2}^n, \varphi_L(\mathbf{v}_{1/2}^n)), \quad \mathbf{g}_{R,0}^n = \mathbf{g}_R(\varphi_R(\mathbf{v}_{-1/2}^n), \mathbf{u}_{1/2}^n)$$

where we define reconstructed states on each side of the interface:  $\mathbf{v}_{1/2}^n = \varphi_R^{-1}(\mathbf{u}_{1/2}^n), \mathbf{v}_{-1/2}^n = \varphi_L^{-1}(\mathbf{u}_{-1/2}^n)$ . This means that the states at the interface are first characterized by their primitive variables and then are reconstructed as conservative states; this reconstruction is necessary in the computation of the flux.

Let us set

$$\mathcal{M}^{(n)} = (\mathbf{g}_R)_0^n - (\mathbf{g}_L)_0^n. \quad (6.24)$$

In case of conservative coupling we have  $\mathcal{M}^{(n)} = 0$ , otherwise  $\mathcal{M}^{(n)}$  appears as a consistent discretization of  $\mathcal{M}(t_n)$ .

*Remark 6.6.* The above approach with two fluxes or reconstructed states can be related to other nonconservative approaches such as the single fluid algorithm introduced in [10] (as already said in Remark 6.3).  $\square$

If we come back to the formulation (6.10), or (6.12), we are led to study the well-balanced character of the above numerical interface coupling procedure, in the spirit of Sect. 3.2.

*Proposition 6.2*

Assume that the fluxes of scheme (6.18)(6.22) are upwind; then the scheme is well-balanced for system (6.10) in the sense that it preserves the equilibria of Definition 6.2.

*Proof.* Let  $\mathbf{u}_\pm$  be a stationary solution of a CRP. Recall that for system (6.10), the stationary waves satisfy (6.13), thus

$$\mathbf{f}_R(\mathbf{u}_+) - \mathbf{f}_L(\mathbf{u}_-) = \mathcal{M}$$

where  $\mathcal{M}$  is defined by the coupling condition. Assume, for instance, that we consider a state coupling condition (6.6) (the case (6.8) is treated in the same way). If we have a stationary solution satisfying the CRP (6.6),

necessarily  $\mathbf{W}_L(0-; \mathbf{u}_-, \mathbf{u}_+) = \mathbf{u}_-$  and  $\mathbf{W}_R(0+; \mathbf{u}_-, \mathbf{u}_+) = \mathbf{u}_+$ . This implies that there is no left going wave associated with  $\mathbf{f}_L$  in  $x < 0$  (the eigenvalues of  $\mathbf{f}'_L(\mathbf{u}_-)$  are positive  $\geq 0$ ) and no right going wave in  $x < 0$  associated with  $\mathbf{f}_R$  (the eigenvalues of  $\mathbf{f}'_R(\mathbf{u}_+)$  are negative  $\leq 0$ ). If the numerical fluxes  $\mathbf{g}_\alpha$  are upwind,  $\mathbf{g}_L(\mathbf{u}_-, \mathbf{u}_+) = \mathbf{f}_L(\mathbf{u}_-)$ ,  $\mathbf{g}_R(\mathbf{u}_-, \mathbf{u}_+) = \mathbf{f}_R(\mathbf{u}_+)$ .

Then set  $\mathbf{u}_{-1/2}^0 = \mathbf{u}_-$ ,  $\mathbf{u}_{1/2}^0 = \mathbf{u}_+$ . By consistency of both numerical fluxes, we see from (6.18)(6.22) that  $\mathbf{u}_{-1/2}^n = \mathbf{u}_-$ ,  $\mathbf{u}_{1/2}^n = \mathbf{u}_+$  if and only if the numerical fluxes satisfy

$$\mathbf{g}_{L,0}^0 \equiv \mathbf{g}_L(\mathbf{u}_-, \mathbf{u}_+) = \mathbf{f}_L(\mathbf{u}_-), \quad \mathbf{g}_{R,0}^0 \equiv \mathbf{g}_R(\mathbf{u}_-, \mathbf{u}_+) = \mathbf{f}_R(\mathbf{u}_+)$$

or  $\mathbf{g}_{L,0}^0 = \mathbf{f}_L(\mathbf{W}_L(0-; \mathbf{u}_-, \mathbf{u}_+))$ ,  $\mathbf{g}_{R,0}^0 = \mathbf{f}_R(\mathbf{W}_R(0+; \mathbf{u}_-, \mathbf{u}_+))$  which is satisfied by the upwind flux.  $\square$

Note that for a conservative system with continuous flux, constants are elementary solutions. A conservative scheme with consistent flux preserves constant states; one simply imposes that there is only one numerical flux at each interface and that this flux coincides with the exact flux if the states on each side are equal. For a discontinuous flux (equivalently for an interface coupling problem), the equilibria play the role of constants.

In [347] an approximate Riemann solver is proposed for the numerical coupling of two Euler systems (with energy), coupled by a measure source term. The solver is based on a relaxation approach, it is designed in order to preserve equilibrium solutions of the coupled problem. Following the method described in Chap. IV, Sect. 8.3.4, the relaxation scheme involves a  $4 \times 4$  relaxation system, so that the source term is also completed by a 4th component which is evaluated in order that the scheme be well-balanced.

## Notes

In this chapter we have chosen to detail a few illustrative examples of systems with source terms and some of the recent numerical treatments that they have received. There are other natural examples [71]. The choice which has been done concerns methods for which the properties can be understood from a continuous PDE point of view; it is thus far from being exhaustive, all the most since there has been a huge literature published on the subject during the last few years (for instance, [448], for stiff source terms; involving GRP [875], etc.). We have tried to emphasize the arguments which make these methods rather general, even if the analysis is done on particular examples, so as to give the reader some tools to address other situations or understand other approaches.

Let us mention that there are several interesting textbooks which address the question of source terms, with different viewpoints which make them all interesting and complementary, in particular those of R. J. Leveque [777] and

F. Bouchut [163] and the more recent textbook of L. Gosse [553]; one can find much more precise results than stated in the previous sections. Let us also mention a few special issues on the subject, in link with conferences or summer schools, [9, 202, 960]. Moreover, for the particular example of the shallow water system with bottom topography, there are plenty of recent publications, but only a few of them have been referred to.

For a review concerning AP schemes for multiscale kinetic and hyperbolic equations, see [661]. A recent paper addresses the weak consistency of (an analog of the Lax-Wendroff theorem) finite volume schemes for balance laws in the multidimensional case and under minimal regularity assumptions for the mesh [505].

# References

- [1] R. Abgrall, *Généralisation du schéma de Roe pour le calcul d'écoulements de gaz à concentrations variables* (La Recherche Aérospatiale, 1988), pp. 31–43
- [2] R. Abgrall, Preliminary results on an extension of Roe's approximate Riemann solver to nonequilibrium flows, INRIA Research Report 987, INRIA Rocquencourt, Le Chesnay, France (1989)
- [3] R. Abgrall, A genuinely multidimensional Riemann solver, INRIA Research Report 1859, INRIA Rocquencourt , Le Chesnay, France (1993)
- [4] R. Abgrall, Approximation du problème de Riemann vraiment multidimensionnel des équations d'Euler par une méthode de type Roe. I. La linéarisation. C. R. Acad. Sci. Paris Sér. I Math. **319**, 499–504 (1994)
- [5] R. Abgrall, Approximation du problème de Riemann vraiment multidimensionnel des équations d'Euler par une méthode de type Roe. II. Solution du problème de Riemann approché. C. R. Acad. Sci. Paris Sér. I Math. **319**, 625–629 (1994)
- [6] R. Abgrall, On essentially non-oscillatory schemes on unstructured meshes: analysis and implementation. J. Comput. Phys. **114**, 45–58 (1994)
- [7] R. Abgrall, How to prevent pressure oscillations in multicomponent flow calculations: a quasi-conservative approach. J. Comput. Phys. **125**, 150–160 (1996)
- [8] R. Abgrall, Toward the ultimate conservative scheme: following the quest. J. Comput. Phys. **167**, 277–315 (2001)
- [9] R. Abgrall, L. Fezoui, J. Talendier, An extension of Osher's Riemann solver for chemical and vibrational non-equilibrium gas flows. Int. J. Numer. Methods Fluids **14**(8), 935–960 (1992)
- [10] R. Abgrall, S. Karni, Computations of compressible multifluids. J. Comput. Phys. **169**, 594–623 (2001)

- [11] R. Abgrall, S. Karni, Two-layer shallow water system: a relaxation approach. *SIAM J. Sci. Comput.* **31**, 1603–1627 (2009)
- [12] R. Abgrall, S. Karni, A comment on the computation of non-conservative products. *J. Comput. Phys.* **229**, 2759–2763 (2010)
- [13] R. Abgrall, S. Mishra, Uncertainty quantification for hyperbolic systems of conservation laws, in *Handbook of Numerical Methods for Hyperbolic Problems*. Handb. Numer. Anal., vol. 18 (Elsevier/North-Holland, Amsterdam, 2017), pp. 507–544
- [14] R. Abgrall, J.-L. Montagné, *Generalization of the Osher Scheme for Calculating Flows of Mixed Gases of Variable Concentrations, and of Real Gases* (Rech. Aérospat., 1989), pp. 1–13
- [15] R. Abgrall, B. Nkonga, R. Saurel, Efficient numerical approximation of compressible multi-material flow for unstructured meshes. *Comput. Fluids* **32**, 571–605 (2003)
- [16] R. Abgrall, V. Perrier, Asymptotic expansion of a multiscale numerical scheme for compressible multiphase flow. *Multiscale Model. Simul.* **5**, 84–115 (2006)
- [17] R. Abgrall, P.L. Roe, High order fluctuation schemes on triangular meshes. *J. Sci. Comput.* **19**, 3–36 (2003)
- [18] R. Abgrall, R. Saurel, Discrete equations for physical and numerical compressible multiphase mixtures. *J. Comput. Phys.* **186**, 361–396 (2003)
- [19] R. Abgrall, C.-W. Shu, eds., *Handbook of Numerical Methods for Hyperbolic Problems*. Handbook of Numerical Analysis, vol. 17 (Elsevier/North-Holland, Amsterdam, 2016). Basic and fundamental issues
- [20] R. Abgrall, C.-W. Shu, eds., *Handbook of Numerical Methods for Hyperbolic Problems*. Handbook of Numerical Analysis, vol. 18 (Elsevier/North-Holland, Amsterdam, 2017). Applied and modern issues
- [21] M. Adamczewski, J.-F. Colombeau, A.Y. LeRoux, Convergence of numerical schemes involving powers of the Dirac delta function. *J. Math. Anal. Appl.* **145**, 172–185 (1990)
- [22] Adimurthi, S. Mishra, G.D.V. Gowda, Optimal entropy solutions for conservation laws with discontinuous flux-functions. *J. Hyperbolic Differ. Equ.* **2**, 783–837 (2005)
- [23] N. Aguillon, F. Lagoutière, N. Seguin, Convergence of finite volumes schemes for the coupling between the inviscid Burgers equation and a particle. *Math. Comput.* **86**, 157–196 (2017)
- [24] K. Ajmani, W.-F. Ng, M.-S. Liou, Preconditioned conjugate gradient methods for the navier-stokes equations. *J. Comput. Phys.* **110**, 68–81 (1994)
- [25] F. Alcrudo, F. Benkhaldoun, Exact solutions to the Riemann problem of the shallow water equations with a bottom step. *Comput. Fluids* **30**, 643–671 (2001)

- [26] F. Alcrudo, P. Garcia-Navarro, A high resolution Godunov-type scheme in finite volumes for the 2d shallow-water equations. *Int. J. Numer. Methods Fluids* **16**, 489–505 (1993)
- [27] S. Alinhac, Existence d’ondes de raréfaction pour des systèmes quasi-linéaires hyperboliques multidimensionnels. *Commun. Partial Differ. Equ.* **14**, 173–230 (1989)
- [28] G. Allaire, A. Zelmanse, Kinetic schemes for gas dynamics of real gases or two-phase mixtures, in *Numerical Methods in Mechanics* (Concepción, 1995). Pitman Res. Notes Math. Ser., Longman, vol. 371 (Harlow, 1997), pp. 13–24
- [29] F. Alouges, B. Merlet, Approximate shock curves for non-conservative hyperbolic systems in one space dimension. *J. Hyperbolic Differ. Equ.* **1**, 769–788 (2004)
- [30] D. Amadori, P. Baiti, A. Corli, E. Dal Santo, Global weak solutions for a model of two-phase flow with a single interface. *J. Evol. Equ.* **15**, 699–726 (2015)
- [31] D. Amadori, A. Corli, On a model of multiphase flow. *SIAM J. Math. Anal.* **40**, 134–166 (2008)
- [32] D. Amadori, L. Gosse, Transient  $L^1$  error estimates for well-balanced schemes on non-resonant scalar balance laws. *J. Differ. Equ.* **255**, 469–502 (2013)
- [33] D. Amadori, L. Gosse, *Error Estimates for Well-Balanced Schemes on Simple Balance Laws*. SpringerBriefs in Mathematics, vol. 2 (Springer, Berlin, 2015). One-Dimensional Position-Dependent Models
- [34] D. Amadori, L. Gosse, G. Guerra, Godunov-type approximation for a general resonant balance law with large data. *J. Differ. Equ.* **198**, 233–274 (2004)
- [35] L. Ambrosio, B. Alberto, H. Dirk, E. Zuazua, *Modelling and Optimisation of Flows on Networks*, ed. by B. Piccoli, M. Rascle. C.I.M.E. Foundation Subseries (Springer, Berlin, 2013). Cetraro, Italy 2009
- [36] A. Ambroso, C. Chalons, F. Coquel, T. Galié, E. Godlewski, P.-A. Raviart, N. Seguin, The drift-flux asymptotic limit of barotropic two-phase two-pressure models. *Commun. Math. Sci.* **6**, 521–529 (2008)
- [37] A. Ambroso, C. Chalons, F. Coquel, E. Godlewski, F. Lagoutière, P.-A. Raviart, N. Seguin, The coupling of homogeneous models for two-phase flows. *Int. J. Finite* **4**, 39 (2007)
- [38] A. Ambroso, C. Chalons, F. Coquel, E. Godlewski, F. Lagoutière, P.-A. Raviart, N. Seguin, Coupling of general Lagrangian systems. *Math. Comput.* **77**, 909–941 (2008)
- [39] A. Ambroso, C. Chalons, F. Coquel, E. Godlewski, F. Lagoutière, P.-A. Raviart, N. Seguin, Relaxation methods and coupling procedures. *Int. J. Numer. Methods Fluids* **56**, 1123–1129 (2008)
- [40] J. Anderson, *Modern Compressible Flow with Historical Perspective*. McGraw-Hill Series in Mechanical Engineering (McGraw-Hill, New York, 1982)

- [41] W. Anderson, A grid generation and flow solution method for the Euler equations on unstructured grids. *J. Comput. Phys.* **110**, 23–38 (1994)
- [42] B. Andreianov, The semigroup approach to conservation laws with discontinuous flux, in *Hyperbolic Conservation Laws and Related Analysis with Applications*, ed. by G.-Q.G. Chen, H. Holden, K.H. Karlsen. Springer Proceedings in Mathematics & Statistics, 2011, pp. 1–22
- [43] B. Andreianov, P. Bénilan, S.N. Kruzhkov,  $L^1$ -theory of scalar conservation law with continuous flux function. *J. Funct. Anal.* **171**, 15–33 (2000)
- [44] B. Andreianov, M. Karimou Gazibo, Explicit formulation for the Dirichlet problem for parabolic-hyperbolic conservation laws. *Netw. Heterog. Media* **11**, 203–222 (2016)
- [45] B. Andreianov, K.H. Karlsen, N.H. Risebro, A theory of  $L^1$ -dissipative solvers for scalar conservation laws with discontinuous flux. *Arch. Ration. Mech. Anal.* **201**, 27–86 (2011)
- [46] B. Andreianov, K. Sbihi, Scalar conservation laws with nonlinear boundary conditions. *C. R. Math. Acad. Sci. Paris* **345**, 431–434 (2007)
- [47] B. Andreianov, K. Sbihi, Strong boundary traces and well-posedness for scalar conservation laws with dissipative boundary conditions, in *Hyperbolic Problems: Theory, Numerics, Applications* (Springer, Berlin, 2008), pp. 937–945
- [48] B. Andreianov, K. Sbihi, Well-posedness of general boundary-value problems for scalar conservation laws. *Trans. Am. Math. Soc.* **367**, 3763–3806 (2015)
- [49] B. Andreianov, N. Seguin, Analysis of a Burgers equation with singular resonant source term and convergence of well-balanced schemes. *Discrete Contin. Dyn. Syst.* **32**, 1939–1964 (2012)
- [50] N. Andrianov, Performance of numerical methods on the non-unique solution to the Riemann problem for the shallow water equations. *Int. J. Numer. Methods Fluids* **47**, 825–831 (2005)
- [51] N. Andrianov, G. Warnecke, On the solution to the Riemann problem for the compressible duct flow. *SIAM J. Appl. Math.* **64**, 878–901 (2004)
- [52] N. Andrianov, G. Warnecke, The Riemann problem for the Baer-Nunziato two-phase flow model. *J. Comput. Phys.* **195**, 434–464 (2004)
- [53] F. Angrand, V. Boulard, A. Dervieux, J. Périaux, G. Vijayasundaram, Triangular finite element methods for the Euler equations, in *Computing Methods in Applied Sciences and Engineering, VI* (Versailles, 1983) (North-Holland, Amsterdam, 1984), pp. 535–563
- [54] F. Angrand, F.C. Lafon, Flux formulation using a fully 2D approximate Roe Riemann solver, in *Nonlinear Hyperbolic Problems: The-*

- retical, Applied, and Computational Aspects (Taormina, 1992). Notes Numer. Fluid Mech., Friedr., vol. 43 (Vieweg, Braunschweig, 1993), pp. 15–22
- [55] D. Aregba-Driollet, R. Natalini, Discrete kinetic schemes for multidimensional systems of conservation laws. SIAM J. Numer. Anal. **37**, 1973–2004 (2000) (electronic)
- [56] P. Arminjon, A. Dervieux, Construction of TVD-like artificial viscosities on two-dimensional arbitrary FEM grids. J. Comput. Phys. **106**, 176–198 (1993)
- [57] P. Arminjon, A. Dervieux, L. Fézoui, H. Steve, B. Stoufflet, Nonoscillatory schemes for multidimensional Euler calculations with unstructured grids, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24. Friedr. Vieweg, Braunschweig, 1989, pp. 1–10
- [58] P. Arminjon, A. Dervieux, L. Fézoui, H. Steve, B. Stoufflet, Nonoscillatory schemes for multidimensional Euler calculations with unstructured grids, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24. Friedr. Vieweg, Braunschweig, 1989, pp. 1–10
- [59] P. Arminjon, M.-C. Viallon, Généralisation du schéma de Nessyahu-Tadmor pour une équation hyperbolique à deux dimensions d'espace, C. R. Acad. Sci. Paris Sér. I Math. **320**, 85–88 (1995)
- [60] P. Arminjon, M.-C. Viallon, Convergence of a finite volume extension of the Nessyahu-Tadmor scheme on unstructured grids for a two-dimensional linear hyperbolic equation. SIAM J. Numer. Anal. **36**, 738–771 (1999)
- [61] F. Asakura, The initial-boundary value problem for hyperbolic systems of conservation laws, in *Hyperbolic Problems: Theory, Numerics, Applications* (Stony Brook, NY, 1994), (World Sci. Publ., River Edge, 1996), pp. 278–283
- [62] E. Audusse, F. Benkhaldoun, S. Sari, M. Seaid, P. Tassi, A fast finite volume solver for multi-layered shallow water flows with mass exchange. J. Comput. Phys. **272**, 23–45 (2014)
- [63] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, B. Perthame, A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comput. **25**, 2050–2065 (2004)
- [64] E. Audusse, F. Bouchut, M.-O. Bristeau, J. Sainte-Marie, Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system. Math. Comput. **85**, 2815–2837 (2016)
- [65] E. Audusse, M.-O. Bristeau, M. Pelanti, J. Sainte-Marie, Approximation of the hydrostatic Navier-Stokes system for density stratified flows by a multilayer model: kinetic interpretation and numerical solution. J. Comput. Phys. **230**, 3453–3478 (2011)

- [66] E. Audusse, M.-O. Bristeau, B. Perthame, J. Sainte-Marie, A multilayer Saint-Venant system with mass exchanges for shallow water flows. Derivation and numerical validation. *ESAIM Math. Model. Numer. Anal.* **45**, 169–200 (2011)
- [67] E. Audusse, B. Perthame, Uniqueness for scalar conservation laws with discontinuous flux via adapted entropies. *Proc. R. Soc. Edinb. Sect. A* **135**, 253–265 (2005)
- [68] A. Aw, M. Rascle, Resurrection of “second order” models of traffic flow. *SIAM J. Appl. Math.* **60**, 916–938 (2000)
- [69] A.V. Azevedo, D. Marchesin, Multiple viscous profile Riemann solutions in mixed elliptic-hyperbolic models for flow in porous media, in *Nonlinear Evolution Equations That Change Type*. IMA Vol. Math. Appl., vol. 27 (Springer, New York, 1990), pp. 1–17
- [70] F. Bachmann, J. Vovelle, Existence and uniqueness of entropy solution of scalar conservation laws with a flux function involving discontinuous coefficients. *Commun. Partial Differ. Equ.* **31**, 371–395 (2006)
- [71] P. Bagnerini, R.M. Colombo, A. Corli, On the role of source terms in continuum traffic flow models. *Math. Comput. Model.* **44**, 917–930 (2006)
- [72] P. Baiti, A. Bressan, H.K. Jenssen, Instability of travelling wave profiles for the Lax-Friedrichs scheme. *Discrete Contin. Dyn. Syst.* **13**, 877–899 (2005)
- [73] D.S. Bale, R.J. Leveque, S. Mitran, J.A. Rossmanith, A wave propagation method for conservation laws and balance laws with spatially varying flux functions. *SIAM J. Sci. Comput.* **24**, 955–978 (2002)
- [74] J.M. Ball, A version of the fundamental theorem for Young measures, in *PDEs and Continuum Models of Phase Transitions* (Nice, 1988). Lecture Notes in Phys., vol. 344 (Springer, Berlin, 1989), pp. 207–215
- [75] J. Ballmann, R. Jeltsch, eds., *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications*. Notes on Numerical Fluid Mechanics, vol. 24. Friedr. Vieweg & Sohn, Braunschweig, 1989
- [76] D.S. Balsara, Riemann solver for relativistic hydrodynamics. *J. Comput. Phys.* **114**, 284–297 (1994)
- [77] D.S. Balsara, Multidimensional HLLE Riemann solver: application to Euler and magnetohydrodynamic flows. *J. Comput. Phys.* **229**, 1970–1993 (2010)
- [78] D.S. Balsara, A two-dimensional HLLC Riemann solver for conservation laws: application to Euler and magnetohydrodynamic flows. *J. Comput. Phys.* **231**, 7476–7503 (2012)
- [79] D.S. Balsara, M. Dubois, R. Abgrall, Multidimensional HLLC Riemann solver for unstructured meshes—with application to Euler and MHD flows. *J. Comput. Phys.* **261**, 172–208 (2014)
- [80] D.S. Balsara, J. Vides, K. Gurski, B. Nkonga, M. Dubois, S. Garain, E. Audit, A two-dimensional Riemann solver with self-similar substructure—alternative formulation based on least squares projection. *J. Comput. Phys.* **304**, 138–161 (2016)

- [81] J.W. Banks, D.W. Schwendeman, A.K. Kapila, W.D. Henshaw, A high-resolution Godunov method for compressible multi-material flow on overlapping grids. *J. Comput. Phys.* **223**, 262–297 (2007)
- [82] R. Baraille, G. Bourdin, F. Dubois, A.Y. LeRoux, Une version à pas fractionnaires du schéma de Godunov pour l'hydrodynamique. *C. R. Acad. Sci. Paris Sér. I Math.* **314**, 147–152 (1992)
- [83] C. Bardos, Introduction aux problèmes hyperboliques non linéaires, in *Fluid Dynamics* (Varenna, 1982). Lecture Notes in Math., vol. 1047 (Springer, Berlin, 1984), pp. 1–74
- [84] C. Bardos, Different approach for the relation between the kinetic and the macroscopic equations, in *Nonlinear hyperbolic problems* (St. Etienne, 1986). Lecture Notes in Math., vol. 1270 (Springer, Berlin, 1987), pp. 308–323
- [85] C. Bardos, Une interprétation des relations existant entre les équations de Boltzmann, de Navier-Stokes et d'Euler à l'aide de l'entropie. *Mat. Apl. Comput.* **6**, 97–117 (1987)
- [86] C. Bardos, A.Y. LeRoux, J.-C. Nédélec, First order quasilinear equations with boundary conditions. *Commun. Partial Differ. Equ.* **4**, 1017–1034 (1979)
- [87] A.J.C. Barré de Saint-Venant, Théorie du mouvement non permanent des eaux avec applications aux crues des rivières et à l'introduction des marées dans leur lit. *C. R. Acad. Sci. Paris* **73**, 147–154 (1871)
- [88] W. Barsukow, P.V.F. Edelmann, C. Klingenberg, F. Miczek, F.K. Röpke, A numerical scheme for the compressible low-Mach number regime of ideal fluid dynamics. *J. Sci. Comput.* **72**, 623–646 (2017)
- [89] W. Barsukow, P.V.F. Edelmann, C. Klingenberg, F.K. Röpke, A low-Mach Roe-type solver for the Euler equations allowing for gravity source terms, in *LMLFN 2015—Low Velocity Flows—Application to Low Mach and Low Froude Regimes*. ESAIM Proc. Surveys, EDP Sci., vol. 58. Les Ulis, 2017, pp. 27–39
- [90] T. Barth, M. Ohlberger, *Finite Volume Methods: Foundation and Analysis* (Wiley, 2004)
- [91] P. Batten, N. Clarke, C. Lambert, D.M. Causon, On the choice of wavespeeds for the HLLC Riemann solver. *SIAM J. Sci. Comput.* **18**, 1553–1570 (1997)
- [92] M. Baudin, C. Berthon, F. Coquel, R. Masson, Q.H. Tran, A relaxation method for two-phase flow models with hydrodynamic closure law. *Numer. Math.* **99**, 411–440 (2005)
- [93] M. Baum, T. Poinsot, D. Thévenin, Accurate boundary conditions for multicomponent reactive flows. *J. Comput. Phys.* **116**, 247–261 (1995)
- [94] A. Bayliss, E. Turkel, Far field boundary conditions for compressible flows. *J. Comput. Phys.* **48**, 182–199 (1982)
- [95] R.M. Beam, R.F. Warming, An implicit finite-difference algorithm for hyperbolic systems in conservation-law form. *J. Comput. Phys.* **22**, 87–110 (1976)

- [96] R.M. Beam, R.F. Warming, H.C. Yee, Stability analysis of numerical boundary conditions and implicit difference approximations of hyperbolic equations. *J. Comput. Phys.* **48**, 200–222 (1982)
- [97] K. Beauchard, E. Zuazua, Large time asymptotics for partially dissipative hyperbolic systems. *Arch. Ration. Mech. Anal.* **199**, 177–227 (2011)
- [98] A. Beccantini, E. Studer, The reactive Riemann problem for thermally perfect gases at all combustion regimes. *Int. J. Numer. Methods Fluids* **64**, 269–313 (2010)
- [99] J.B. Bell, P. Colella, J.A. Trangenstein, Higher order Godunov methods for general systems of hyperbolic conservation laws. *J. Comput. Phys.* **82**, 362–397 (1989)
- [100] J.B. Bell, C. Dawson, J. Quirk, G. Shubin, An unsplit higher order godunov method for scalar conservation laws in multiple dimension. *J. Comput. Phys.* **106**, 1–24 (1988)
- [101] M. Ben-Artzi, The generalized Riemann problem for reactive flows. *J. Comput. Phys.* **81**, 70–101 (1989)
- [102] M. Ben-Artzi, A. Birman, Computation of reactive duct flows in external fields. *J. Comput. Phys.* **86**, 225–255 (1990)
- [103] M. Ben-Artzi, J. Falcovitz, A second-order Godunov-type scheme for compressible fluid dynamics. *J. Comput. Phys.* **55**, 1–32 (1984)
- [104] M. Ben-Artzi, J. Falcovitz, *Generalized Riemann Problems in Computational Fluid Dynamics*. Cambridge Monographs on Applied and Computational Mathematics, vol. 11. (Cambridge University Press, Cambridge, 2003)
- [105] A. Benabdallah, Le “ $p$  système” dans un intervalle. *C. R. Acad. Sci. Paris Sér. I Math.* **303**, 123–126 (1986)
- [106] A. Benabdallah, D. Serre, Problèmes aux limites pour des systèmes hyperboliques non linéaires de deux équations à une dimension d'espace. *C. R. Acad. Sci. Paris Sér. I Math.* **305**, 677–680 (1987)
- [107] S. Benharbit, Sur la théorie et l'approximation numérique des problèmes hyperboliques non linéaires. Application à la dynamique des gaz compressible, Ph.D. thesis, Université J. Fourier, Grenoble, France (1992)
- [108] S. Benharbit, A. Chalabi, J.-P. Vila, Numerical viscosity and convergence of finite volume methods for conservation laws with boundary conditions. *SIAM J. Numer. Anal.* **32**, 775–796 (1995)
- [109] P. Bénilan, J. Carrillo, P. Wittbold, Renormalized entropy solutions of scalar conservation laws. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **29**, 313–327 (2000)
- [110] P. Bénilan, S. Kružkov, Conservation laws with continuous flux functions. *NoDEA Nonlinear Differ. Equ. Appl.* **3**, 395–419 (1996)
- [111] S. Benzoni-Gavage, D. Serre, Compensated compactness for a class of hyperbolic systems of  $p$  conservation laws with  $p \geq 3$ , in *Progress in Partial Differential Equations: the Metz Surveys 2* (1992). Pitman Res. Notes Math. Ser., vol. 296 (Longman Sci. Tech., Harlow, 1993), pp. 3–11

- [112] S. Benzoni-Gavage, D. Serre, *Multidimensional Hyperbolic Partial Differential Equations: First-Order Systems and Applications*. (The Clarendon Press, Oxford University Press, Oxford, 2007)
- [113] B. Berde, M. Borrel, Comparison of high-order Godunov-type schemes for the Euler equations on irregular meshes, in *Proceedings of the European conference, Eccomas-94*, Stuttgart (1994)
- [114] F. Béreux, Zero-relaxation limit versus operator splitting for two-phase fluid flow computations. *Comput. Methods Appl. Mech. Eng.* **133**, 93–124 (1996)
- [115] F. Béreux, L. Sainsaulieu, Un schéma numérique de type Roe pour les systèmes hyperboliques avec relaxation. *C. R. Acad. Sci. Paris Sér. I Math.* **320**, 379–384 (1995)
- [116] F. Béreux, L. Sainsaulieu, A Roe-type Riemann solver for hyperbolic systems with relaxation based on time-dependent wave decomposition. *Numer. Math.* **77**, 143–185 (1997)
- [117] F. Berger, J.-F. Colombeau, Numerical solutions of one-pressure models in multifluid flows. *SIAM J. Numer. Anal.* **32**, 1139–1154 (1995)
- [118] A. Bermúdez, X. López, M.E. Vázquez-Cendón, Numerical solution of non-isothermal non-adiabatic flow of real gases in pipelines. *J. Comput. Phys.* **323**, 126–148 (2016)
- [119] A. Bermúdez, M.E. Vázquez, Upwind methods for hyperbolic conservation laws with source terms. *Comput. Fluids* **23**, 1049–1071 (1994)
- [120] M. Bernard, S. Dellacherie, G. Faccanoni, B. Grec, Y. Penel, Study of a low Mach nuclear core model for two-phase flows with phase transition I: stiffened gas law. *ESAIM Math. Model. Numer. Anal.* **48**, 1639–1679 (2014)
- [121] A. Bernard-Champmartin, O. Poujade, J. Mathiaud, J.-M. Ghidaglia, Modelling of an homogeneous equilibrium mixture model (HEM). *Acta Appl. Math.* **129**, 1–21 (2014)
- [122] R. Bernetti, V.A. Titarev, E.F. Toro, Exact solution of the Riemann problem for the shallow water equations with discontinuous bottom geometry. *J. Comput. Phys.* **227**, 3212–3243 (2008)
- [123] F. Berthelin, Convergence of flux vector splitting schemes with single entropy inequality for hyperbolic systems of conservation laws. *Numer. Math.* **99**, 585–604 (2005)
- [124] F. Berthelin, F. Bouchut, Relaxation to isentropic gas dynamics for a BGK system with single kinetic entropy. *Methods Appl. Anal.* **9**, 313–327 (2002)
- [125] F. Berthelin, T. Goudon, S. Minjeaud, Kinetic schemes on staggered grids for barotropic Euler models: entropy-stability analysis. *Math. Comput.* **84**, 2221–2262 (2015)
- [126] F. Berthelin, T. Goudon, B. Polizzi, M. Ribot, Asymptotic problems and numerical schemes for traffic flows with unilateral constraints describing the formation of jams. *Netw. Heterog. Media* **12**, 591–617 (2017)

- [127] F. Berthelin, A.E. Tzavaras, A. Vasseur, From discrete velocity Boltzmann equations to gas dynamics before shocks. *J. Stat. Phys.* **135**, 153–173 (2009)
- [128] F. Berthelin, A. Vasseur, From kinetic equations to multidimensional isentropic gas dynamics before shocks. *SIAM J. Math. Anal.* **36**, 1807–1835 (2005) (electronic)
- [129] F. Berthelin, J. Vovelle, A Bhatnagar-Gross-Krook approximation to scalar conservation laws with discontinuous flux. *Proc. R. Soc. Edinb. Sect. A* **140**, 953–972 (2010)
- [130] C. Berthon, Robustness of MUSCL schemes for 2D unstructured meshes. *J. Comput. Phys.* **218**, 495–509 (2006)
- [131] C. Berthon, M. Bessemoulin-Chatard, H. Mathis, Numerical convergence rate for a diffusive limit of hyperbolic systems:  $p$ -system with damping. *SMAI J. Comput. Math.* **2**, 99–119 (2016)
- [132] C. Berthon, C. Chalons, A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. *Math. Comput.* **85**, 1281–1307 (2016)
- [133] C. Berthon, C. Chalons, R. Turpault, Asymptotic-preserving Godunov-type numerical schemes for hyperbolic systems with stiff and nonstiff relaxation terms. *Numer. Methods Partial Differ. Equ.* **29**, 1149–1172 (2013)
- [134] C. Berthon, P. Charrier, B. Dubroca, An HLLC scheme to solve the  $M_1$  model of radiative transfer in two space dimensions. *J. Sci. Comput.* **31**, 347–389 (2007)
- [135] C. Berthon, F. Coquel, P.G. LeFloch, Why many theories of shock waves are necessary: kinetic relations for non-conservative systems. *Proc. R. Soc. Edinburgh Sect. A* **142**, 1–37 (2012)
- [136] C. Berthon, Y. Coudière, V. Desveaux, Development of DDFV methods for the Euler equations, in *Finite Volumes for Complex Applications. VI. Problems & Perspectives*. Springer Proc. Math., vols. 1, 2, (Springer, Heidelberg, 2011), pp. 117–124
- [137] C. Berthon, Y. Coudière, V. Desveaux, Second-order MUSCL schemes based on dual mesh gradient reconstruction (DMGR). *ESAIM Math. Model. Numer. Anal.* **48**, 583–602 (2014)
- [138] C. Berthon, V. Desveaux, An entropy preserving MOOD scheme for the Euler equations. *Int. J. Finite* **11**, 39 (2014)
- [139] C. Berthon, F. Foucher, Efficient well-balanced hydrostatic upwind schemes for shallow-water equations. *J. Comput. Phys.* **231**, 4993–5015 (2012)
- [140] C. Berthon, P.G. LeFloch, R. Turpault, Late-time/stiff-relaxation asymptotic-preserving approximations of hyperbolic equations. *Math. Comput.* **82**, 831–860 (2013)
- [141] C. Berthon, F. Marche, A positive preserving high order VFRoe scheme for shallow water equations: a class of relaxation schemes. *SIAM J. Sci. Comput.* **30**, 2587–2612 (2008)

- [142] C. Berthon, F. Marche, R. Turpault, An efficient scheme on wet/dry transitions for shallow water equations with friction. *Comput. Fluids* **48**, 192–201 (2011)
- [143] C. Berthon, C. Sarazin, R. Turpault, Space-time generalized Riemann problem solvers of order  $k$  for linear advection with unrestricted time step. *J. Sci. Comput.* **55**, 268–308 (2013)
- [144] C. Berthon, R. Turpault, Asymptotic preserving HLL schemes. *Numer. Methods Partial Differ. Equ.* **27**, 1396–1422 (2011)
- [145] F. Beux, S. Lantéri, A. Dervieux, B. Larrouy-Millet, *Upwind stabilization of Navier-Stokes solvers*, INRIA Research Report, INRIA Rocquencourt, 78153 Le Chesnay, France (1993)
- [146] F. Bezard, B. Després, An entropic solver for ideal Lagrangian magnetohydrodynamics. *J. Comput. Phys.* **154**, 65–89 (1999)
- [147] S. Bianchini, A. Bressan, Vanishing viscosity solutions of nonlinear hyperbolic systems. *Ann. Math.* (2) **161**, 223–342 (2005)
- [148] H. Bijl, D. Lucor, S. Mishra, C. Schwab, eds., *Uncertainty Quantification in Computational Fluid Dynamics*. Lecture Notes in Computational Science and Engineering, vol. 92 (Springer, Heidelberg, 2013)
- [149] G. Billet, Finite-difference schemes with dissipation control joined to a generalization of van Leer flux splitting, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24. Friedr. Vieweg, Braunschweig, 1989, pp. 11–20
- [150] S.J. Billett, E.F. Toro, On WAF-type schemes for multidimensional hyperbolic conservation laws. *J. Comput. Phys.* **130**, 1–24 (1997)
- [151] V. Billey, J. Périoux, P. Perrier, B. Stoufflet, in 2-D and 3-D Euler computations with finite element methods in aerodynamics, in *Nonlinear Hyperbolic Problems* (St. Etienne, 1986). Lecture Notes in Math., vol. 1270 (Springer, Berlin, 1987), pp. 64–81
- [152] M. Blunt, B. Rubin, Implicit flux limiting schemes for petroleum reservoir simulation. *J. Comput. Phys.* **102**, 194–210 (1992)
- [153] G. Boillat, Chocs caractéristiques. *C. R. Acad. Sci. Paris Sér. A-B* **274**, A1018–A1021 (1972)
- [154] H. Böing, K. Werner, H. Jackisch, Construction of the entropy solution of hyperbolic conservation laws by a geometrical interpretation of the conservation principle. *J. Comput. Phys.* **95**, 40–58 (1991)
- [155] J.P. Boris, D.L. Book, Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. **11**(1), 38–69 (1973)], **135**, 170–186 (1997). With an introduction by Steven T. Zalesak, Commemoration of the 30th Anniversary of J. Comput. Phys.
- [156] S. Boscarino, P.G. LeFloch, G. Russo, High-order asymptotic-preserving methods for fully nonlinear relaxation problems. *SIAM J. Sci. Comput.* **36**, A377–A395 (2014)

- [157] S. Boscarino, G. Russo, Flux-explicit IMEX Runge-Kutta schemes for hyperbolic to parabolic relaxation problems. *SIAM J. Numer. Anal.* **51**, 163–190 (2013)
- [158] R. Botchorishvili, B. Perthame, A. Vasseur, Equilibrium schemes for scalar conservation laws with stiff sources. *Math. Comput.* **72**, 131–157 (2003) (electronic)
- [159] D. Bouche, J.-M. Ghidaglia, F. Pascal, Error estimate and the geometric corrector for the upwind finite volume method applied to the linear advection equation. *SIAM J. Numer. Anal.* **43**, 578–603 (2005) (electronic)
- [160] D. Bouche, J.-M. Ghidaglia, F. Pascal, An optimal error estimate for upwind finite volume methods for nonlinear hyperbolic conservation laws. *Appl. Numer. Math.* **61**, 1114–1131 (2011)
- [161] F. Bouchut, On zero pressure gas dynamics, in *Advances in Kinetic Theory and Computing*. Ser. Adv. Math. Appl. Sci., vol. 22 (World Sci. Publ., River Edge, 1994), pp. 171–190.
- [162] F. Bouchut, Construction of BGK models with a family of kinetic entropies for a given system of conservation laws. *J. Statist. Phys.* **95**, 113–170 (1999)
- [163] F. Bouchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources*. Frontiers in Mathematics (Birkhäuser, Basel, 2004)
- [164] F. Bouchut, A reduced stability condition for nonlinear relaxation to conservation laws. *J. Hyperbolic Differ. Equ.* **1**, 149–170 (2004)
- [165] F. Bouchut, Stability of relaxation models for conservation laws, in *European Congress of Mathematics* (Eur. Math. Soc., Zürich, 2005), pp. 95–101
- [166] F. Bouchut, C. Bourdarias, B. Perthame, A MUSCL method satisfying all the numerical entropy inequalities. *Math. Comput.* **65**, 1439–1461 (1996)
- [167] F. Bouchut, Y. Brenier, J. Cortes, J.-F. Ripoll, A hierarchy of models for two-phase flows. *J. Nonlinear Sci.* **10**, 639–660 (2000)
- [168] F. Bouchut, E.D. Fernández-Nieto, A. Mangeney, G. Narbona-Reina, A two-phase shallow debris flow model with energy balance. *ESAIM Math. Model. Numer. Anal.* **49**, 101–140 (2015)
- [169] F. Bouchut, F. Golse, M. Pulvirenti, *Kinetic Equations and Asymptotic Theory*. Series in Applied Mathematics (Paris), vol. 4. Gauthier-Villars, Éditions Scientifiques et Médicales Elsevier, Paris, 2000. Edited and with a foreword by Benoît Perthame and Laurent Desvillettes
- [170] F. Bouchut, F. James, One-dimensional transport equations with discontinuous coefficients. *Nonlinear Anal.* **32**, 891–933 (1998)
- [171] F. Bouchut, F. James, Duality solutions for pressureless gases, monotone scalar conservation laws, and uniqueness. *Commun. Partial Differ. Equ.* **24**, 2173–2189 (1999)

- [172] F. Bouchut, C. Klingenberg, K. Waagan, A multiwave approximate Riemann solver for ideal MHD based on relaxation. I. Theoretical framework. *Numer. Math.* **108**, 7–42 (2007)
- [173] F. Bouchut, C. Klingenberg, K. Waagan, A multiwave approximate Riemann solver for ideal MHD based on relaxation II: numerical implementation with 3 and 5 waves. *Numer. Math.* **115**, 647–679 (2010)
- [174] F. Bouchut, T. Morales de Luna, An entropy satisfying scheme for two-layer shallow water equations with uncoupled treatment. *M2AN Math. Model. Numer. Anal.* **42**, 683–698 (2008)
- [175] F. Bouchut, T. Morales de Luna, Semi-discrete entropy satisfying approximate Riemann solvers. The case of the Suliciu relaxation approximation. *J. Sci. Comput.* **41**, 483–509 (2009)
- [176] F. Bouchut, H. Ounaissa, B. Perthame, Upwinding of the source term at interfaces for Euler equations with high friction. *Comput. Math. Appl.* **53**, 361–375 (2007)
- [177] F. Bouchut, B. Perthame, Kružkov’s estimates for scalar conservation laws revisited. *Trans. Am. Math. Soc.* **350**, 2847–2870 (1998)
- [178] F. Bouchut, M. Westdickenberg, Gravity driven shallow water models for arbitrary topography. *Commun. Math. Sci.* **2**, 359–389 (2004)
- [179] T. Boukadia, A.Y. LeRoux, A new version of the two-dimensional Lax-Friedrichs scheme. *Math. Comput.* **63**, 541–553 (1994)
- [180] A.-C. Boulanger, C. Cancès, H. Mathis, K. Saleh, and N. Seguin, OS-AMOAL: Optimized Simulations by Adapted MOdels using Asymptotic Limits, in *CEMRACS’11: Multiscale Coupling of Complex Models in Scientific Computing*, vol. 38 of *ESAIM Proc.*, EDP Sci., Les Ulis, 2012, pp. 183–201
- [181] C. Bourdarias, S. Gerbi, A finite volume scheme for a model coupling free surface and pressurised flows in pipes. *J. Comput. Appl. Math.* **209**, 109–131 (2007)
- [182] C. Bourdarias, S. Gerbi, M. Gisclon, A kinetic formulation for a model coupling free surface and pressurised flows in closed pipes. *J. Comput. Appl. Math.* **218**, 522–531 (2008)
- [183] F. Bourdel, P. Delorme, P.-A. Mazet, Convexity in hyperbolic problems. Application to a discontinuous Galerkin method for the resolution of the polydimensional Euler equations, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24. Friedr. Vieweg, Braunschweig, 1989, pp. 31–42
- [184] A. Bourgeade, P. LeFloch, P.-A. Raviart, An asymptotic expansion for the solution of the generalized Riemann problem. II. Application to the equations of gas dynamics. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **6**, 437–480 (1989)
- [185] A. Bourgeat, J. Koebbe, Minimization of grid orientation effects in simulation of oil recovery processes through use of an unsplit higher order scheme. *Numer. Methods Partial Differ. Equ.* **12**, 161–189 (1996)

- [186] B. Boutin, C. Chalons, P.-A. Raviart, Existence result for the coupling problem of two scalar conservation laws with Riemann initial data. *Math. Models Methods Appl. Sci.* **20**, 1859–1898 (2010)
- [187] B. Boutin, F. Coquel, E. Godlewski, Dafermos regularization for interface coupling of conservation laws, in *Hyperbolic Problems: Theory, Numerics, Applications* (Springer, Berlin, 2008), pp. 567–575
- [188] B. Boutin, F. Coquel, P.G. LeFloch, Coupling techniques for nonlinear hyperbolic equations. III. The well-balanced approximation of thick interfaces. *SIAM J. Numer. Anal.* **51**, 1108–1133 (2013)
- [189] M. Boutoumet, L. Chupin, P. Noble, J.P. Vila, Shallow water viscous flows for arbitrary topography. *Commun. Math. Sci.* **6**, 29–55 (2008)
- [190] Y. Brenier, Résolution d'équations d'évolution quasilinéaires en dimension  $N$  d'espace à l'aide d'équations linéaires en dimension  $N + 1$ . *J. Differ. Equ.* **50**, 375–390 (1983)
- [191] Y. Brenier, Averaged multivalued solutions for scalar conservation laws. *SIAM J. Numer. Anal.* **21**, 1013–1037 (1984)
- [192] Y. Brenier, *Systèmes hyperboliques de lois de conservation*, cours de DEA d'Analyse Numérique (1992–93), Université Pierre et Marie Curie, Paris (France) (1992)
- [193] Y. Brenier, S. Osher, Approximate Riemann solvers and numerical flux functions. *SIAM J. Numer. Anal.* **23**, 259–273 (1986)
- [194] Y. Brenier, S. Osher, The discrete one-sided Lipschitz condition for convex scalar conservation laws. *SIAM J. Numer. Anal.* **25**, 8–23 (1988)
- [195] D. Bresch, B. Desjardins, J.-M. Ghidaglia, E. Grenier, M. Hillairet, Multi-fluid models including compressible fluids, in *Handbook of Mathematical Analysis in Mechanics of Viscous Fluids* (Springer, Cham, 2018), pp. 2927–2978
- [196] D. Bresch, R. Klein, C. Lucas, Multiscale analyses for the shallow water equations, in *Computational Science and High Performance Computing IV*. Notes Numer. Fluid Mech. Multidiscip. Des., vol. 115 (Springer, Berlin, 2011), pp. 149–164
- [197] A. Bressan, *Hyperbolic Systems of Conservation Laws*. Oxford Lecture Series in Mathematics and Its Applications, vol. 20 (Oxford University Press, Oxford, 2000). The one-dimensional Cauchy problem
- [198] A. Bressan, Front tracking method for systems of conservation laws, in *Evolutionary Equations*. Handb. Differ. Equ., Vol. I (North-Holland, Amsterdam, 2004), pp. 87–168
- [199] A. Bressan, S. Čanić, M. Garavello, M. Herty, B. Piccoli, Flows on networks: recent results and perspectives. *EMS Surv. Math. Sci.* **1**, 47–111 (2014)
- [200] A. Bressan, R. M. Colombo, The semigroup generated by  $2 \times 2$  conservation laws. *Arch. Rational Mech. Anal.* **133**, 1–75 (1995)
- [201] A. Bressan, H.K. Jenssen, P. Baiti, An instability of the Godunov scheme. *Commun. Pure Appl. Math.* **59**, 1604–1638 (2006)

- [202] A. Bressan, D. Serre, M. Williams, K. Zumbrun, *Hyperbolic Systems of Balance Laws*. Lecture Notes in Mathematics, vol. 1911 (Springer, Berlin, 2007). Fondazione C.I.M.E., Florence, 2007. Lectures given at the C.I.M.E. Summer School held in Cetraro, July 14–21, 2003, Edited and with a preface by Pierangelo Marcati
- [203] M. Breuss, An analysis of the influence of data extrema on some first and second order central approximations of hyperbolic conservation laws. *M2AN Math. Model. Numer. Anal.* **39**, 965–993 (2005)
- [204] M. Brio, Admissibility conditions for weak solutions of nonstrictly hyperbolic systems, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24. Friedr. Vieweg, Braunschweig, 1989, pp. 43–50
- [205] M. Brio, C.C. Wu, An upwind differencing scheme for the equations of ideal magnetohydrodynamics. *J. Comput. Phys.* **75**, 400–422 (1988)
- [206] M.-O. Bristeau, N. Goutal, J. Sainte-Marie, Numerical simulations of a non-hydrostatic shallow water model. *Comput. Fluids* **47**, 51–64 (2011)
- [207] M.-O. Bristeau, A. Mangeney, J. Sainte-Marie, N. Seguin, An energy-consistent depth-averaged Euler system: derivation and properties. *Discrete Contin. Dyn. Syst. Ser. B* **20**, 961–988 (2015)
- [208] M.-O. Bristeau, J. Sainte-Marie, Derivation of a non-hydrostatic shallow water model; comparison with Saint-Venant and Boussinesq systems. *Discrete Contin. Dyn. Syst. Ser. B* **10**, 733–759 (2008)
- [209] S. Bryson, Y. Epshteyn, A. Kurganov, G. Petrova, Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system. *ESAIM Math. Model. Numer. Anal.* **45**, 423–446 (2011)
- [210] C. Buet, S. Cordier, An asymptotic preserving scheme for hydrodynamics radiative transfer models: numerics for radiative transfer. *Numer. Math.* **108**, 199–221 (2007)
- [211] C. Buet, B. Després, Asymptotic preserving and positive schemes for radiation hydrodynamics. *J. Comput. Phys.* **215**, 717–740 (2006)
- [212] T. Buffard, Analyse de quelques méthodes de volumes finis non structurées pour la résolution des équations d'Euler, Ph.D. thesis, UPMC-Paris06, France (1993)  
Analyse de quelques méthodes de volumes finis non structures pour la résolution des équations d'Euler,
- [213] T. Buffard, S. Clain, Monoslope and multislope MUSCL methods for unstructured meshes. *J. Comput. Phys.* **229**, 3745–3776 (2010)
- [214] T. Buffard, T. Gallouët, J.-M. Hérard, Un schéma simple pour les équations de Saint-Venant, *C. R. Acad. Sci. Paris Sér. I Math.* **326**, 385–390 (1998)
- [215] T. Buffard, T. Gallouët, J.-M. Hérard, A sequel to a rough Godunov scheme: application to real gases. *Comput. Fluids* **29**, 813–847 (2000)

- [216] T. Buffard, J. M. Hérard, A conservative fractional step method to solve non-isentropic Euler equations, *Comput. Methods Appl. Mech. Eng.* **144**, 199–225 (1997)
- [217] B. Bokiet, Application of front tracking to two-dimensional curved detonation fronts. *SIAM J. Sci. Statist. Comput.* **9**, 80–99 (1988)
- [218] R. Bürger, H. Frid, K.H. Karlsen, On the well-posedness of entropy solutions to conservation laws with a zero-flux boundary condition. *J. Math. Anal. Appl.* **326**, 108–120 (2007)
- [219] R. Bürger, K.H. Karlsen, N.H. Risebro, J.D. Towers, Well-posedness in  $BV_t$  and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units. *Numer. Math.* **97**, 25–65 (2004)
- [220] R. Bürger, K.H. Karlsen, J.D. Towers, A conservation law with discontinuous flux modelling traffic flow with abruptly changing road surface conditions, in *Hyperbolic Problems: Theory, Numerics and Applications. Proceedings of Symposia in Applied Mathematics*. vol. 67 (Amer. Math. Soc., Providence, 2009), pp. 455–464
- [221] E. Burman, L. Sainsaulieu, Numerical analysis of two operator splitting methods for an [a] hyperbolic system of conservation laws with stiff relaxation terms. *Comput. Methods Appl. Mech. Eng.* **128**, 291–314 (1995)
- [222] F. Caetano, *Sur certains problèmes de linéarisation et de couplage pour les systèmes hyperboliques non linéaires (French)*, Ph.D. thesis, UPMC-Paris06 (France), 2006
- [223] F. Caetano, The linearization of a boundary value problem for a scalar conservation law. *Commun. Math. Sci.* **6**, 651–667 (2008)
- [224] R.E. Caflisch, S. Jin, G. Russo, Uniformly accurate schemes for hyperbolic systems with relaxation. *SIAM J. Numer. Anal.* **34**, 246–281 (1997)
- [225] D. Calhoun, R.J. Le Veque, An accuracy study of mesh refinement on mapped grids, in *Adaptive Mesh Refinement—Theory and Applications*. Lect. Notes Comput. Sci. Eng., vol. 41 (Springer, Berlin, 2005), pp. 91–101
- [226] L. Cambier, B. Escande, J.-P. Veuillot, *Calculs d'écoulements internes à grand nombre de Reynolds par résolution numérique des équations de Navier–Stokes*, Rech. Aérospat. (1986), pp. 415–432
- [227] L. Cambier, W. Ghazzi, J.-P. Veuillot, H. Viviand, A multidomain approach for the computation of viscous transonic flows by unsteady type methods, in *Computational Methods in Viscous Flows*. Recent Adv. Numer. Methods Fluids, vol. 3 (Pineridge, Swansea, 1984), pp. 513–539
- [228] C. Cancès, F. Coquel, E. Godlewski, H. Mathis, N. Seguin, Error analysis of a dynamic model adaptation procedure for nonlinear hyperbolic equations. *Commun. Math. Sci.* (2015). [10.4310/CMS.2016.v14.n1.a1](https://doi.org/10.4310/CMS.2016.v14.n1.a1)

- [229] C. Cancès, H. Mathis, N. Seguin, Error estimate for time-explicit finite volume approximation of strong solutions to systems of conservation laws. SIAM J. Numer. Anal. **54**, 1263–1287 (2016)
- [230] S. Čanić, B.L. Keyfitz, Riemann problems for the two-dimensional unsteady transonic small disturbance equation. SIAM J. Appl. Math. **58**, 636–665 (1998)
- [231] S. Čanić, B.J. Plohr, Shock wave admissibility for quadratic conservation laws. J. Differ. Equ. **118**, 293–335 (1995)
- [232] C. Carasso, P.-A. Raviart, D. Serre, eds., *Nonlinear Hyperbolic Problems, Proceedings St. Etienne 1986*. Lecture Notes in Mathematics, vol. 1270 (Springer, Berlin, 1987)
- [233] G. Carbou, B. Hanouzet, Relaxation approximation of some initial-boundary value problem for  $p$ -systems. Commun. Math. Sci. **5**, 187–203 (2007)
- [234] G. Carbou, B. Hanouzet, R. Natalini, Semilinear behavior for totally linearly degenerate hyperbolic systems with relaxation. J. Differ. Equ. **246**, 291–319 (2009)
- [235] P. Cargo, G. Gallice, Un solveur de Roe pour les équations de la magnétohydrodynamique. C. R. Acad. Sci. Paris Sér. I Math. **320**, 1269–1272 (1995)
- [236] P. Cargo, G. Gallice, Roe matrices for ideal MHD and systematic construction of Roe matrices for systems of conservation laws. J. Comput. Phys. **136**, 446–466 (1997)
- [237] P. Cargo, A.Y. LeRoux, Un schéma équilibre adapté au modèle d’atmosphère avec termes de gravité. C. R. Acad. Sci. Paris Sér. I Math. **318**, 73–76 (1994)
- [238] F. Caro, F. Coquel, D. Jamet, S. Kokh, A simple finite-volume method for compressible isothermal two-phase flows simulation. Int. J. Finite **3**, 37 (2006)
- [239] J. Casper, H.L. Atkins, A finite-volume high-order ENO scheme for two-dimensional hyperbolic systems. J. Comput. Phys. **106**, 62–76 (1993)
- [240] C.E. Castro, E.F. Toro, A Riemann solver and upwind methods for a two-phase flow model in non-conservative form. Int. J. Numer. Methods Fluids **50**, 275–307 (2006)
- [241] M.J. Castro Díaz, U.S. Fjordholm, S. Mishra, C. Parés, Entropy conservative and entropy stable schemes for nonconservative hyperbolic systems. SIAM J. Numer. Anal. **51**, 1371–1391 (2013)
- [242] M.J. Castro Díaz, J.M. Gallardo, A. Marquina, Approximate Osher-Solomon schemes for hyperbolic systems. Appl. Math. Comput. **272**, 347–368 (2016)
- [243] M.J. Castro Díaz, J.M. González-Vida, C. Parés, Numerical treatment of wet/dry fronts in shallow flows with a modified Roe scheme. Math. Models Methods Appl. Sci. **16**, 897–931 (2006)

- [244] M.J. Castro Díaz, P.G. LeFloch, M.L. Muñoz-Ruiz, C. Parés, Why many theories of shock waves are necessary: convergence error in formally path-consistent schemes. *J. Comput. Phys.* **227**, 8107–8129 (2008)
- [245] M.J. Castro Díaz, P.G. LeFloch, M.L. Muñoz-Ruiz, C. Parés, Numerical investigation of finite difference schemes for nonconservative hyperbolic systems, in *Hyperbolic Problems: Theory, Numerics and Applications. Proceedings of Symposia in Applied Mathematics*, vol. 67 (Amer. Math. Soc., Providence, 2009), pp. 465–475
- [246] M.J. Castro Díaz, C. Parés, G. Puppo, G. Russo, Central schemes for nonconservative hyperbolic systems. *SIAM J. Sci. Comput.* **34**, B523–B558 (2012)
- [247] M.J. Castro Díaz, T. Chacón Rebollo, E.D. Fernández-Nieto, J.M. González Vida, C. Parés, Well-balanced finite volume schemes for 2D non-homogeneous hyperbolic systems. Application to the dam break of Aznalcóllar. *Comput. Methods Appl. Mech. Eng.* **197**, 3932–3950 (2008)
- [248] M.J. Castro Díaz, E.D. Fernández-Nieto, J.M. González-Vida, C. Parés-Madroñal, Numerical treatment of the loss of hyperbolicity of the two-layer shallow-water system. *J. Sci. Comput.* **48**, 16–40 (2011)
- [249] M.J. Castro Díaz, E.D. Fernández-Nieto, T. Morales de Luna, G. Narbona-Reina, C. Parés, A HLLC scheme for nonconservative hyperbolic problems. Application to turbidity currents with sediment transport. *ESAIM Math. Model. Numer. Anal.* **47**, 1–32 (2013)
- [250] M.J. Castro Díaz, E.D. Fernández-Nieto, G. Narbona-Reina, M. de la Asunción, A second order PVM flux limiter method. Application to magnetohydrodynamics and shallow stratified flows. *J. Comput. Phys.* **262**, 172–193 (2014)
- [251] M.J. Castro Díaz, J.A. López-García, C. Parés, High order exactly well-balanced numerical methods for shallow water systems. *J. Comput. Phys.* **246**, 242–264 (2013)
- [252] J.-J. Cauret, J.-F. Colombeau, A.Y. LeRoux, Discontinuous generalized solutions of nonlinear nonconservative hyperbolic equations. *J. Math. Anal. Appl.* **139**, 552–573 (1989)
- [253] C. Cercignani, *The Boltzmann Equation and Its Applications*. Applied Mathematical Sciences, vol. 67 (Springer, New York, 1988)
- [254] C. Cercignani, R. Illner, M. Pulvirenti, *The Mathematical Theory of Dilute Gases*. Applied Mathematical Sciences, vol. 106 (Springer, New York, 1994)
- [255] C. Chainais-Hillairet, Finite volume schemes for a nonlinear hyperbolic equation. Convergence towards the entropy solution and error estimate. *M2AN Math. Model. Numer. Anal.* **33**, 129–156 (1999)

- [256] C. Chainais-Hillairet, Second-order finite-volume schemes for a non-linear hyperbolic equation: error estimate. *Math. Methods Appl. Sci.* **23**, 467–490 (2000)
- [257] C. Chainais-Hillairet, S. Champier, Finite volume schemes for non-homogeneous scalar conservation laws: error estimate. *Numer. Math.* **88**, 607–639 (2001)
- [258] S.R. Chakravarthy, S. Osher, Computing with high-resolution upwind schemes for hyperbolic equations, in *Large-Scale Computations in Fluid Mechanics, Part 1* (La Jolla, Calif., 1983). Lectures in Appl. Math., vol. 22 (Amer. Math. Soc., Providence, 1985), pp. 57–86
- [259] A. Chalabi, Stable upwind schemes for hyperbolic conservation laws with source terms. *IMA J. Numer. Anal.* **12**, 217–241 (1992)
- [260] A. Chalabi, Convergence of relaxation schemes for hyperbolic conservation laws with stiff source terms. *Math. Comput.* **68**, 955–970 (1999)
- [261] A. Chalabi, D. Seghir, Convergence of relaxation schemes for initial boundary value problems for conservation laws. *Comput. Math. Appl.* **43**(8-9), 1079–1093 (2002)
- [262] A. Chalabi, J.-P. Vila, Operator splitting, fractional steps method and entropy condition, in *Third International Conference on Hyperbolic Problems*, vols. I, II (Uppsala, 1990), Studentlitteratur, Lund, 1991, pp. 226–240
- [263] N. Chalmers, E. Lorin, Approximation of nonconservative hyperbolic systems based on different shock curve definitions. *Can. Appl. Math. Q.* **17**, 447–485 (2009)
- [264] N. Chalmers, E. Lorin, On the numerical approximation of one-dimensional nonconservative hyperbolic systems. *J. Comput. Sci.* **4**, 111–124 (2013)
- [265] C. Chalons, F. Coquel, Modified Suliciu relaxation system and exact resolution of isolated shock waves. *Math. Models Methods Appl. Sci.* **24**, 937–971 (2014)
- [266] C. Chalons, F. Coquel, P. Engel, C. Rohde, Fast relaxation solvers for hyperbolic-elliptic phase transition problems. *SIAM J. Sci. Comput.* **34**, A1753–A1776 (2012)
- [267] C. Chalons, F. Coquel, E. Godlewski, P.-A. Raviart, N. Seguin, Godunov-type schemes for hyperbolic systems with parameter-dependent source. The case of Euler system with friction. *Math. Models Methods Appl. Sci.* **20**, 2109–2166 (2010)
- [268] C. Chalons, J.-F. Coulombel, Relaxation approximation of the Euler equations. *J. Math. Anal. Appl.* **348**, 872–893 (2008)
- [269] C. Chalons, M. Girardin, S. Kokh, Large time step and asymptotic preserving numerical schemes for the gas dynamics equations with source terms. *SIAM J. Sci. Comput.* **35**, A2874–A2902 (2013)

- [270] C. Chalons, M. Girardin, S. Kokh, Operator-splitting based asymptotic preserving scheme for the gas dynamics equations with stiff source terms, in *Proceedings of the 2012 International Conference on Hyperbolic Problems: Theory, Numerics, Applications, no. 8 in AIMS on Applied Mathematics* (American Institute of Mathematical Sciences, 2014), pp. 607–614
- [271] C. Chalons, M. Girardin, S. Kokh, An all-regime Lagrange-projection like scheme for the gas dynamics equations on unstructured meshes. *Commun. Comput. Phys.* **20**, 188–233 (2016)
- [272] C. Chalons, M. Girardin, S. Kokh, An all-regime Lagrange-projection like scheme for 2D homogeneous models for two-phase flows on unstructured meshes. *J. Comput. Phys.* **335**, 885–904 (2017)
- [273] C. Chalons, P. Goatin, Transport-equilibrium schemes for computing contact discontinuities in traffic flow modeling. *Commun. Math. Sci.* **5**, 533–551 (2007)
- [274] C. Chalons, P.-A. Raviart, N. Seguin, The interface coupling of the gas dynamics equations. *Q. Appl. Math.* **66**, 659–705 (2008)
- [275] S. Champier, T. Gallouët, Convergence d'un schéma décentré amont sur un maillage triangulaire pour un problème hyperbolique linéaire. *RAIRO Modél. Math. Anal. Numér.* **26**, 835–853 (1992)
- [276] S. Champier, T. Gallouët, R. Herbin, Convergence of an upstream finite volume scheme for a nonlinear hyperbolic equation on a triangular mesh. *Numer. Math.* **66**, 139–157 (1993)
- [277] C.-L. Chang, C.L. Merkle, The relation between flux vector splitting and parabolized schemes. *J. Comput. Phys.* **80**, 344–361 (1989)
- [278] T. Chang and L. Hsiao, *The Riemann Problem and Interaction of Waves in Gas Dynamics*. Pitman Monographs and Surveys in Pure and Applied Mathematics, vol. 41 Longman Scientific & Technical, Harlow; copublished in the United States with John Wiley & Sons, New York, 1989
- [279] G. Chanteperdrix, P. Villedieu, J.-P. Vila, A compressible model for separated two-phase flows computations, in *ASME 2002 Joint US-European Fluids Engineering Division Conference* (American Society of Mechanical Engineers, 2002), pp. 809–816
- [280] D. Chargy, R. Abgrall, L. Fézoui, B. Larrouy, *Conservative Numerical Schemes for Multicomponent Inviscid Flows* (Rech. Aérospat., 1992), pp. 61–80
- [281] P. Charrier, B. Dubroca, L. Flandrin, Un solveur de Riemann approché pour l'étude d'écoulements hypersoniques bidimensionnels. *C. R. Acad. Sci. Paris Sér. I Math.* **317**, 1083–1086 (1993)
- [282] P. Charrier, B. Tessieras, On front-tracking methods applied to hyperbolic systems of nonlinear conservation laws. *SIAM J. Numer. Anal.* **23**, 461–472 (1986)
- [283] G.Q. Chen, The method of quasidecoupling for discontinuous solutions to conservation laws. *Arch. Ration. Mech. Anal.* **121**, 131–185 (1992)

- [284] G.-Q. Chen, H. Frid, Vanishing viscosity limit for initial-boundary value problems for conservation laws, in *Nonlinear Partial Differential Equations* (Evanston, IL, 1998). Contemp. Math., vol. 238 (Amer. Math. Soc., Providence, 1999), pp. 35–51
- [285] G.-Q. Chen, H. Frid, Uniqueness and asymptotic stability of Riemann solutions for the compressible Euler equations. Trans. Am. Math. Soc. **353**, 1103–1117 (2001)
- [286] G.-Q. Chen, J. Glimm, Shock capturing and global solutions to the compressible Euler equations with geometrical structure, in *Hyperbolic problems: theory, numerics, applications* (Stony Brook, NY, 1994) (World Sci. Publ., River Edge, 1996), pp. 101–109
- [287] G.-Q. Chen, D. Hoff, K. Trivisa, Global solutions to a model for exothermically reacting, compressible flows with large discontinuous initial data. Arch. Ration. Mech. Anal. **166**, 321–358 (2003)
- [288] G.-Q. Chen, P.T. Kan, Hyperbolic conservation laws with umbilic degeneracy. I. Arch. Ration. Mech. Anal. **130**, 231–276 (1995)
- [289] G.-Q. Chen, P.T. Kan, Global solutions to hyperbolic conservation laws with umbilic degeneracy, in *Hyperbolic Problems: Theory, Numerics, Applications* (Stony Brook, NY, 1994) (World Sci. Publ., River Edge, 1996), pp. 368–374
- [290] G.Q. Chen, P.G. LeFloch, Entropy flux-splittings for hyperbolic conservation laws. I. General framework. Commun. Pure Appl. Math. **48**, 691–729 (1995)
- [291] G.Q. Chen, C.D. Levermore, T.-P. Liu, Hyperbolic conservation laws with stiff relaxation terms and entropy. Commun. Pure Appl. Math. **47**, 787–830 (1994)
- [292] G.Q. Chen, J.-G. Liu, Convergence of second-order schemes for isentropic gas dynamics. Math. Comput. **61**, 607–627 (1993)
- [293] G.Q. Chen, T.-P. Liu, Zero relaxation and dissipation limits for hyperbolic conservation laws. Commun. Pure Appl. Math. **46**, 755–781 (1993)
- [294] G.Q. Chen, D.H. Wagner, Large time, weak solutions to reacting Euler equations, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects* (Taormina, 1992). Notes Numer. Fluid Mech., vol. 43. Friedr. Vieweg, Braunschweig, 1993, pp. 144–149
- [295] G.Q. Chen, D.H. Wagner, Global entropy solutions to exothermically reacting, compressible Euler equations. J. Differ. Equ. **191**, 277–322 (2003)
- [296] Y. Chen, S. Jiang, A non-oscillatory kinetic scheme for multi-component flows with the equation of state for a stiffened gas. J. Comput. Math. **29**, 661–683 (2011)
- [297] R. Chéret, *La Détonation des Explosifs Condensés*, Masson, Paris, 1988. Tome 1. Série scientifique, Collection CEA.
- [298] I.-L. Chern, J. Glimm, O. McBryan, B. Plohr, S. Yaniv, Front tracking for gas dynamics. J. Comput. Phys. **62**, 83–110 (1986)

- [299] A. Chertock, S. Cui, A. Kurganov, C.N. Özcan, E. Tadmor, Well-balanced schemes for the Euler equations with gravitation: conservative formulation using global fluxes. *J. Comput. Phys.* **358**, 36–52 (2018)
- [300] P. Chévrier, H. Galley, A van Leer finite volume scheme for the Euler equations on unstructured meshes. *RAIRO Modél. Math. Anal. Numér.* **27**, 183–201 (1993)
- [301] A. Chinnayya, A.-Y. LeRoux, N. Seguin, A well-balanced numerical scheme for the approximation of the shallow-water equations with topography: the resonance phenomenon. *Int. J. Finite* **1**, 33 (2004)
- [302] E. Chiodaroli, C. De Lellis, O.R. Kreml, Global ill-posedness of the isentropic system of gas dynamics. *Commun. Pure Appl. Math.* **68**, 1157–1190 (2015)
- [303] A.J. Chorin, Random choice solution of hyperbolic systems. *J. Computational Phys.* **22**, 517–533 (1976)
- [304] A.J. Chorin, Random choice methods with application to reacting gas flow. *J. Comput. Phys.* **25**, 253–272 (1977)
- [305] A.J. Chorin, J.E. Marsden, *A Mathematical Introduction to Fluid Mechanics* (Springer, New York, 1979)
- [306] A.J. Chorin, M.F. McCracken, T.J.R. Hughes, J.E. Marsden, Product formulas and numerical algorithms. *Commun. Pure Appl. Math.* **31**, 205–256 (1978)
- [307] K.N. Chueh, C.C. Conley, J.A. Smoller, Positively invariant regions for systems of nonlinear diffusion equations. *Indiana Univ. Math. J.* **26**, 373–392 (1977)
- [308] M.C. Ciccoli, L. Fézou, J.-A. Désidéri, *Efficient Methods for Inviscid Nonequilibrium Hypersonic Flow Fields* (Rech. Aérospat., 1992), pp. 37–52
- [309] P. Cinnella, B. Grossman, Flux-split algorithms for hypersonic flows, in *Computational methods in hypersonic aerodynamics*. *Fluid Mech. Appl.*, vol. 9 (Kluwer Acad. Publ., Dordrecht, 1991), pp. 153–202
- [310] J.-P. Cioni, L. Fézou, H. Steve, A parallel time-domain Maxwell solver using upwind schemes and triangular meshes. *Impact Comput. Sci. Eng.* **5**, 215–247 (1993)
- [311] S. Clain, S. Diot, R. Loubère, A high-order finite volume method for systems of conservation laws—Multi-dimensional Optimal Order Detection (MOOD). *J. Comput. Phys.* **230**, 4028–4050 (2011)
- [312] S. Clain, D. Rochette, First- and second-order finite volume methods for the one-dimensional nonconservative Euler system. *J. Comput. Phys.* **228**, 8214–8248 (2009)
- [313] J.F. Clarke, S. Karni, J. Quirk, P. Roe, L. Simmonds, E. Toro, Numerical computation for two-dimensional unsteady detonation waves in high energy solids. *J. Comput. Phys.* **106**, 215–233 (1993)

- [314] J.-P. Cocchi, R. Saurel, A Riemann problem based method for the resolution of compressible multimaterial flows. *J. Comput. Phys.* **137**, 265–298 (1997)
- [315] B. Cockburn, Quasimonotone schemes for scalar conservation laws. I. *SIAM J. Numer. Anal.* **26**, 1325–1341 (1989)
- [316] B. Cockburn, F. Coquel, P.G. LeFloch, An error estimate for finite volume methods for multidimensional conservation laws. *Math. Comput.* **63**, 77–103 (1994)
- [317] B. Cockburn, F. Coquel, P.G. LeFloch, Convergence of the finite volume method for multidimensional conservation laws. *SIAM J. Numer. Anal.* **32**, 687–705 (1995)
- [318] B. Cockburn, P.-A. Gremaud, A priori error estimates for numerical methods for scalar conservation laws. I. The general approach. *Math. Comput.* **65**, 533–573 (1996)
- [319] B. Cockburn, S. Hou, C.-W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case. *Math. Comput.* **54**, 545–581 (1990)
- [320] B. Cockburn, C. Johnson, C.-W. Shu, E. Tadmor, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*. Lecture Notes in Mathematics, vol. 1697 (Springer, Berlin, 1998); Centro Internazionale Matematico Estivo (C.I.M.E.), Florence, 1998. Papers from the C.I.M.E. Summer School held in Cetraro, June 23–28, 1997, Edited by Alfio Quarteroni, Fondazione CIME/CIME Foundation Subseries.
- [321] B. Cockburn, G.E. Karniadakis, C.-W. Shu, The development of discontinuous Galerkin methods, in *Discontinuous Galerkin Methods* (Newport, RI, 1999). Lect. Notes Comput. Sci. Eng., vol. 11 (Springer, Berlin, 2000), pp. 3–50
- [322] B. Cockburn, C.-W. Shu, Nonlinearly stable compact schemes for shock calculations. *SIAM J. Numer. Anal.* **31**, 607–627 (1994)
- [323] A. Cohen, S.M. Kaber, S. Müller, M. Postel, Fully adaptive multiresolution finite volume schemes for conservation laws. *Math. Comput.* **72**, 183–225 (2003)
- [324] W.J. Coirier, K. G. Powell, An accuracy assessment of Cartesian-mesh approaches for the Euler equations. *J. Comput. Phys.* **117**, 121–131 (1995)
- [325] J.D. Cole, On a quasi-linear parabolic equation occurring in aerodynamics. *Q. Appl. Math.* **9**, 225–236 (1951)
- [326] P. Colella, Glimm's method for gas dynamics. *SIAM J. Sci. Statist. Comput.* **3**, 76–110 (1982)
- [327] P. Colella, A direct Eulerian MUSCL scheme for gas dynamics. *SIAM J. Sci. Stat. Comput.* **6**, 104–117 (1985)
- [328] P. Colella, Multidimensional upwind methods for hyperbolic conservation laws. *J. Comput. Phys.* **87**, 171–200 (1990)
- [329] P. Colella, H. M. Glaz, Efficient solution algorithms for the Riemann problem for real gases. *J. Comput. Phys.* **59**, 264–289 (1985)

- [330] P. Colella, A. Majda, V. Roytburd, Fractional step methods for reacting shock waves, in *Reacting Flows: Combustion and Chemical Reactors, Part 2* (Ithaca, N.Y., 1985). Lectures in Appl. Math., vol. 24 (Amer. Math. Soc., Providence, 1986), pp. 459–477
- [331] P. Colella, A. Majda, V. Roytburd, Theoretical and numerical structure for reacting shock waves. SIAM J. Sci. Stat. Comput. **7**, 1059–1080 (1986)
- [332] P. Colella, P. Woodward, The piecewise parabolic method (PPM) for gas dynamics simulations. J. Comput. Phys. **54**, 174–201 (1984)
- [333] J.P. Collins, P. Colella, H.M. Glaz, An implicit-explicit Eulerian Godunov scheme for compressible flow. J. Comput. Phys. **116**, 195–211 (1995)
- [334] J.-F. Colombeau, A.Y. LeRoux, A. Noussaïr, B. Perrot, Microscopic profiles of shock waves and ambiguities in multiplications of distributions. SIAM J. Numer. Anal. **26**, 871–883 (1989)
- [335] R.M. Colombo, Wave front tracking in systems of conservation laws. Appl. Math. **49**, 501–537 (2004)
- [336] R.M. Colombo, A. Corli, Sonic hyperbolic phase transitions and Chapman-Jouguet detonations. J. Differ. Equ. **184**, 321–347 (2002)
- [337] R.M. Colombo, A. Corli, On the operator splitting method: nonlinear balance laws and a generalization of Trotter-Kato formulas, in *Hyperbolic problems and regularity questions, Trends Math.* (Birkhäuser, Basel, 2007), pp. 91–100
- [338] R.M. Colombo, M. Garavello, A well posed Riemann problem for the  $p$ -system at a junction. Netw. Heterog. Media **1**, 495–511 (2006)
- [339] R.M. Colombo, G. Guerra, A coupling between a non-linear 1D compressible-incompressible limit and the 1D  $p$ -system in the non smooth case. Netw. Heterog. Media **11**, 313–330 (2016)
- [340] R.M. Colombo, M. Herty, Nodal conditions for hyperbolic systems of balance laws, in *Hyperbolic Problems: Theory, Numerics, Applications, no. 8 in AIMS on Applied Mathematics* (American Institute of Mathematical Sciences, 2014), pp. 147–161. Proceedings of the fourteenth International Conference on Hyperbolic Problems
- [341] R.M. Colombo, F. Marcellini, Coupling conditions for the  $3 \times 3$  Euler system. Netw. Heterog. Media **5**, 675–690 (2010)
- [342] R.M. Colombo, F. Marcellini, Smooth and discontinuous junctions in the  $p$ -system and in the  $3 \times 3$  Euler system. Riv. Math. Univ. Parma (N.S.) **3**, 55–69 (2012)
- [343] R.M. Colombo, M. Mercier, M.D. Rosini, Stability and total variation estimates on general scalar balance laws. Commun. Math. Sci. **7**, 37–65 (2009)
- [344] R.M. Colombo, M. Mercier, M.D. Rosini, Stability estimates on general scalar balance laws. C. R. Math. Acad. Sci. Paris **347**, 45–48 (2009)

- [345] F. Coquel, K. El Amine, E. Godlewski, B. Perthame, P. Rascle, A numerical method using upwind schemes for the resolution of two-phase flows. *J. Comput. Phys.* **136**, 272–288 (1997)
- [346] F. Coquel, T. Gallouët, P. Helluy, J.-M. Hérard, O. Hurisse, N. Seguin, Modelling compressible multiphase flows, in *Applied mathematics in Savoie—AMIS 2012: Multiphase Flow in Industrial and Environmental Engineering, of ESAIM Proc., EDP Sci.*, Les Ulis, 2013, pp. 34–50
- [347] F. Coquel, E. Godlewski, K. Haddaoui, C. Marmignon, F. Renac, Choice of measure source terms in interface coupling for a model problem in gas dynamics. *Math. Comput.* **85**, 2305–2339 (2016)
- [348] F. Coquel, E. Godlewski, B. Perthame, A. In, P. Rascle, Some new Godunov and relaxation methods for two-phase flow problems, in *Godunov Methods* (Oxford, 1999) (Kluwer/Plenum, New York, 2001), pp. 179–188
- [349] F. Coquel, E. Godlewski, N. Seguin, Relaxation of fluid systems. *Math. Models Methods Appl. Sci.* **22**, 1250014, 52 (2012)
- [350] F. Coquel, J.-M. Hérard, K. Saleh, A splitting method for the isentropic Baer-Nunziato two-phase flow model, in *CEMRACS'11: Multiscale Coupling of Complex Models in Scientific Computing, vol. 38 of ESAIM Proc., EDP Sci.*, Les Ulis, 2012, pp. 241–256
- [351] F. Coquel, J.-M. Hérard, K. Saleh, A positive and entropy-satisfying finite volume scheme for the Baer-Nunziato model. *J. Comput. Phys.* **330**, 401–435 (2017)
- [352] F. Coquel, J.-M. Hérard, K. Saleh, N. Seguin, A robust entropy-satisfying finite volume scheme for the isentropic Baer-Nunziato model. *ESAIM Math. Model. Numer. Anal.* **48**, 165–206 (2014)
- [353] F. Coquel, S. Jin, J.-G. Liu, L. Wang, Well-posedness and singular limit of a semilinear hyperbolic relaxation system with a two-scale discontinuous relaxation rate. *Arch. Ration. Mech. Anal.* **214**, 1051–1084 (2014)
- [354] F. Coquel, S. Jin, J.-G. Liu, L. Wang, Entropic sub-cell shock capturing schemes via Jin-Xin relaxation and Glimm front sampling for scalar conservation laws. *Math. Comput.* **87**, 1083–1126 (2018)
- [355] F. Coquel, P. LeFloch, Convergence of finite difference schemes for conservation laws in several space dimensions: the corrected antidiiffusive flux approach. *Math. Comput.* **57**, 169–210 (1991)
- [356] F. Coquel, P. LeFloch, Convergence of finite difference schemes for conservation laws in several space dimensions: a general theory. *SIAM J. Numer. Anal.* **30**, 675–700 (1993)
- [357] F. Coquel, P. LeFloch, An entropy satisfying MUSCL scheme for systems of conservation laws. *Numer. Math.* **74**, 1–33 (1996)
- [358] F. Coquel, M.-S. Liou, Stable and low diffusive hybrid upwind splitting methods, in *Computational Fluid Dynamics' 92, Proceedings of the First European Computational Fluid Dynamics Conference*, vol. 2, 7–11 September 1992, Brussels, Belgium, ed. by C. Hirsch, J. Périaux, W. Kordulla (Elsevier, Amsterdam, 1992)

- [359] F. Coquel, M.-S. Liou, *Hybrid upwind splitting (HUS) by a field-by-field decomposition*, National Aeronautics and Space Administration, ICOMP, NASA technical memorandum (1995)
- [360] F. Coquel, B. Perthame, Relaxation of energy and approximate Riemann solvers for general pressure laws in fluid dynamics. SIAM J. Numer. Anal. **35**, 2223–2249 (1998) (electronic)
- [361] F. Coquel, K. Saleh, N. Seguin, A robust and entropy-satisfying numerical scheme for fluid flows in discontinuous nozzles. Math. Models Methods Appl. Sci. **24**, 2043–2083 (2014)
- [362] F. Cordier, P. Degond, A. Kumbaro, An asymptotic-preserving all-speed scheme for the Euler and Navier-Stokes equations. J. Comput. Phys. **231**, 5685–5704 (2012)
- [363] F. Coron, B. Perthame, Numerical passage from kinetic to fluid equations. SIAM J. Numer. Anal. **28**, 26–42 (1991)
- [364] J. Cortes, A. Debussche, I. Toumi, A density perturbation method to study the eigenstructure of two-phase flow equation systems. J. Comput. Phys. **147**, 463–484 (1998)
- [365] J.-F. Coulombel, Stability of finite difference schemes for hyperbolic initial boundary value problems. SIAM J. Numer. Anal. **47**, 2844–2871 (2009)
- [366] J.-F. Coulombel, The hyperbolic region for hyperbolic boundary value problems. Osaka J. Math. **48**, 457–469 (2011)
- [367] J.-F. Coulombel, Stability of finite difference schemes for hyperbolic initial boundary value problems II. Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **10**, 37–98 (2011)
- [368] J.-F. Coulombel On the strong stability of finite difference schemes for hyperbolic systems in two space dimensions. Calcolo **51**, 97–108 (2014)
- [369] J.-F. Coulombel, Fully discrete hyperbolic initial boundary value problems with nonzero initial data. Confluentes Math. **7**, 17–47 (2015)
- [370] J.-F. Coulombel, T. Goudon, The strong relaxation limit of the multi-dimensional isothermal Euler equations. Trans. Am. Math. Soc. **359**, 637–648 (2007)
- [371] R. Courant, K.O. Friedrichs, *Supersonic Flow and Shock Waves* (Springer, New York-Heidelberg, 1976). Reprinting of the 1948 original, Applied Mathematical Sciences, vol. 21
- [372] M. Crandall, A. Majda, The method of fractional steps for conservation laws. Numer. Math. **34**, 285–314 (1979)
- [373] M. Crandall, A. Majda, Monotone difference approximations for scalar conservation laws. Math. Comput. **34**, 1–21 (1980)
- [374] J.-P. Croisille, *Contribution à l'étude théorique et à l'approximation par éléments finis du système hyperbolique de la dynamique des gaz multidimensionnelle et multiespèces*, PhD thesis, UPMC-Paris06 (France), 1990

- [375] J.-P. Croisille, P. Delorme, Kinetic symmetrizations and pressure laws for the Euler equations. *Phys. D* **57**, 395–416 (1992)
- [376] J.-P. Croisille, R. Khanfir, G. Chanteur, Numerical simulation of the MHD equations by a kinetic-type methods. *J. Sci. Comput.* **10**, 81–92 (1995)
- [377] J.-P. Croisille, P. Villedieu, Entropies de Lax pour les équations d'Euler en déséquilibre thermochimique. *C. R. Acad. Sci. Paris Sér. I Math.* **318**, 723–727 (1994)
- [378] C.M. Dafermos, Polygonal approximations of solutions of the initial value problem for a conservation law. *J. Math. Anal. Appl.* **38**, 33–41 (1972)
- [379] C.M. Dafermos, The entropy rate admissibility criterion for solutions of hyperbolic conservation laws. *J. Differ. Equ.* **14**, 202–212 (1973)
- [380] C.M. Dafermos, Solution of the Riemann problem for a class of hyperbolic systems of conservation laws by the viscosity method. *Arch. Rational Mech. Anal.* **52**, 1–9 (1973)
- [381] C.M. Dafermos, *Hyperbolic systems of conservation laws*, in Systems of nonlinear partial differential equations (Oxford, 1982), vol. 111 of NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. (Reidel, Dordrecht, 1983), pp. 25–70
- [382] C.M. Dafermos, Generalized characteristics in hyperbolic systems of conservation laws. *Arch. Rational Mech. Anal.* **107**, 127–155 (1989)
- [383] C.M. Dafermos, Equivalence of referential and spatial field equations in continuum physics, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects* (Taormina, 1992). Notes Numer. Fluid Mech., vol. 43, Friedr. Vieweg, Braunschweig, 1993, pp. 179–183
- [384] C.M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 2nd edn., vol. 325 (Springer, Berlin, 2005)
- [385] W. Dahmen, S. Müller, A. Voß, Riemann problem for the Euler equation with non-convex equation of state including phase transitions, in *Analysis and Numerics for Conservation Laws* (Springer, Berlin, 2005), pp. 137–162
- [386] W. Dai, P.R. Woodward, Extension of the piecewise parabolic method to multidimensional ideal magnetohydrodynamics. *J. Comput. Phys.* **115**, 485–514 (1994)
- [387] G. Dal Maso, P.G. Lefloch, F. Murat, Definition and weak stability of nonconservative products. *J. Math. Pures Appl. (9)* **74**, 483–548 (1995)
- [388] V.G. Danilov, V.M. Shelkovich, Delta-shock wave type solution of hyperbolic systems of conservation laws. *Q. Appl. Math.* **63**, 401–427 (2005)

- [389] V. Daru, A. Lerat, Analysis of an implicit Euler solver, in *Numerical Methods for the Euler Equations of Fluid Dynamics* (Rocquencourt, 1983) (SIAM, Philadelphia, 1985), pp. 246–280
- [390] R. Dautray, J.-L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques. Vol. 1*, INSTN: Collection Enseignement. [INSTN: Teaching Collection], Masson, Paris, 1987. Modèles physiques. [Physical models], With the collaboration of Michel Cessenat, André Gervat and Hélène Lanchon, Reprinted from the 1984 edition
- [391] S.F. Davis, A rotationally biased upwind difference scheme for the Euler equations. *J. Comput. Phys.* **56**, 65–92 (1984)
- [392] S.F. Davis, A simplified TVD finite difference scheme via artificial viscosity. *SIAM J. Sci. Stat. Comput.* **8**, 1–18 (1987)
- [393] S.F. Davis, Simplified second-order Godunov-type methods. *SIAM J. Sci. Stat. Comput.* **9**, 445–473 (1988)
- [394] S.F. Davis, An interface tracking method for hyperbolic systems of conservation laws. *Appl. Numer. Math.* **10**, 447–472 (1992)
- [395] C. Dawson, Godunov-mixed methods for advection-diffusion equations in multidimensions. *SIAM J. Numer. Anal.* **30**, 1315–1332 (1993)
- [396] F. De Vuyst, A new implicit second order scheme based on a kinetic interpretation for solving the Euler equations, in *Numerical Methods for Fluid Dynamics*, vol. 4 (Reading, 1992) (Oxford Univ. Press, New York, 1993), pp. 425–433
- [397] F. De Vuyst, Traitement décentré des variables caractéristiques pour le schéma de Roe. Une approche macroscopique ou cinétique selon la nature des champs caractéristiques. *C. R. Acad. Sci. Paris Sér. I Math.* **320**, 743–748 (1995)
- [398] A. Decoene, L. Bonaventura, E. Miglio, F. Saleri, Asymptotic derivation of the section-averaged shallow water equations for natural river hydraulics. *Math. Models Methods Appl. Sci.* **19**, 387–417 (2009)
- [399] H. Deconinck, P.L. Roe, R. Struijs, A multidimensional generalization of Roe's flux difference splitter for the Euler equations. *Comput. Fluids* **22**, 215–222 (1993)
- [400] H. Deconinck, R. Struijs, H. Paillère, et al., Development of cell-vertex multidimensional upwind solvers for the compressible flow equations. *CWI Q.* **6**, 1–28 (1993)
- [401] P. Degond, G. Dimarco, L. Mieussens, A moving interface method for dynamic kinetic-fluid coupling. *J. Comput. Phys.* **227**, 1176–1208 (2007)
- [402] P. Degond, S. Jin, A smooth transition model between kinetic and diffusion equations. *SIAM J. Numer. Anal.* **42**, 2671–2687 (2005) (electronic)
- [403] P. Degond, M. Tang, All speed scheme for the low Mach number limit of the isentropic Euler equations. *Commun. Comput. Phys.* **10**, 1–31 (2011)

- [404] M. Deininger, J. Jung, R. Skoda, P. Helluy, C.-D. Munz, Evaluation of interface models for 3D-1D coupling of compressible Euler methods for the application on cavitating flows, in *CEMRACS'11: Multiscale Coupling of Complex Models in Scientific Computing*, vol. 38 of *ESAIM Proc., EDP Sci.*, Les Ulis, 2012, pp. 298–318
- [405] O. Delestre, S. Cordier, F. Darboux, F. James, A limitation of the hydrostatic reconstruction technique for Shallow Water equations. *C. R. Math. Acad. Sci. Paris* **350**, 677–681 (2012)
- [406] O. Delestre, F. Marche, A numerical scheme for a viscous shallow water model with friction. *J. Sci. Comput.* **48**, 41–51 (2011)
- [407] S. Dellacherie, Relaxation schemes for the multicomponent Euler system. *M2AN Math. Model. Numer. Anal.* **37**, 909–936 (2003)
- [408] S. Dellacherie, J. Jung, P. Omnes, P.-A. Raviart, Construction of modified Godunov-type schemes accurate at any Mach number for the compressible Euler system. *Math. Models Methods Appl. Sci.* **26**, 2525–2615 (2016)
- [409] S. Dellacherie, P. Omnes, F. Rieper, The influence of cell geometry on the Godunov scheme applied to the linear wave equation. *J. Comput. Phys.* **229**, 5315–5338 (2010)
- [410] A. Dervieux, G. Vijayasundaram, On numerical schemes for solving the Euler equations of gas dynamics, in *Numerical Methods for the Euler Equations of Fluid Dynamics* (Rocquencourt, 1983) (SIAM, Philadelphia, 1985), pp. 121–144
- [411] S. Deshpande, J. Mandal, Kinetic theory based new upwind methods for inviscid compressible flows, in *Proceedings of Euromekh Colloquium 224 on Kinetic Theory Aspects of Evaporation/Condensation Phenomena*, vol. 19, 1988, pp. 3, 6, 9, 32–38
- [412] J.-A. Désidéri, The computation over unstructured grids of inviscid hypersonic reactive flow by upwind finite-volume schemes, in *Computational methods in hypersonic aerodynamics*, vol. 9 of *Fluid Mech. Appl.* (Kluwer Acad. Publ., Dordrecht, 1991), pp. 387–446
- [413] J.-A. Désidéri, A. Goudjo, V. Selmin, *Third-order numerical schemes for hyperbolic problems*, INRIA Research Report, INRIA Rocquencourt, 78153 Le Chesnay, France (1987)
- [414] B. Després, Lagrangian systems of conservation laws. Invariance properties of Lagrangian systems of conservation laws, approximate Riemann solvers and the entropy condition. *Numer. Math.* **89**, 99–134 (2001)
- [415] B. Després, An explicit a priori estimate for a finite volume approximation of linear advection on non-Cartesian grids. *SIAM J. Numer. Anal.* **42**, 484–504 (2004) (electronic)
- [416] B. Després, *Lois de Conservations Eulériennes, Lagagiennes et Méthodes Numériques*. Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 68 (Springer, Berlin, 2010)

- [417] B. Després, *Numerical Methods for Eulerian and Lagrangian Conservation Laws*. Frontiers in Mathematics (Birkhäuser/Springer, Cham, 2017)
- [418] B. Després, F. Lagoutière, Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. *J. Sci. Comput.* **16**, 479–524 (2001)
- [419] B. Després, C. Mazeran, Lagrangian gas dynamics in two dimensions and Lagrangian systems. *Arch. Ration. Mech. Anal.* **178**, 327–372 (2005)
- [420] V. Desveaux, M. Zenk, C. Berthon, C. Klingenberg, A well-balanced scheme to capture non-explicit steady states in the Euler equations with gravity. *Int. J. Numer. Methods Fluids* **81**, 104–127 (2016)
- [421] G. Dimarco, R. Loubère, M.-H. Vignal, Study of a new asymptotic preserving scheme for the Euler system in the low Mach number limit. *SIAM J. Sci. Comput.* **39**, A2099–A2128 (2017)
- [422] R.J. DiPerna, Existence in the large for quasilinear hyperbolic conservation laws. *Arch. Rational Mech. Anal.* **52**, 244–257 (1973)
- [423] R.J. DiPerna, The structure of solutions to hyperbolic conservation laws, in *Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, vol. 39 of Res. Notes in Math.* (Pitman, Boston, London, 1979), pp. 1–16
- [424] R.J. DiPerna, Convergence of approximate solutions to conservation laws. *Arch. Ration. Mech. Anal.* **82**, 27–70 (1983)
- [425] R.J. DiPerna, Convergence of the viscosity method for isentropic gas dynamics. *Commun. Math. Phys.* **91**, 1–30 (1983)
- [426] R.J. DiPerna, Measure-valued solutions to conservation laws. *Arch. Ration. Mech. Anal.* **88**, 223–270 (1985)
- [427] V. Dolejší, T. Gallouët, A numerical study of a particular non-conservative hyperbolic problem. *Comput. Fluids* **37**, 1077–1091 (2008)
- [428] A. Dolezal, S.S.M. Wong, Relativistic hydrodynamics and essentially non-oscillatory shock capturing schemes. *J. Comput. Phys.* **120**, 266–277 (1995)
- [429] R. Donat, A. Marquina, Capturing shock reflections: an improved flux formula. *J. Comput. Phys.* **125**, 42–58 (1996)
- [430] A. Donato, F. Oliveri, eds., *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects*. Notes on Numerical Fluid Mechanics, vol. 43 (Friedr. Vieweg & Sohn, Braunschweig, 1993)
- [431] F. Dubois, *Code Euler 3d implicite en hypersonique et supersonique élévé. Evaluation du flux d’Osher pour l’air à l’équilibre chimique*, Aérospatiale Report, 41747 (1989)
- [432] F. Dubois, *Concavity of Thermostatic Entropy and Convexity of Lax’s Mathematical Entropy* (Rech. Aérospace, 1990), pp. 77–80

- [433] F. Dubois, Flux vector splitting and stationary contact discontinuity, in *Finite Volumes for Complex Applications II* (Hermes Sci. Publ., Paris, 1999), pp. 133–140
- [434] F. Dubois, Décomposition de flux et discontinuité de contact stationnaire. C. R. Acad. Sci. Paris Sér. I Math. **330**, 847–850 (2000)
- [435] F. Dubois, Partial Riemann problem, boundary conditions, and gas dynamics, in *Absorbing boundaries and layers, domain decomposition methods*, ed. by L. Tourrette L. Halpern (Nova Sci. Publ., Huntington, 2001), pp. 16–77
- [436] F. Dubois, B. Després, *Systèmes hyperboliques de lois de conservation. Application à la dynamique des gaz* (Editions de l’École polytechnique, 2005)
- [437] F. Dubois, P. LeFloch, Boundary conditions for nonlinear hyperbolic systems of conservation laws. J. Differ. Equ. **71**, 93–122 (1988)
- [438] F. Dubois, G. Mehlman, A non-parameterized entropy correction for Roe’s approximate Riemann solver. Numer. Math. **73**, 169–208 (1996)
- [439] F. Dubois, G. Michaux, Solution of the Euler equations around a double ellipsoidal shape using unstructured meshes and including real gas effects, in *Proceedings of the Workshop on Hypersonic Flows and Reentry Problems*, Antibes France (1990), ed. by J.-A. Désidéri, R. Glowinski, J. Périaux, vol. 2 (Springer, Berlin, 1992), pp. 358–373
- [440] B. Dubroca, Solveur de Roe positivement conservatif. C. R. Acad. Sci. Paris Sér. I Math. **329**, 827–832 (1999)
- [441] B. Dubroca, G. Gallice, Problème mixte hyperbolique pour un système de lois de conservation monodimensionnel. C. R. Acad. Sci. Paris Sér. I Math. **306**, 317–320 (1988)
- [442] B. Dubroca, G. Gallice, Résultats d’existence et d’unicité du problème mixte pour des systèmes hyperboliques de lois de conservation monodimensionnels. Commun. Partial Differ. Equ. **15**, 59–80 (1990)
- [443] B. Dubroca, J.-P. Morreeuw, An extension of Roe’s approximate Riemann solver for the approximation of Navier-Stokes equations in chemical nonequilibrium cases, in *Proceedings of the Tenth International Conference on Computing Methods in Applied Sciences and Engineering, France*, ed. by R. Glowinski (Nova Science Publishers, Inc, New York, 1992), pp. 345–372
- [444] J.K. Dukowicz, M.C. Cline, F.L. Addessio, A general topology Godunov method. J. Comput. Phys. **82**, 29–63 (1989)
- [445] J.K. Dukowicz, A.S. Dvinsky, Approximate factorization as a high order splitting for the implicit incompressible flow equations. J. Comput. Phys. **102**, 336–347 (1992)
- [446] M. Dumbser, D.S. Balsara, A new efficient formulation of the HLLEM Riemann solver for general conservative and non-conservative hyperbolic systems. J. Comput. Phys. **304**, 275–319 (2016)

- [447] M. Dumbser, D.S. Balsara, E.F. Toro, C.-D. Munz, A unified framework for the construction of one-step finite volume and discontinuous Galerkin schemes on unstructured meshes. *J. Comput. Phys.* **227**, 8209–8253 (2008)
- [448] M. Dumbser, C. Enaux, E.F. Toro, Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws. *J. Comput. Phys.* **227**, 3971–4001 (2008)
- [449] M. Dumbser, E.F. Toro, A simple extension of the Osher Riemann solver to non-conservative hyperbolic systems. *J. Sci. Comput.* **48**, 70–88 (2011)
- [450] A. Duran, Q. Liang, F. Marche, On the well-balanced numerical discretization of shallow water equations on unstructured meshes. *J. Comput. Phys.* **235**, 565–586 (2013)
- [451] L.J. Durlofsky, A triangle based mixed finite element-finite volume technique for modeling two phase flow through porous media. *J. Comput. Phys.* **105**, 252–266 (1993)
- [452] L.J. Durlofsky, B. Engquist, S. Osher, Triangle based adaptive stencils for the solution of hyperbolic conservation laws. *J. Comput. Phys.* **98**, 64–73 (1992)
- [453] P. Dutt, Stable boundary conditions and difference schemes for Navier-Stokes equations. *SIAM J. Numer. Anal.* **25**, 245–267 (1988)
- [454] W. E., Homogenization of scalar conservation laws with oscillatory forcing terms. *SIAM J. Appl. Math.* **52**, 959–972 (1992)
- [455] W. E., R.V. Kohn, The initial value problem for measure-valued solutions of a canonical  $2 \times 2$  system with linearly degenerate fields. *Commun. Pure Appl. Math.* **44**, 981–1000 (1991)
- [456] W. E., C.-W. Shu, A numerical resolution study of high order essentially non-oscillatory schemes applied to incompressible flow. *J. Comput. Phys.* **110**, 39–46 (1994)
- [457] B. Einfeldt, On Godunov-type methods for gas dynamics. *SIAM J. Numer. Anal.* **25**, 294–318 (1988)
- [458] B. Einfeldt, C.-D. Munz, P.L. Roe, B. Sjögreen, On Godunov-type methods near low densities. *J. Comput. Phys.* **92**, 273–295 (1991)
- [459] P. Embid, J. Hunter, A. Majda, Simplified asymptotic equations for the transition to detonation in reactive granular materials. *SIAM J. Appl. Math.* **52**, 1199–1237 (1992)
- [460] B. Engquist, A. Majda, Absorbing boundary conditions for numerical simulation of waves. *Proc. Nat. Acad. Sci. U.S.A.* **74**, 1765–1766 (1977)
- [461] B. Engquist, S. Osher, Stable and entropy satisfying approximations for transonic flow calculations. *Math. Comput.* **34**, 45–75 (1980)
- [462] J. Ernest, P.G. LeFloch, S. Mishra, Schemes with well-controlled dissipation. *SIAM J. Numer. Anal.* **53**, 674–699 (2015)
- [463] J.L. Estivalezes, P. Villedieu, High-order positivity-preserving kinetic schemes for the compressible Euler equations. *SIAM J. Numer. Anal.* **33**, 2050–2067 (1996)

- [464] R. Eymard, T. Gallouët, Convergence d'un schéma de type éléments finis–volumes finis pour un système formé d'une équation elliptique et d'une équation hyperbolique, *RAIRO Modél. Math. Anal. Numér.* **27**, 843–861 (1993)
- [465] R. Eymard, T. Gallouët, M. Ghilani, R. Herbin, Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes. *IMA J. Numer. Anal.* **18**, 563–594 (1998)
- [466] R. Eymard, T. Gallouët, R. Herbin, Finite volume methods, *Handbook of Numerical Analysis, Vol. VII, Handb. Numer. Anal., VII* (North-Holland, Amsterdam, 2000), pp. 713–1020
- [467] H. Fan, S. Jin, Z.-h. Teng, Zero reaction limit for hyperbolic conservation laws with source terms. *J. Differ. Equ.* **168**, 270–294 (2000). Special issue in celebration of Jack K. Hale's 70th birthday, Part 2 (Atlanta, GA/Lisbon, 1998)
- [468] H.T. Fan, J.K. Hale, Large time behavior in inhomogeneous conservation laws. *Arch. Rational Mech. Anal.* **125**, 201–216 (1993)
- [469] G. Fernandez, B. Larrouy, Hyperbolic schemes for multi-component Euler equations, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 128–138
- [470] E.D. Fernández-Nieto, D. Bresch, J. Monnier, A consistent intermediate wave speed for a well-balanced HLLC solver. *C. R. Math. Acad. Sci. Paris* **346**, 795–800 (2008)
- [471] E.D. Fernández-Nieto, J. Marin, J. Monnier, Coupling superposed 1D and 2D shallow-water models: source terms in finite volume schemes. *Comput. Fluids* **39**, 1070–1082 (2010)
- [472] E.D. Fernández-Nieto, G. Narbona-Reina, Extension of WAF type methods to non-homogeneous shallow water equations with pollutant. *J. Sci. Comput.* **36**, 193–217 (2008)
- [473] M. Fey, Multidimensional upwinding. I. The method of transport for solving the Euler equations. *J. Comput. Phys.* **143**, 159–180 (1998)
- [474] M. Fey, Multidimensional upwinding. II. Decomposition of the Euler equations into advection equations. *J. Comput. Phys.* **143**, 181–199 (1998)
- [475] M. Fey, R. Jeltsch, S. Müller, The influence of a source term, an example: chemically reacting hypersonic flow, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects* (Taormina, 1992). Notes Numer. Fluid Mech., vol. 43 (Friedr. Vieweg, Braunschweig, 1993), pp. 235–245
- [476] L. Fezoui, Résolution des équations d'Euler par un schéma de Van Leer en éléments finis, INRIA Research Report 358, INRIA Rocquencourt, Le Chesnay, France (1985)

- [477] L. Fezoui, H. Steve, Décomposition de flux de Van Leer en éléments finis, INRIA Research Report 830, INRIA Rocquencourt, 78153 Le Chesnay, France (1988)
- [478] L. Fezoui, H. Steve, V. Selmin, Simulation numérique d'écoulements compressibles 3-D par un schéma décentré en maillage non structure, INRIA Research Report 825, INRIA Rocquencourt, Le Chesnay, France (1988)
- [479] L. Fézou, B. Stoufflet, A class of implicit upwind schemes for Euler simulations with unstructured meshes. *J. Comput. Phys.* **84**, 174–206 (1989)
- [480] W. Fickett, W. Davis, *Detonation* (University of California Press, Berkeley-Los Angeles-London, 1979)
- [481] F. Filbet, S. Jin, A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. *J. Comput. Phys.* **229**, 7625–7648 (2010)
- [482] F. Filbet, A. Rambaud, Analysis of an asymptotic preserving scheme for relaxation systems. *ESAIM Math. Model. Numer. Anal.* **47**, 609–633 (2013)
- [483] P. Finaud-Guyot, C. Delenne, J. Lhomme, V. Guinot, C. Llovel, An approximate-state Riemann solver for the two-dimensional shallow water equations with porosity. *Int. J. Numer. Methods Fluids* **62**, 1299–1331 (2010)
- [484] U.S. Fjordholm, S. Mishra, E. Tadmor, Well-balanced and energy stable schemes for the shallow water equations with discontinuous topography. *J. Comput. Phys.* **230**, 5587–5609 (2011)
- [485] U.S. Fjordholm, S. Mishra, E. Tadmor, On the computation of measure-valued solutions. *Acta Numer.* **25**, 567–679 (2016)
- [486] T. Flåtten, H. Lund, Relaxation two-phase flow models and the sub-characteristic condition. *Math. Models Methods Appl. Sci.* **21**, 2379–2407 (2011)
- [487] T. Flåtten, A. Morin, S.T. Munkejord, Wave propagation in multi-component flow models. *SIAM J. Appl. Math.* **70**, 2861–2882 (2010)
- [488] A. Forestier, S. Gavrilyuk, Criterion of hyperbolicity for non-conservative quasilinear systems admitting a partially convex conservation law. *Math. Methods Appl. Sci.* **34**, 2148–2158 (2011)
- [489] H. Freistühler, A standard model of generic rotational degeneracy, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 149–158
- [490] H. Freistühler, Instability of vanishing viscosity approximation to hyperbolic systems of conservation laws with rotational invariance. *J. Differ. Equ.* **87**, 205–226 (1990)
- [491] H. Freistühler, On the Cauchy problem for a class of hyperbolic systems of conservation laws. *J. Differ. Equ.* **112**, 170–178 (1994)

- [492] H. Freistühler, E.B. Pitman, A numerical study of a rotationally degenerate hyperbolic system. I. The Riemann problem. *J. Comput. Phys.* **100**, 306–321 (1992)
- [493] D. Friedlander, D. Serre, eds., *The Handbook of Mathematical Fluid Dynamics* (Elsevier, Amsterdam, 2002)
- [494] K.O. Friedrichs, P.D. Lax, Systems of conservation equations with a convex extension. *Proc. Nat. Acad. Sci. U.S.A.* **68**, 1686–1688 (1971)
- [495] F.G. Fuchs, A.D. McMurry, S. Mishra, N.H. Risebro, K. Waagan, High order well-balanced finite volume schemes for simulating wave propagation in stratified magnetic atmospheres. *J. Comput. Phys.* **229**, 4033–4058 (2010)
- [496] G. Gallice, Matrices de Roe pour des lois de conservation générales sous forme eulérienne ou lagrangienne: application à la dynamique des gaz et à la magnétohydrodynamique. *C. R. Acad. Sci. Paris Sér. I Math.* **321**, 1069–1072 (1995)
- [497] G. Gallice, Positive and entropy stable Godunov-type schemes for gas dynamics and MHD equations in Lagrangian or Eulerian coordinates. *Numer. Math.* **94**, 673–713 (2003)
- [498] T. Gallouët, Resonance and nonlinearities, in *Hyperbolic Problems: Theory, Numerics, Applications* (Springer, Berlin, 2008), pp. 113–124
- [499] T. Gallouët, J.-M. Hérard, O. Hurisse, A.-Y. LeRoux, Well-balanced schemes versus fractional step method for hyperbolic systems with source terms. *Calcolo* **43**, 217–251 (2006)
- [500] T. Gallouët, J.-M. Hérard, N. Seguin, A hybrid scheme to compute contact discontinuities in one-dimensional Euler systems. *M2AN Math. Model. Numer. Anal.* **36**, 1133–1159 (2002)
- [501] T. Gallouët, J.-M. Hérard, N. Seguin, Some recent finite volume schemes to compute euler equations using real gas EOS. *Int. J. Numer. Methods Fluids* **39**, 1073–1138 (2002)
- [502] T. Gallouët, J.-M. Hérard, N. Seguin, On the use of symmetrizing variables for vacuums. *Calcolo* **40**, 163–194 (2003)
- [503] T. Gallouët, J.-M. Hérard, N. Seguin, Some approximate Godunov schemes to compute shallow-water equations with topography. *Comput. Fluids* **32**, 479–513 (2003)
- [504] T. Gallouët, R. Herbin, A uniqueness result for measure-valued solutions of nonlinear hyperbolic equations. *Differ. Integr. Equ.* **6**, 1383–1394 (1993)
- [505] T. Gallouët, R. Herbin, J.-C. Latché, On the weak consistency of finite volumes schemes for conservation laws on general meshes. *SeMA J.* **76**, 581–594 (2019)
- [506] T. Gallouët, J.-P. Vila, Finite volume schemes for conservation laws of mixed type. *SIAM J. Numer. Anal.* **28**, 1548–1573 (1991)
- [507] M. Garavello, B. Piccoli, *Traffic Flow on Networks*. AIMS Series on Applied Mathematics, vol. 1 (American Institute of Mathematical Sciences (AIMS), Springfield, 2006). Conservation laws models

- [508] M. Garavello, B. Piccoli, Conservation laws on complex networks. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **26**, 1925–1951 (2009)
- [509] M. Garavello, B. Piccoli, Coupling of Lighthill-Whitham-Richards and phase transition models. *J. Hyperbolic Differ. Equ.* **10**, 577–636 (2013)
- [510] I. Gasser, P. Szmolyan, A geometric singular perturbation analysis of detonation and deflagration waves. *SIAM J. Math. Anal.* **24**, 968–986 (1993)
- [511] S. Gavrilyuk, Multiphase flow modeling via Hamilton’s principle, in *Variational Models and Methods in Solid and Fluid Mechanics*. CISM Courses and Lect., vol. 535 (SpringerWienNewYork, Vienna, 2011), pp. 163–210
- [512] J.-F. Gerbeau, B. Perthame, Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation. *Discrete Contin. Dyn. Syst. Ser. B* **1**, 89–102 (2001)
- [513] H. Gilquin, J. Laurens, C. Rosier, Multi-dimensional Riemann problems for linear hyperbolic systems. I, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects* (Taormina, 1992). Notes Numer. Fluid Mech., vol. 43 (Friedr. Vieweg, Braunschweig, 1993), pp. 276–283
- [514] H. Gilquin, J. Laurens, C. Rosier, *Multi-dimensional Riemann problems for linear hyperbolic systems. II*, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects* (Taormina, 1992). Notes Numer. Fluid Mech., vol. 43 (Friedr. Vieweg, Braunschweig, 1993), pp. 284–290
- [515] V. Giovangigli, *Multicomponent Flow Modeling*. Modeling and Simulation in Science, Engineering and Technology (Birkhäuser Boston, Boston, 1999)
- [516] M. Gisclon, *Étude des conditions aux limites pour des systèmes strictement hyperboliques, via l’approximation parabolique*, PhD thesis, Université Claude Bernard-Lyon 1 (France), 1994
- [517] M. Gisclon, Étude des conditions aux limites pour un système strictement hyperbolique, via l’approximation parabolique. *J. Math. Pures Appl. (9)* **75**, 485–508 (1996)
- [518] M. Gisclon, D. Serre, Étude des conditions aux limites pour un système strictement hyperbolique via l’approximation parabolique. *C. R. Acad. Sci. Paris Sér. I Math.* **319**, 377–382 (1994)
- [519] M. Gisclon, D. Serre, Conditions aux limites pour un système strictement hyperbolique fournies par le schéma de Godunov. *RAIRO Modél. Math. Anal. Numér.* **31**, 359–380 (1997)
- [520] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*. Monographs in Mathematics, vol. 80 (Birkhäuser, Basel, 1984)
- [521] D. Givoli, Nonreflecting boundary conditions. *J. Comput. Phys.* **94**, 1–29 (1991)
- [522] D. Givoli, D. Cohen, Nonreflecting boundary conditions based on Kirchhoff-type formulae. *J. Comput. Phys.* **117**, 102–113 (1995)

- [523] P. Glaister, An approximate linearized Riemann solver for the Euler equations for real gases. *J. Comput. Phys.* **74**, 382–408 (1988)
- [524] P. Glaister, An efficient algorithm for compressible flows with real gases. *Int. J. Numer. Methods Fluids* **9**, 1269–1283 (1989)
- [525] P. Glaister, Flux difference splitting for hyperbolic systems of conservation laws with source terms. *Comput. Math. Appl.* **26**, 79–96 (1993)
- [526] P. Glaister, A weak formulation of Roe’s approximate Riemann solver applied to “barotropic” flows. *Comput. Math. Appl.* **27**, 87–90 (1994)
- [527] P. Glaister, A weak formulation of Roe’s approximate Riemann solver applied to the St. Venant equations. *J. Comput. Phys.* **116**, 189–191 (1995)
- [528] H.M. Glaz, Self-similar shock reflection in two space dimensions, in *Multidimensional Hyperbolic Problems and Computations* (Minneapolis, MN, 1989). IMA Vol. Math. Appl., vol. 29 (Springer, New York, 1991), pp. 70–88
- [529] H.M. Glaz, A.B. Wardlaw, A high-order Godunov scheme for steady supersonic gas dynamics. *J. Comput. Phys.* **58**, 157–187 (1985)
- [530] J. Glimm, Solutions in the large for nonlinear hyperbolic systems of equations. *Commun. Pure Appl. Math.* **18**, 697–715 (1965)
- [531] J. Glimm, Nonuniqueness of solutions for Riemann problems, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 169–178
- [532] J. Glimm, M.J. Graham, J.W. Grove, B.J. Plohr, eds., *Hyperbolic Problems: Theory, Numerics, Applications* World (Scientific Publishing, River Edge, 1996)
- [533] J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, S. Yaniv, Front tracking and two-dimensional Riemann problems. *Adv. in Appl. Math.* **6**, 259–290 (1985)
- [534] J. Glimm, B. Lindquist, Q. Zhang, Front tracking, oil reservoirs, engineering scale problems and mass conservation, in *Multidimensional Hyperbolic Problems and Computations* (Minneapolis, MN, 1989). IMA Vol. Math. Appl., vol. 29 (Springer, New York, 1991), pp. 123–139
- [535] J. Glimm, A.J. Majda, eds., *Multidimensional Hyperbolic Problems and Computations*. The IMA Volumes in Mathematics and Its Applications, vol. 29 (Springer, New York, 1991). Papers from the IMA Workshop held in Minneapolis, Minnesota, April 3–14, 1989
- [536] P. Goatin, P.G. LeFloch, The Riemann problem for a class of resonant hyperbolic systems of balance laws. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **21**, 881–902 (2004)
- [537] E. Godlewski, K.-C. Le Thanh, P.-A. Raviart, The numerical interface coupling of nonlinear hyperbolic systems of conservation laws. II. The case of systems. *M2AN Math. Model. Numer. Anal.* **39**, 649–692 (2005)

- [538] E. Godlewski, M. Olazabal, P.-A. Raviart, On the linearization of systems of conservation laws for fluids at a material contact discontinuity. *J. Math. Pures Appl.* (9) **78**, 1013–1042 (1999)
- [539] E. Godlewski, P.-A. Raviart, *Hyperbolic Systems of Conservation Laws*. Mathématiques & Applications (Paris) [Mathematics and Applications], vol. 3/4 (Ellipses, Paris, 1991)
- [540] E. Godlewski, P.-A. Raviart, The numerical interface coupling of nonlinear hyperbolic systems of conservation laws. I. The scalar case. *Numer. Math.* **97**, 81–130 (2004)
- [541] E. Godlewski, N. Seguin, The Riemann problem for a simple model of phase transition. *Commun. Math. Sci.* **4**, 227–247 (2006)
- [542] S.K. Godunov, A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics, *Mat. Sb. (N.S.)* **47**(89), 271–306 (1959)
- [543] S.K. Godunov, Lois de conservation d'intégrales d'énergie des équations hyperboliques, in *Nonlinear hyperbolic problems* (St. Etienne, 1986). Lecture Notes in Math., vol. 1270 (Springer, Berlin, 1987), pp. 135–149
- [544] S.K. Godunov, A. Zabrodin, M. Ivanov, A. Kraiko, G. Prokopov, *Résolution Numérique des Problèmes Multidimensionnels de la Dynamique des gaz* (“Mir”, Moscow, 1979). Translated from the Russian by Valéri Platonov
- [545] C.R. Goetz, M. Dumbser, A novel solver for the generalized Riemann problem based on a simplified LeFloch-Raviart expansion and a local space-time discontinuous Galerkin formulation. *J. Sci. Comput.* **69**, 805–840 (2016)
- [546] M. Goldberg, E. Tadmor, Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comput.* **36**, 603–626 (1981)
- [547] M. Goldberg, E. Tadmor, Convenient stability criteria for difference approximations of hyperbolic initial-boundary value problems. Math. Comput. **44**, 361–377 (1985)
- [548] M. Goldberg, E. Tadmor, Convenient stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comput.* **48**, 503–520 (1987)
- [549] J. Goodman, R.J. LeVeque, On the accuracy of stable schemes for 2D scalar conservation laws. *Math. Comput.* **45**, 15–21 (1985)
- [550] J. Goodman, A. Majda, The validity of the modified equation for nonlinear shock waves. *J. Comput. Phys.* **58**, 336–348 (1985)
- [551] J. Goodman, Z. P. Xin, Viscous limits for piecewise smooth solutions to systems of conservation laws. *Arch. Ration. Mech. Anal.* **121**, 235–265 (1992)
- [552] L. Gosse, A well-balanced scheme using non-conservative products designed for hyperbolic systems of conservation laws with source terms. *Math. Models Methods Appl. Sci.* **11**, 339–365 (2001)

- [553] L. Gosse, *Computing Qualitatively Correct Approximations of Balance Laws*. SIMAI Springer Series, vol. 2 (Springer, Milan, 2013). Exponential-fit, well-balanced and asymptotic-preserving.
- [554] L. Gosse, A.-Y. LeRoux, Un schéma-équilibre adapté aux lois de conservation scalaires non-homogènes. *C. R. Acad. Sci. Paris Sér. I Math.* **323**, 543–546 (1996)
- [555] L. Gosse, C. Makridakis, Two a posteriori error estimates for one-dimensional scalar conservation laws. *SIAM J. Numer. Anal.* **38**, 964–988 (2000)
- [556] L. Gosse, G. Toscani, An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations. *C. R. Math. Acad. Sci. Paris* **334**, 337–342 (2002)
- [557] L. Gosse, G. Toscani, Space localization and well-balanced schemes for discrete kinetic models in diffusive regimes. *SIAM J. Numer. Anal.* **41**, 641–658 (2003) (electronic)
- [558] N. Goutal, J. Sainte-Marie, A kinetic interpretation of the section-averaged Saint-Venant system for natural river hydraulics. *Int. J. Numer. Methods Fluids* **67**, 914–938 (2011)
- [559] M. F. Göz, C.-D. Munz, Approximate Riemann solvers for fluid flow with material interfaces, in *Numerical Methods for Wave Propagation* (Manchester, 1995). *Fluid Mech. Appl.*, vol. 47 (Kluwer Acad. Publ., Dordrecht, 1998), pp. 211–235
- [560] J.M. Greenberg, A.Y. LeRoux, A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM J. Numer. Anal.* **33**, 1–16 (1996)
- [561] D.F. Griffiths, A.M. Stuart, H.C. Yee, Numerical wave propagation in an advection equation with a nonlinear source term. *SIAM J. Numer. Anal.* **29**, 1244–1260 (1992)
- [562] B. Grossman, P. Cinnella, Fux-split algorithms for flows with non-equilibrium chemistry and vibrational relaxation. *J. Comput. Phys.* **88**, 131–168 (1990)
- [563] J.-L. Guermond, B. Popov, Viscous regularization of the Euler equations and entropy principles. *SIAM J. Appl. Math.* **74**, 284–305 (2014)
- [564] G. Guerra, Well-posedness for a scalar conservation law with singular nonconservative source, *J. Differ. Equ.* **206**, 438–469 (2004)
- [565] H. Guillard, R. Abgrall, *Modélisation numérique des fluides compressibles*. Series in Applied Mathematics (Paris), vol. 5 (Gauthier-Villars, Paris; North-Holland, Amsterdam, 2001)
- [566] H. Guillard, A. Murrone, On the behavior of upwind schemes in the low Mach number limit. II. *Comput. Fluids* **33**, 655–675 (2004)
- [567] H. Guillard, C. Viozat, On the behaviour of upwind schemes in the low Mach number limit. *Comput. Fluids* **28**, 63–86 (1999)
- [568] V. Guinot, Riemann solvers and boundary conditions for two-dimensional shallow water simulations. *Int. J. Numer. Methods Fluids* **41**, 1191–1219 (2003)

- [569] K.F. Gurski, An HLLC-type approximate Riemann solver for ideal magnetohydrodynamics. *SIAM J. Sci. Comput.* **25**, 2165–2187 (2004) (electronic)
- [570] B. Gustafsson, The choice of numerical boundary conditions for hyperbolic systems. *J. Comput. Phys.* **48**, 270–283 (1982)
- [571] B. Gustafsson, L. Ferm, Far field boundary conditions for steady state solutions to hyperbolic systems, in *Nonlinear Hyperbolic Problems* (St. Etienne, 1986). Lecture Notes in Math., vol. 1270 (Springer, Berlin, 1987), pp. 238–252
- [572] B. Gustafsson, H.-O. Kreiss, Boundary conditions for time-dependent problems with an artificial boundary. *J. Comput. Phys.* **30**, 333–351 (1979)
- [573] B. Gustafsson, H.-O. Kreiss, J. Oliger, *Time Dependent Problems and Difference Methods*. Pure and Applied Mathematics (New York) (Wiley, New York, 1995). A Wiley-Interscience Publication
- [574] B. Gustafsson, H.-O. Kreiss, A. Sundström, Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comput.* **26**, 649–686 (1972)
- [575] J. Haack, S. Jin, J.-G. Liu, An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations. *Commun. Comput. Phys.* **12**, 955–980 (2012)
- [576] T. Hagstrom, S.I. Hariharan, Accurate boundary conditions for exterior problems in gas dynamics. *Math. Comput.* **51**, 581–597 (1988)
- [577] M. Hall, A comparison of first and second order rezoned and Lagrangian Godunov solutions. *J. Comput. Phys.* **90**, 458–485 (1990)
- [578] E. Han, M. Hantke, G. Warnecke, Exact Riemann solutions to compressible Euler equations in ducts with discontinuous cross-section. *J. Hyperbolic Differ. Equ.* **9**, 403–449 (2012)
- [579] E. Han, M. Hantke, G. Warnecke, Criteria for nonuniqueness of Riemann solutions to compressible duct flows. *ZAMM Z. Angew. Math. Mech.* **93**, 465–475 (2013)
- [580] E. Han, G. Warnecke, Exact Riemann solutions to shallow water equations. *Q. Appl. Math.* **72**, 407–453 (2014)
- [581] R. Hannapel, T. Hauser, R. Friedrich, A comparison of ENO and TVD schemes for the computation of shock-turbulence interaction. *J. Comput. Phys.* **121**, 176–184 (1995)
- [582] B. Hanouzet, R. Natalini, Global existence of smooth solutions for partially dissipative hyperbolic systems with a convex entropy. *Arch. Ration. Mech. Anal.* **169**, 89–117 (2003)
- [583] P. Hansbo, Explicit streamline diffusion finite element methods for the compressible Euler equations in conservation variables. *J. Comput. Phys.* **109**, 274–288 (1993)
- [584] P. Hansbo, Aspects of conservation in finite element flow computations. *Comput. Methods Appl. Mech. Engrg.* **117**, 423–437 (1994)

- [585] E. Harabetian, A numerical method for computing viscous shock layers, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 220–229
- [586] E. Harabetian, A numerical method for viscous perturbations of hyperbolic conservation laws. SIAM J. Numer. Anal. **27**, 870–884 (1990)
- [587] E. Harabetian, A subcell resolution method for viscous systems of conservation laws. J. Comput. Phys. **103**, 350–358 (1992)
- [588] E. Harabetian, R. Pego, Nonconservative hybrid shock capturing schemes. J. Comput. Phys. **105**, 1–13 (1993)
- [589] A. Harten, High resolution schemes for hyperbolic conservation laws. J. Comput. Phys. **49**, 357–393 (1983)
- [590] A. Harten, On the symmetric form of systems of conservation laws with entropy. J. Comput. Phys. **49**, 151–164 (1983)
- [591] A. Harten, On a class of high resolution total-variation-stable finite-difference schemes. SIAM J. Numer. Anal. **21**, 1–23 (1984) With an appendix by Peter D. Lax.
- [592] A. Harten, Preliminary results on the extension of ENO schemes to two-dimensional problems, in *Nonlinear Hyperbolic Problems* (St. Etienne, 1986). Lecture Notes in Math., vol. 1270 (Springer, Berlin, 1987), pp. 23–40
- [593] A. Harten, J.M. Hyman, Self-adjusting grid methods for one-dimensional hyperbolic conservation laws. J. Comput. Phys. **50**, 235–269 (1983)
- [594] A. Harten, J.M. Hyman, P.D. Lax, On finite-difference approximations and entropy conditions for shocks. Commun. Pure Appl. Math. **29**, 297–322 (1976). With an appendix by B. Keyfitz
- [595] A. Harten, P.D. Lax, B. van Leer, On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**, 35–61 (1983)
- [596] H. Hattori, The Riemann problem for a van der Waals fluid with entropy rate admissibility criterion—isothermal case. Arch. Rational Mech. Anal. **92**, 247–263 (1986)
- [597] G.W. Hedstrom, Nonreflecting boundary conditions for nonlinear hyperbolic systems. J. Comput. Phys. **30**, 222–237 (1979)
- [598] A. Heibig, Existence et unicité des solutions pour certains systèmes de lois de conservation. C. R. Acad. Sci. Paris Sér. I Math. **311**, 861–866 (1990)
- [599] A. Heibig, Existence and uniqueness of solutions for some hyperbolic systems of conservation laws. Arch. Rational Mech. Anal. **126**, 79–101 (1994)
- [600] A. Heibig, D. Serre, Étude variationnelle du problème de Riemann. J. Differ. Equ. **96**, 56–88 (1992)

- [601] P. Helluy, J.-M. Hérard, H. Mathis, A well-balanced approximate Riemann solver for compressible flows in variable cross-section ducts. *J. Comput. Appl. Math.* **236**, 1976–1992 (2012)
- [602] P. Helluy, N. Seguin, Relaxation models of phase transition flows. *M2AN Math. Model. Numer. Anal.* **40**, 331–352 (2006)
- [603] C. Helzel, R.J. LeVeque, G. Warnecke, A modified fractional step method for the accurate approximation of detonation waves. *SIAM J. Sci. Comput.* **22**, 1489–1510 (2000)
- [604] P.W. Hemker, S.P. Spekreijse, Multiple grid and Osher’s scheme for the efficient solution of the steady Euler equations. *Appl. Numer. Math.* **2**, 475–493 (1986)
- [605] W. Henshaw, A scheme for the numerical solution of hyperbolic systems of conservation laws. *J. Comput. Phys.* **68**, 25–47 (1987)
- [606] J.-M. Hérard, Un modèle hyperbolique diphasique bi-fluide en milieu poreux. *C. R. Mecanique. Acad. Sci. Paris* **336**, 650–655 (2008)
- [607] J.-M. Hérard, O. Hurisse, Coupling two and one-dimensional unsteady euler equations through a thin interface. *Comput. Fluids* **5**, 651–666 (2007)
- [608] R. Herbin, J.-C. Latché, T.T. Nguyen, Explicit staggered schemes for the compressible Euler equations, in *Applied Mathematics in Savoie—AMIS 2012: Multiphase Flow in Industrial and Environmental Engineering*, vol. 40 of *ESAIM Proc., EDP Sci.*, Les Ulis, 2013, pp. 83–102
- [609] G. Hernández-Dueñas, S. Karni, Shallow water flows in channels. *J. Sci. Comput.* **48**, 190–208 (2011)
- [610] M. Herty, S. Moutari, M. Rascle, Optimization criteria for modelling intersections of vehicular traffic flow. *Netw. Heterog. Media* **1**, 275–294 (2006)
- [611] M. Herty, M. Rascle, Coupling conditions for a class of second-order models for traffic flow. *SIAM J. Math. Anal.* **38**, 595–616 (2006)
- [612] J.S. Hesthaven, *Numerical Methods for Conservation Laws*. Computational Science & Engineering, vol. 18 (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2018). From analysis to algorithms
- [613] J.S. Hesthaven, T. Warburton, *Nodal Discontinuous Galerkin Methods*. Texts in Applied Mathematics, vol. 54 (Springer, New York, 2008). Algorithms, analysis, and applications
- [614] R.L. Higdon, Absorbing boundary conditions for difference approximations to the multidimensional wave equation. *Math. Comput.* **47**, 437–459 (1986)
- [615] R.L. Higdon, Initial-boundary value problems for linear hyperbolic systems. *SIAM Rev.* **28**, 177–217 (1986)
- [616] J. Hilditch, P. Colella, A front tracking method for compressible flames in one dimension. *SIAM J. Sci. Comput.* **16**, 755–772 (1995)

- [617] C. Hirsch, *Numerical Computation of Internal and External Flows. Computational Methods for Inviscid and Viscous Flows*, vol. 2 (Wiley, Chichester, 1995). Reprint of the 1974 original publication
- [618] C. Hirsch, *Numerical Computation of Internal and External Flows. The Fundamentals of Computational Fluid Dynamics*, 2nd edn. (Elsevier, Amsterdam, 2007)
- [619] J.A.F. Hittinger, P.L. Roe, Asymptotic analysis of the Riemann problem for constant coefficient hyperbolic systems with relaxation. *ZAMM Z. Angew. Math. Mech.* **84**, 452–471 (2004)
- [620] D. Hoff, Invariant regions for systems of conservation laws. *Trans. Am. Math. Soc.* **289**, 591–610 (1985)
- [621] D. Hoff, J. Smoller, Error bounds for Glimm difference approximations for scalar conservation laws. *Trans. Am. Math. Soc.* **289**, 611–642 (1985)
- [622] H. Holden, L. Holden, R. Hoegh-Krohn, A numerical method for first order nonlinear scalar conservation laws in one dimension. *Comput. Math. Appl.* **15**, 595–602 (1988)
- [623] H. Holden, L. Holden, N.H. Risebro, Some qualitative properties of  $2 \times 2$  systems of conservation laws of mixed type, in *Nonlinear Evolution Equations That Change Type*, vol. 27 of *IMA Vol. Math. Appl.* (Springer, New York, 1990), pp. 67–78
- [624] H. Holden, K.H. Karlsen, K.-A. Lie, N.H. Risebro, *Splitting Methods for Partial Differential Equations with Rough Solutions*. EMS Series of Lectures in Mathematics, (European Mathematical Society (EMS), Zürich, 2010). Analysis and MATLAB programs
- [625] H. Holden, N.H. Risebro, A mathematical model of traffic flow on a network of roads, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects* (Taormina, 1992). Notes Numer. Fluid Mech., vol. 43 (Friedr. Vieweg, Braunschweig, 1993), pp. 329–335
- [626] H. Holden, N.H. Risebro, *Front Tracking for Hyperbolic Conservation Laws*. Applied Mathematical Sciences, 2nd edn., vol. 152 (Springer, Heidelberg, 2015)
- [627] L. Holden, On the strict hyperbolicity of the Buckley-Leverett equations for three-phase flow in a porous medium. *SIAM J. Appl. Math.* **50**, 667–682 (1990)
- [628] M. Holt, *Numerical Methods in Fluid Dynamics*. Springer Series in Computational Physics, 2nd edn. (Springer, Berlin, 1984)
- [629] J. Hong, B. Temple, The generic solution of the Riemann problem in a neighborhood of a point of resonance for systems of nonlinear balance laws. *Methods Appl. Anal.* **10**, 279–294 (2003)
- [630] E. Hopf, The partial differential equation  $u_t + uu_x = \mu u_{xx}$ . *Commun. Pure Appl. Math.* **3**, 201–230 (1950)
- [631] T.Y. Hou, P.G. LeFloch, Why nonconservative schemes converge to wrong solutions: error analysis. *Math. Comput.* **62**, 497–530 (1994)

- [632] L. Hsiao, Qualitative behavior of solutions for Riemann problems of conservation laws of mixed type, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 246–256
- [633] L. Hsiao, Admissibility criteria and admissible weak solutions of Riemann problems for conservation laws of mixed type: a summary, in *Nonlinear Evolution Equations That Change Type*. IMA Vol. Math. Appl., vol. 27 (Springer, New York, 1990), pp. 85–88
- [634] L. Hsiao, T.-P. Liu, Convergence to nonlinear diffusion waves for solutions of a system of hyperbolic conservation laws with damping. *Commun. Math. Phys.* **143**, 599–605 (1992)
- [635] L. Hsiao, D. Serre, Global existence of solutions for the system of compressible adiabatic flow through porous media. *SIAM J. Math. Anal.* **27**, 70–77 (1996)
- [636] C.-H. Hsu, S.-S. Lin, Some qualitative properties of the Riemann problem in gas dynamical combustion. *J. Differ. Equ.* **140**, 10–43 (1997)
- [637] F. Huang, P. Marcati, R. Pan, Convergence to the Barenblatt solution for the compressible Euler equations with damping and vacuum. *Arch. Ration. Mech. Anal.* **176**, 1–24 (2005)
- [638] L.P. Huang, T.-P. Liu, A conservative, piecewise-steady difference scheme for transonic nozzle flow, *Comput. Math. Appl. A* **12**, 377–388 (1986). Hyperbolic partial differential equations, III
- [639] W.H. Hui, S. Kudriakov, A unified coordinate system for solving the three-dimensional Euler equations. *J. Comput. Phys.* **172**, 235–260 (2001)
- [640] W.H. Hui, C.Y. Loh, A new Lagrangian method for steady supersonic flow computation. II. Slip-line resolution. *J. Comput. Phys.* **103**, 450–464 (1992)
- [641] W.H. Hui, C.Y. Loh, A new Lagrangian method for steady supersonic flow computation. III. Strong shocks. *J. Comput. Phys.* **103**, 465–471 (1992)
- [642] W. Hundsdorfer, B. Koren, M. van Loon, J.G. Verwer, A positive finite-difference advection scheme. *J. Comput. Phys.* **117**, 35–46 (1995)
- [643] H.T. Huynh, Accurate upwind methods for the Euler equations. *SIAM J. Numer. Anal.* **32**, 1565–1619 (1995)
- [644] E.L. Isaacson, D. Marchesin, B.J. Plohr, The structure of the Riemann solution for nonstrictly hyperbolic conservation laws, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 269–278
- [645] E.L. Isaacson, J.B. Temple, Analysis of a singular hyperbolic system of conservation laws. *J. Differ. Equ.* **65**, 250–268 (1986)

- [646] E.L. Isaacson, J.B. Temple, Nonlinear resonance in systems of conservation laws. *SIAM J. Appl. Math.* **52**, 1260–1278 (1992)
- [647] E.L. Isaacson, J.B. Temple, Convergence of the  $2 \times 2$  Godunov method for a general resonant nonlinear balance law. *SIAM J. Appl. Math.* **55**, 625–640 (1995)
- [648] A. Jameson, Numerical solution of the Euler equation for compressible inviscid fluids, in *Numerical Methods for the Euler Equations of Fluid Dynamics* (Rocquencourt, 1983) (SIAM, Philadelphia, 1985), pp. 199–245
- [649] P. Janhunen, A positive conservative method for magnetohydrodynamics based on HLL and Roe methods. *J. Comput. Phys.* **160**, 649–661 (2000)
- [650] A. Jeffrey, *Quasilinear Hyperbolic Systems and Waves* (Pitman Publishing, London-San Francisco, Calif.-Melbourne, 1976). Research Notes in Mathematics, No. 5
- [651] A. Jeffrey, T. Taniuti, *Non-Linear Wave Propagation. With Applications to Physics and Magnetohydrodynamics* (Academic Press, New York-London, 1964)
- [652] Y.N. Jeng, J.L. Chen, Truncation error analysis of the finite volume method for a model steady convective equation. *J. Comput. Phys.* **100**, 64–76 (1992)
- [653] Y.N. Jeng, U.J. Payne, An adaptive TVD limiter. *J. Comput. Phys.* **118**, 229–241 (1995)
- [654] H.K. Jenssen, I.A. Kogan, Extensions for systems of conservation laws. *Commun. Partial Differ. Equ.* **37**, 1096–1140 (2012)
- [655] H. Jiang, P.-A. Forsyth, Robust linear and nonlinear strategies for solution of the transonic Euler equations. *Int. J. Comput. Fluids* **24**(7), 753–770 (1995)
- [656] H. Jiang, Y.S. Wong, Absorbing boundary conditions for second-order hyperbolic equations. *J. Comput. Phys.* **88**, 205–231 (1990)
- [657] B.X. Jin, On an essentially conservative scheme for hyperbolic conservation laws. *J. Comput. Phys.* **112**, 308–315 (1994)
- [658] S. Jin, Runge-Kutta methods for hyperbolic conservation laws with stiff relaxation terms. *J. Comput. Phys.* **122**, 51–67 (1995)
- [659] S. Jin, Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.* **21**, 441–454 (1999) (electronic)
- [660] S. Jin, A steady-state capturing method for hyperbolic systems with geometrical source terms. *M2AN Math. Model. Numer. Anal.* **35**, 631–645 (2001)
- [661] S. Jin, Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. *Riv. Math. Univ. Parma (N.S.)* **3**, 177–216 (2012)
- [662] S. Jin, Y.J. Kim, On the computation of roll waves. *M2AN Math. Model. Numer. Anal.* **35**, 463–480 (2001)

- [663] S. Jin, C.D. Levermore, Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. *J. Comput. Phys.* **126**, 449–467 (1996)
- [664] S. Jin, J.-G. Liu, The effects of numerical viscosities. I. Slowly moving shocks. *J. Comput. Phys.* **126**, 373–389 (1996)
- [665] S. Jin, J.-G. Liu, L. Wang, A domain decomposition method for semi-linear hyperbolic systems with two-scale relaxations. *Math. Comput.* **82**, 749–779 (2013)
- [666] S. Jin, L. Pareschi, G. Toscani, Uniformly accurate diffusive relaxation schemes for multiscale transport equations. *SIAM J. Numer. Anal.* **38**, 913–936 (2000) (electronic)
- [667] S. Jin, X. Wen, Two interface-type numerical methods for computing hyperbolic systems with geometrical source terms having concentrations. *SIAM J. Sci. Comput.* **26**, 2079–2101 (2005) (electronic)
- [668] S. Jin, Z. P. Xin, The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Commun. Pure Appl. Math.* **48**, 235–276 (1995)
- [669] F. John, Formation of singularities in one-dimensional nonlinear wave propagation. *Commun. Pure Appl. Math.* **27**, 377–405 (1974)
- [670] F. John, *Partial Differential Equations*. Applied Mathematical Sciences, 4th edn., vol. 1 (Springer, New York, 1991)
- [671] P. Jorgenson, E. Turkel, Central difference TVD schemes for time dependent and steady state problems. *J. Comput. Phys.* **107**, 297–308 (1993)
- [672] V. Jovanović, C. Rohde, Error estimates for finite volume approximations of classical solutions for nonlinear systems of hyperbolic balance laws. *SIAM J. Numer. Anal.* **43**, 2423–2449 (2006) (electronic)
- [673] D. Kamowitz, *Some observations on boundary conditions for numerical conservation laws*, ICASE Report No 88–67, ICASE NASA Langley Research Center, Hampton, VA (1989)
- [674] S. Kaniel, A kinetic model for the compressible flow equations. *Indiana Univ. Math. J.* **37**, 537–563 (1988)
- [675] R. Käppeli, S. Mishra, Well-balanced schemes for the Euler equations with gravitation. *J. Comput. Phys.* **259**, 199–219 (2014)
- [676] K.H. Karlsen, M. Rascle, E. Tadmor, On the existence and compactness of a two-dimensional resonant system of conservation laws. *Commun. Math. Sci.* **5**, 253–265 (2007)
- [677] K.H. Karlsen, N.H. Risebro, J.D. Towers, Front tracking for scalar balance equations. *J. Hyperbolic Differ. Equ.* **1**, 115–148 (2004)
- [678] S. Karni, To the boundary and back—a numerical study. *Int. J. Numer. Methods Fluids* **13**, 201–216 (1991)
- [679] S. Karni, Viscous shock profiles and primitive formulations. *SIAM J. Numer. Anal.* **29**, 1592–1609 (1992)
- [680] S. Karni, Multicomponent flow calculations by a consistent primitive algorithm. *J. Comput. Phys.* **112**, 31–43 (1994)

- [681] S. Karni, S. Čanić, Computations of slowly moving shocks. *J. Comput. Phys.* **136**, 132–139 (1997)
- [682] T. Kato, The Cauchy problem for quasi-linear symmetric hyperbolic systems. *Arch. Rational Mech. Anal.* **58**, 181–205 (1975)
- [683] T. Kato, Trotter's product formula for an arbitrary pair of self-adjoint contraction semigroups, in *Topics in Functional Analysis (Essays Dedicated to M. G. Kreĭn on the Occasion of his 70th Birthday)*. Adv. in Math. Suppl. Stud., vol. 3 (Academic Press, New York-London, 1978), pp. 185–195
- [684] S. Kawashima, A. Matsumura, Stability of shock profiles in viscoelasticity with non-convex constitutive relations. *Commun. Pure Appl. Math.* **47**, 1547–1569 (1994)
- [685] J. Kevorkian, J. Yu, L. Wang, Weakly nonlinear waves for a class of linearly unstable hyperbolic conservation laws with source terms. *SIAM J. Appl. Math.* **55**, 446–484 (1995). Perturbation methods in physical mathematics (Troy, NY, 1993)
- [686] B.L. Keyfitz, Some elementary connections among nonstrictly hyperbolic conservation laws, in *Nonstrictly Hyperbolic Conservation Laws* (Anaheim, Calif., 1985). Contemp. Math., vol. 60 (Amer. Math. Soc., Providence, 1987), pp. 67–77
- [687] B.L. Keyfitz, A survey of nonstrictly hyperbolic conservation laws, in *Nonlinear Hyperbolic Problems* (St. Etienne, 1986). Lecture Notes in Math., vol. 1270 (Springer, Berlin, 1987), pp. 152–162
- [688] B.L. Keyfitz, Admissibility conditions for shocks in conservation laws that change type. *SIAM J. Math. Anal.* **22**, 1284–1292 (1991)
- [689] B.L. Keyfitz, H.C. Kranzer, A system of nonstrictly hyperbolic conservation laws arising in elasticity theory. *Arch. Rational Mech. Anal.* **72**, 219–241 (1979/1980)
- [690] B.L. Keyfitz, H.C. Kranzer, The Riemann problem for a class of hyperbolic conservation laws exhibiting a parabolic degeneracy. *J. Differ. Equ.* **47**, 35–65 (1983)
- [691] B.L. Keyfitz, H.C. Kranzer, eds., *Nonstrictly Hyperbolic Conservation Laws*. Contemporary Mathematics, vol. 60 (American Mathematical Society, Providence, 1987)
- [692] B.L. Keyfitz, H.C. Kranzer, Spaces of weighted measures for conservation laws with singular shock solutions. *J. Differ. Equ.* **118**, 420–451 (1995)
- [693] B.L. Keyfitz, R. Sanders, M. Sever, Lack of hyperbolicity in the two-fluid model for two-phase incompressible flow. *3*, 541–563 (2003). Non-linear differential equations, mechanics and bifurcation (Durham, NC, 2002)
- [694] B.L. Keyfitz, M. Shearer, eds., *Nonlinear Evolution Equations That Change Type*. The IMA Volumes in Mathematics and Its Applications, vol. 27 (Springer, New York, 1990)

- [695] B. Khobalatte, B. Perthame, Maximum principle on the entropy and second-order kinetic schemes. *Math. Comput.* **62**, 119–131 (1994)
- [696] C.A. Kim, A. Jameson, Flux limited dissipation schemes for high speed unsteady flows, in *12th AIAA Computational Fluid Dynamics Conference (San Diego, 1995)* (1995), pp. 1040–1053
- [697] S. Klainerman, A. Majda, Compressible and incompressible fluids. *Commun. Pure Appl. Math.* **35**, 629–651 (1982)
- [698] A. Klar, Convergence of alternating domain decomposition schemes for kinetic and aerodynamic equations. *Math. Methods Appl. Sci.* **18**, 649–670 (1995)
- [699] A. Klar, R. Wegener, A hierarchy of models for multilane vehicular traffic. I. Modeling. *SIAM J. Appl. Math.* **59**, 983–1001 (1999) (electronic)
- [700] R. Klein, Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics. I. One-dimensional flow. *J. Comput. Phys.* **121**, 213–237 (1995)
- [701] R. Klein, N. Botta, T. Schneider, C.D. Munz, S. Roller, A. Meister, L. Hoffmann, T. Sonar, Asymptotic adaptive methods for multi-scale problems in fluid mechanics. *J. Eng. Math.* **39**, 261–343 (2001). Special issue on practical asymptotics
- [702] C. Klingenberg, Y.-g. Lu, Existence of solutions to resonant systems of conservation laws, in *Hyperbolic Problems: Theory, Numerics, Applications* (Stony Brook, NY, 1994) (World Sci. Publ., River Edge, 1996), pp. 383–389
- [703] C. Klingenberg, S. Osher, Nonconvex scalar conservation laws in one and two space dimensions, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 289–299
- [704] P. Klingensteiner, *Hyperbolic conservation laws with source terms: errors of the shock location*, SAM Research Report 94-07, ETH Zürich, Switzerland (1994)
- [705] B. Koren, Upwind schemes for the Navier-Stokes equations, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 300–309
- [706] B. Koren, P.W. Hemker, Damped, direction-dependent multigrid for hypersonic flow computations. *Appl. Numer. Math.* **7**, 309–328 (1991)
- [707] B. Koren, S. Spekreijse, Multigrid and defect correction for the efficient solution of the steady Euler equations, in *Research in Numerical Fluid Mechanics* (Delft, 1986). Notes Numer. Fluid Mech., vol. 17 (Friedr. Vieweg, Braunschweig, 1987), pp. 87–100
- [708] H.C. Kranzer, B.L. Keyfitz, A strictly hyperbolic system of conservation laws admitting singular shocks, in *Nonlinear Evolution Equations That Change Type*. IMA Vol. Math. Appl., vol. 27 (Springer, New York, 1990), pp. 107–125

- [709] H.-O. Kreiss, On difference approximations of the dissipative type for hyperbolic differential equations. *Commun. Pure Appl. Math.* **17**, 335–353 (1964)
- [710] H.-O. Kreiss, Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comput.* **22**, 703–714 (1968)
- [711] H.-O. Kreiss, Initial boundary value problems for hyperbolic systems. *Commun. Pure Appl. Math.* **23**, 277–298 (1970)
- [712] D. Kröner, Absorbing boundary conditions for the linearized Euler equations in 2-D. *Math. Comput.* **57**, 153–167 (1991)
- [713] D. Kröner, *Numerical Schemes for Conservation Laws*. Wiley-Teubner Series Advances in Numerical Mathematics (Wiley, Chichester; B. G. Teubner, Stuttgart, 1997)
- [714] D. Kröner, P.G. LeFloch, M.-D. Thanh, The minimum entropy principle for compressible fluid flows in a nozzle with discontinuous cross-section. *M2AN Math. Model. Numer. Anal.* **42**, 425–442 (2008)
- [715] D. Kröner, S. Noelle, M. Rokyta, Convergence of higher order upwind finite volume schemes on unstructured grids for scalar conservation laws in several space dimensions. *Numer. Math.* **71**, 527–560 (1995)
- [716] D. Kröner, M. Ohlberger, A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions. *Math. Comput.* **69**, 25–39 (2000)
- [717] D. Kröner, M. Rokyta, Convergence of upwind finite volume schemes for scalar conservation laws in two dimensions. *SIAM J. Numer. Anal.* **31**, 324–343 (1994)
- [718] D. Kröner, M.D. Thanh, Numerical solutions to compressible flows in a nozzle with variable cross-section. *SIAM J. Numer. Anal.* **43**, 796–824 (2005)
- [719] S.N. Kružkov, First order quasilinear equations with several independent variables. *Mat. Sb. (N.S.)* **81**(123) 228–255 (1970)
- [720] A. Kumbaro, *Modélisation, analyse mathématique et numérique des modèles bi-fluides d'écoulement diphasique*, PhD thesis, Université Paris-Sud, Orsay, France, 1992
- [721] E. Kunhardt, C. Wu, Towards a more accurate flux corrected transport algorithm. *J. Comput. Phys.* **68**, 127–150 (1987)
- [722] A. Kurganov, S. Noelle, G. Petrova, Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton-Jacobi equations. *SIAM J. Sci. Comput.* **23**, 707–740 (2001)
- [723] A. Kurganov, G. Petrova, Central-upwind schemes for two-layer shallow water equations. *SIAM J. Sci. Comput.* **31**, 1742–1773 (2009)
- [724] D. Kuzmin, R. Löhner, eds., *Flux-Corrected Transport*. Scientific Computation (Springer, Berlin, 2005). Principles, algorithms, and applications
- [725] Y.-S. Kwon, Well-posedness for entropy solutions to multidimensional scalar conservation laws with a strong boundary condition. *J. Math. Anal. Appl.* **340**, 543–549 (2008)

- [726] Y.-S. Kwon, A. Vasseur, Strong traces for solutions to scalar conservation laws with general flux. *Arch. Ration. Mech. Anal.* **185**, 495–513 (2007)
- [727] F. Lagoutière, Stability of reconstruction schemes for scalar hyperbolic conservation laws. *Commun. Math. Sci.* **6**, 57–70 (2008)
- [728] G. Lai, On the Riemann problem for a scalar Zeldovich–von Neumann–Döring combustion model. *SIAM J. Math. Anal.* **46**, 2404–2443 (2014)
- [729] M.-H. Lallemand, Dissipative properties of Runge–Kutta schemes with upwind spatial approximation for the Euler equations, INRIA Research Report 1173, INRIA Rocquencourt, Le Chesnay, France (1990)
- [730] M.-H. Lallemand, A. Chinnayya, O. Le Metayer, Pressure relaxation procedures for multiphase compressible flows. *Int. J. Numer. Methods Fluids* **49**, 1–56 (2005)
- [731] M.-H. Lallemand, L. Fezoui, E. Perez, Un schéma multigrille en éléments finis décentré pour les équations d’Euler, INRIA Research Report 602, INRIA Rocquencourt, Le Chesnay, France (1987)
- [732] J.O. Langseth, R.J. LeVeque, A wave propagation method for three-dimensional hyperbolic conservation laws. *J. Comput. Phys.* **165**, 126–166 (2000)
- [733] J.O. Langseth, A. Tveito, R. Winther, On the convergence of operator splitting applied to conservation laws with source terms. *SIAM J. Numer. Anal.* **33**, 843–863 (1996)
- [734] D. Lannes, F. Marche, A new class of fully nonlinear and weakly dispersive Green-Naghdi models for efficient 2D simulations. *J. Comput. Phys.* **282**, 238–268 (2015)
- [735] B. Larroutuou, How to preserve the mass fractions positivity when computing compressible multi-component flows. *J. Comput. Phys.* **95**, 59–84 (1991)
- [736] B. Larroutuou, Modélisation physique, numérique et mathématique des phénomènes de propagation de flammes, in *Recent Advances in Combustion Modelling* (Rocquencourt, 1989). Ser. Adv. Math. Appl. Sci., vol. 6 (World Sci. Publ., Teaneck, 1991), pp. 65–119
- [737] B. Larroutuou, Recent progress in reactive flow computations, in *Computing Methods in Applied Sciences and Engineering (Proceedings of the Ninth Conference on Computing Methods in Applied Sciences and Engineering, INRIA, Paris 1990)*, ed. by R. Glowinski, A. Lichnewsky (SIAM, Philadelphia, 1990), pp. 249–272
- [738] B. Larroutuou, On upwind approximations of multi-dimensional multi-species flows, in *Proceedings of the first European CFD Conference*, ed. by Ch. Hirsh, J. Périault, E. Onate (Elsevier, Amsterdam, 1992), pp. 117–126
- [739] B. Larroutuou, L. Fézoui, On the equations of multi-component perfect or real gas inviscid flow, in *Nonlinear Hyperbolic Problems* (Bordeaux, 1988). Lecture Notes in Math., vol. 1402 (Springer, Berlin, 1989), pp. 69–98

- [740] C. Lattanzio, B. Piccoli, Coupling of microscopic and macroscopic traffic models at boundaries. *Math. Models Methods Appl. Sci.* **20**, 2349–2370 (2010)
- [741] C. Lattanzio, D. Serre, Convergence of a relaxation scheme for hyperbolic systems of conservation laws. *Numer. Math.* **88**, 121–134 (2001)
- [742] C. Lattanzio, A.E. Tzavaras, Structural properties of stress relaxation and convergence from viscoelasticity to polyconvex elastodynamics. *Arch. Ration. Mech. Anal.* **180**, 449–492 (2006)
- [743] C. Lattanzio, A.E. Tzavaras, Relative entropy in diffusive relaxation. *SIAM J. Math. Anal.* **45**, 1563–1584 (2013)
- [744] P.D. Lax, Shock waves and entropy, in *Contributions to nonlinear functional analysis (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1971)* (Academic Press, New York, 1971), pp. 603–634
- [745] P.D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves* (Society for Industrial and Applied Mathematics, Philadelphia, 1973). Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 11
- [746] P.D. Lax, B. Wendroff, Systems of conservation laws. *Commun. Pure Appl. Math.* **13**, 217–237 (1960)
- [747] P.D. Lax, B. Wendroff, Difference schemes for hyperbolic equations with high order of accuracy. *Commun. Pure Appl. Math.* **17**, 381–398 (1964)
- [748] E. Le Gruyer, A.Y. LeRoux, A two-dimensional Lagrange-Euler technique for gas dynamics, in *Numerical Methods for the Euler Equations of Fluid Dynamics* (Rocquencourt, 1983) (SIAM, Philadelphia, 1985), pp. 176–195
- [749] O. Le Métayer, J. Massoni, R. Saurel, Modelling evaporation fronts with reactive Riemann solvers. *J. Comput. Phys.* **205**, 567–610 (2005)
- [750] P. Le Tallec, F. Mallinger, Coupling Boltzmann and Navier-Stokes equations by half fluxes. *J. Comput. Phys.* **136**, 51–67 (1997)
- [751] M. Lécureux-Mercier, Improved stability estimates for general scalar conservation laws. *J. Hyperbolic Differ. Equ.* **8**, 727–757 (2011)
- [752] P.G. LeFloch, An existence and uniqueness result for two nonstrictly hyperbolic systems, in *Nonlinear Evolution Equations That Change Type*. IMA Vol. Math. Appl., vol. 27 (Springer, New York, 1990), pp. 126–138
- [753] P.G. LeFloch, *Hyperbolic Systems of Conservation Laws*. Lectures in Mathematics (ETH Zürich, Birkhäuser, Basel, 2002). The theory of classical and nonclassical shock waves
- [754] P.G. LeFloch, J.-G. Liu, Discrete entropy and monotonicity criteria for hyperbolic conservation laws. *C. R. Acad. Sci. Paris Sér. I Math.* **319**, 881–886 (1994)
- [755] P.G. LeFloch, T.-P. Liu, Existence theory for nonlinear hyperbolic systems in nonconservative form. *Forum Math.* **5**, 261–280 (1993)

- [756] P.G. LeFloch, P.-A. Raviart, An asymptotic expansion for the solution of the generalized Riemann problem. I. General theory. Ann. Inst. H. Poincaré Anal. Non Linéaire **5**, 179–207 (1988)
- [757] P.G. LeFloch, M.D. Thanh, The Riemann problem for fluid flows in a nozzle with discontinuous cross-section. Commun. Math. Sci. **1**, 763–797 (2003)
- [758] P.G. LeFloch, M.D. Thanh, The Riemann problem for the shallow water equations with discontinuous topography. Commun. Math. Sci. **5**, 865–885 (2007)
- [759] P.G. Lefloch, A.E. Tzavaras, Representation of weak limits and definition of nonconservative products. SIAM J. Math. Anal. **30**, 1309–1342 (1999) (electronic)
- [760] P.G. LeFloch, Z.P. Xin, Uniqueness via the adjoint problems for systems of conservation laws. Commun. Pure Appl. Math. **46**, 1499–1533 (1993)
- [761] A. Lerat, *Sur le calcul des solutions faibles des systèmes hyperboliques de lois de conservation à l'aide de schémas aux différences*, vol. 1 of ONERA Publication 1981, Office National d'Études et de Recherches Aérospatiales, Chatillon, 1981. With an English summary
- [762] A. Lerat, Propriété d'homogénéité et décomposition des flux en dynamique des gaz. J. Méc. Théor. Appl. **2**, 185–213 (1983)
- [763] A. Lerat, Difference schemes for nonlinear hyperbolic systems—a general framework, in Nonlinear hyperbolic problems (Bordeaux, 1988). Lecture Notes in Math., vol. 1402 (Springer, Berlin, 1989), pp. 12–29
- [764] A. Lerat, J. Sidès, Implicit transonic calculations without artificial viscosity or upwinding, in *Numerical Simulation of Compressible Euler Flows* (Rocquencourt, 1986). Notes Numer. Fluid Mech., vol. 26 (Friedr. Vieweg, Braunschweig, 1989), pp. 227–250
- [765] A. Lerat, Z.N. Wu, Stable conservative multidomain treatments for implicit Euler solvers. J. Comput. Phys. **123**, 45–64 (1996)
- [766] A.Y. LeRoux, On the convergence of the Godounov's scheme for first order quasi linear equations. Proc. Jpn Acad. **52**, 488–491 (1976)
- [767] A.Y. LeRoux, Convergence d'un schéma quasi d'ordre deux pour une équation quasi linéaire du premier ordre. C. R. Acad. Sci. Paris Sér. A-B **289**, A575–A577 (1979)
- [768] A.Y. LeRoux, Convergence of an accurate scheme for first order quasi-linear equations. RAIRO Anal. Numér. **15**, 151–170 (1981)
- [769] A.Y. LeRoux, Approximation of initial and boundary value problems for quasilinear first order equations, in *Computational Mathematics* (Warsaw, 1980), vol. 13 (Banach Center Publ., PWN, Warsaw, 1984), pp. 21–31
- [770] A.Y. LeRoux, Riemann solvers for some hyperbolic problems with a source term, in *Actes du 30ème Congrès d'Analyse Numérique: CANum '98* (Arles, 1998), vol. 6 of ESAIM Proc., Soc. Math. Appl. Indust., Paris, 1999, pp. 75–90 (electronic)

- [771] A.-Y. LeRoux, L. Vignon, Sur la modélisation de la stabilité d'une colonne atmosphérique avec gravité dépendant de l'altitude. *C. R. Math. Acad. Sci. Paris* **346**, 239–242 (2008)
- [772] R.J. LeVeque, Intermediate boundary conditions for time-split methods applied to hyperbolic partial differential equations. *Math. Comput.* **47**, 37–54 (1986)
- [773] R.J. LeVeque, High resolution finite volume methods on arbitrary grids via wave propagation. *J. Comput. Phys.* **78**, 36–63 (1988)
- [774] R.J. LeVeque, *Numerical Methods for Conservation Laws*. Lectures in Mathematics, 2nd edn. (ETH Zürich, Birkhäuser, Basel, 1992)
- [775] R.J. LeVeque, Wave propagation algorithms for multidimensional hyperbolic systems. *J. Comput. Phys.* **131**, 327 (1997)
- [776] R.J. LeVeque, Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. *J. Comput. Phys.* **146**, 346–365 (1998)
- [777] R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics (Cambridge University Press, Cambridge, 2002)
- [778] R.J. LeVeque, J.B. Goodman, TVD schemes in one and two space dimensions, in *Large-Scale Computations in Fluid Mechanics, Part 2* (La Jolla, Calif., 1983). Lectures in Appl. Math., vol. 22 (Amer. Math. Soc., Providence, 1985), pp. 51–62
- [779] R.J. LeVeque, M. Pelanti, A class of approximate Riemann solvers and their relation to relaxation schemes. *J. Comput. Phys.* **172**, 572–591 (2001)
- [780] R.J. LeVeque, K.-M. Shyue, One-dimensional front tracking based on high resolution wave propagation methods. *SIAM J. Sci. Comput.* **16**, 348–377 (1995)
- [781] R.J. LeVeque, K.-M. Shyue, Two-dimensional front tracking based on high resolution wave propagation methods. *J. Comput. Phys.* **123**, 354–368 (1996)
- [782] R.J. LeVeque, B. Temple, Stability of Godunov's method for a class of  $2 \times 2$  systems of conservation laws. *Trans. Am. Math. Soc.* **288**, 115–123 (1985)
- [783] R.J. LeVeque, L.N. Trefethen, On the resolvent condition in the Kreiss matrix theorem. *BIT* **24**, 584–591 (1984)
- [784] R.J. LeVeque, R. Walder, Grid alignment effects and rotated methods for computing complex flows in astrophysics, in *Proceedings of the Ninth GAMM-Conference on Numerical Methods in Fluid Mechanics* (Lausanne, 1991). Notes Numer. Fluid Mech., vol. 35 (Friedr. Vieweg, Braunschweig, 1992), pp. 376–385
- [785] R.J. LeVeque, J. Wang, A linear hyperbolic system with stiff source terms, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects* (Taormina, 1992). Notes Numer. Fluid Mech., vol. 43 (Friedr. Vieweg, Braunschweig, 1993), pp. 401–408

- [786] R.J. LeVeque, H.C. Yee, A study of numerical methods for hyperbolic conservation laws with stiff source terms. *J. Comput. Phys.* **86**, 187–210 (1990)
- [787] D. Levy, K. Powell, B. van Leer, Use of a rotated Riemann solver for the two-dimensional Euler equations. *J. Comput. Phys.* **106**, 201–214 (1993)
- [788] J. Lhomme, V. Guinot, A general approximate-state Riemann solver for hyperbolic systems of conservation laws with source terms. *Int. J. Numer. Methods Fluids* **53**, 1509–1540 (2007)
- [789] H. Li, R. Pan, Zero relaxation limit for piecewise smooth solutions to a rate-type viscoelastic system in the presence of shocks. *J. Math. Anal. Appl.* **252**, 298–324 (2000)
- [790] J. Li, P. Zhang, The transition from Zeldovich-von Neumann-Doring to Chapman-Jouguet theories for a nonconvex scalar combustion model. *SIAM J. Math. Anal.* **34**, 675–699 (2002) (electronic)
- [791] J. Li, T. Zhang, S. Yang, *The Two-Dimensional Riemann Problem in Gas Dynamics*. Pitman Monographs and Surveys in Pure and Applied Mathematics, vol. 98 (Longman, Harlow, 1998)
- [792] S. Li, An HLLC Riemann solver for magneto-hydrodynamics. *J. Comput. Phys.* **203**, 344–357 (2005)
- [793] T. Li, C.-W. Shu, M. Zhang, Stability analysis of the inverse Lax-Wendroff boundary treatment for high order upwind-biased finite difference schemes. *J. Comput. Appl. Math.* **299**, 140–158 (2016)
- [794] T.T. Li, *Global Classical Solutions for Quasilinear Hyperbolic Systems*. RAM: Research in Applied Mathematics, vol. 32 (Masson, Paris; Wiley, Chichester, 1994)
- [795] T.T. Li, W.C. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*. Duke University Mathematics Series, V (Duke University, Mathematics Department, Durham, 1985)
- [796] X.-S. Li, C.-W. Gu, An all-speed Roe-type scheme and its asymptotic analysis of low Mach number behaviour. *J. Comput. Phys.* **227**, 5144–5159 (2008)
- [797] X.-S. Li, C.-W. Gu, Mechanism of Roe-type schemes for all-speed flows and its application. *Comput. Fluids* **86**, 56–70 (2013)
- [798] C. Lian, G. Xia, C.L. Merkle, Impact of source terms on reliability of CFD algorithms. *Comput. Fluids* **39**, 1909–1922 (2010)
- [799] S.M. Liang, J.J. Chan, An improved upwind scheme for the Euler equations. *J. Comput. Phys.* **84**, 461–473 (1989)
- [800] J. Lighthill, *Waves in Fluids*. Cambridge Mathematical Library (Cambridge University Press, Cambridge, 2001). Reprint of the 1978 original
- [801] C. Lin, J.-F. Coulombel, The strong relaxation limit of the multidimensional Euler equations. *NoDEA Nonlinear Differ. Equ. Appl.* **20**, 447–461 (2013)

- [802] H.C. Lin, Dissipation additions to flux-difference splitting. *J. Comput. Phys.* **117**, 20–27 (1995)
- [803] L.W. Lin, B. Temple, J.H. Wang, Suppression of oscillations in Godunov’s method for a resonant non-strictly hyperbolic system. *SIAM J. Numer. Anal.* **32**, 841–864 (1995)
- [804] S.Y. Lin, T.M. Wu, Y.S. Chin, Upwind finite-volume method with a triangular mesh for conservation laws. *J. Comput. Phys.* **107**, 324–337 (1993)
- [805] X.-B. Lin, S. Schecter, Stability of self-similar solutions of the Dafermos regularization of a system of conservation laws. *SIAM J. Math. Anal.* **35**, 884–921 (2003) (electronic)
- [806] W.B. Lindquist, Construction of solutions for two-dimensional Riemann problems. *Comput. Math. Appl.* **12**, 615–630 (1986). Hyperbolic partial differential equations, III
- [807] W.B. Lindquist, The scalar Riemann problem in two spatial dimensions: piecewise smoothness of solutions and its breakdown. *SIAM J. Math. Anal.* **17**, 1178–1197 (1986)
- [808] W.B. Lindquist, ed., *Current Progress in Hyperbolic Systems: Riemann Problems and Computations*. Contemporary Mathematics, vol. 100 (American Mathematical Society, Providence, 1989)
- [809] P.-L. Lions, On kinetic equations, in *Proceedings of the International Congress of Mathematicians*, Vol. I, II (Kyoto, 1990) (Math. Soc. Japan, Tokyo, 1991), pp. 1173–1185
- [810] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. Volume 2: Compressible Models*, (Oxford Univ. Press, Oxford, 2013)
- [811] P.-L. Lions, B. Perthame, P.E. Souganidis, Existence and stability of entropy solutions for the hyperbolic systems of isentropic gas dynamics in Eulerian and Lagrangian coordinates. *Commun. Pure Appl. Math.* **49**, 599–638 (1996)
- [812] P.-L. Lions, B. Perthame, E. Tadmor, A kinetic formulation of multidimensional scalar conservation laws and related equations. *J. Am. Math. Soc.* **7**, 169–191 (1994)
- [813] P.-L. Lions, B. Perthame, E. Tadmor, Kinetic formulation of the isentropic gas dynamics and  $p$ -systems. *Commun. Math. Phys.* **163**, 415–431 (1994)
- [814] P.-L. Lions, P.E. Souganidis, Convergence of MUSCL and filtered schemes for scalar conservation laws and Hamilton-Jacobi equations. *Numer. Math.* **69**, 441–470 (1995)
- [815] P.L. Lions, G. Toscani, Diffusive limit for finite velocity Boltzmann kinetic models. *Rev. Mat. Iberoam.* **13**, 473–513 (1997)
- [816] M.-S. Liou, A sequel to AUSM: AUSM<sup>+</sup>. *J. Comput. Phys.* **129**, 364–382 (1996)
- [817] M.-S. Liou, C.J. Steffen, Jr., A new flux splitting scheme. *J. Comput. Phys.* **107**, 23–39 (1993)

- [818] M.-S. Liou, B. van Leer, J.-S. Shuen, Splitting of inviscid fluxes for real gases. *J. Comput. Phys.* **87**, 1–24 (1990)
- [819] R. Liska, B. Wendroff, Comparison of several difference schemes on 1D and 2D test problems for the Euler equations. *SIAM J. Sci. Comput.* **25**, 995–1017 (2003)
- [820] J.-G. Liu, Z.P. Xin, Nonlinear stability of discrete shocks for systems of conservation laws. *Arch. Ration. Mech. Anal.* **125**, 217–256 (1993)
- [821] T.P. Liu, The Riemann problem for general systems of conservation laws. *J. Differ. Equ.* **18**, 218–234 (1975)
- [822] T.P. Liu, The entropy condition and the admissibility of shocks. *J. Math. Anal. Appl.* **53**, 78–88 (1976)
- [823] T.P. Liu, Solutions in the large for the equations of nonisentropic gas dynamics. *Indiana Univ. Math. J.* **26**, 147–177 (1977)
- [824] T.P. Liu, Admissible solutions of hyperbolic conservation laws. *Mem. Am. Math. Soc.* **30**, iv+78 (1981)
- [825] T.P. Liu, Nonlinear resonance for quasilinear hyperbolic equation. *J. Math. Phys.* **28**, 2593–2602 (1987)
- [826] T.-P. Liu, Z.P. Xin, Overcompressive shock waves, in *Nonlinear Evolution Equations That Change Type*. IMA Vol. Math. Appl., vol. 27 (Springer, New York, 1990), pp. 139–145
- [827] T.-P. Liu, L.A. Ying, Nonlinear stability of strong detonations for a viscous combustion model. *SIAM J. Math. Anal.* **26**, 519–528 (1995)
- [828] X.-D. Liu, A maximum principle satisfying modification of triangle based adaptative stencils for the solution of scalar hyperbolic conservation laws. *SIAM J. Numer. Anal.* **30**, 701–716 (1993)
- [829] Y. Liu, M. Vinokur, Nonequilibrium flow computations. I. An analysis of numerical formulations of conservation laws. *J. Comput. Phys.* **83**, 373–397 (1989)
- [830] C.Y. Loh and W. H. Hui, A new Lagrangian method for steady supersonic flow computation. I. Godunov scheme. *J. Comput. Phys.* **89**, 207–240 (1990)
- [831] C.-Y. Loh, M.-S. Liou, Lagrangian solution of supersonic real gas flows. *J. Comput. Phys.* **104**, 150–161 (1993)
- [832] C. Lowe, J. Clarke, A class of exact solutions for the Euler equations with sources. I. *Math. Comput. Model.* **36**, 275–291 (2002). Mathematical modelling of nonlinear systems (Leeds, 1999)
- [833] B.J. Lucier, Error bounds for the methods of Glimm, Godunov and LeVeque. *SIAM J. Numer. Anal.* **22**, 1074–1081 (1985)
- [834] B.J. Lucier, A moving mesh numerical method for hyperbolic conservation laws. *Math. Comput.* **46**, 59–69 (1986)
- [835] H. Lund, A hierarchy of relaxation models for two-phase flow. *SIAM J. Appl. Math.* **72**, 1713–1741 (2012)
- [836] C. Lytton, Solution of the Euler equations for transonic flow over a lifting aerofoil—the Bernoulli formulation (Roe-Lytton method). *J. Comput. Phys.* **73**, 395–431 (1987)

- [837] M. Macrossan, The equilibrium flux method for the calculation of flows with nonequilibrium chemical reactions. *J. Comput. Phys.* **80**, 204–231 (1989)
- [838] P.-H. Maire, R. Abgrall, J. Breil, J. Ovadia, A cell-centered Lagrangian scheme for two-dimensional compressible flow problems. *SIAM J. Sci. Comput.* **29**, 1781–1824 (2007) (electronic)
- [839] A. Majda, A qualitative model for dynamic combustion. *SIAM J. Appl. Math.* **41**, 70–93 (1981)
- [840] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*. Applied Mathematical Sciences, vol. 53 (Springer, New York, 1984)
- [841] A. Majda, One perspective on open problems in multi-dimensional conservation laws, in *Multidimensional Hyperbolic Problems and Computations* (Minneapolis, MN, 1989). IMA Vol. Math. Appl., vol. 29 (Springer, New York, 1991), pp. 217–238
- [842] A. Majda, S. Osher, Numerical viscosity and the entropy condition. *Commun. Pure Appl. Math.* **32**, 797–838 (1979)
- [843] A. Majda, R.L. Pego, Stable viscosity matrices for systems of conservation laws. *J. Differ. Equ.* **56**, 229–262 (1985)
- [844] A. Majda, J. Sethian, The derivation and numerical solution of the equations for zero mach number combustion. *Combust. Sci. Technol.* **42**, 185–205 (1985)
- [845] J.C. Mandal, S.M. Deshpande, Higher order accurate kinetic flux vector splitting method for Euler equations, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 384–392
- [846] J.C. Mandal, S.M. Deshpande, Kinetic flux vector splitting for Euler equations. *Comput. Fluids* **23**, 447–478 (1994)
- [847] D.K. Mao, A treatment of discontinuities for finite difference methods. *J. Comput. Phys.* **103**, 359–369 (1992)
- [848] D.K. Mao, A treatment of discontinuities for finite difference methods in the two-dimensional case. *J. Comput. Phys.* **104**, 377–397 (1993)
- [849] P. Marcati, A. Milani, The one-dimensional Darcy’s law as the limit of a compressible Euler flow. *J. Differ. Equ.* **84**, 129–147 (1990)
- [850] F. Marche, P. Bonneton, P. Fabrie, N. Seguin, Evaluation of well-balanced bore-capturing schemes for 2D wetting and drying processes. *Int. J. Numer. Methods Fluids* **53**, 867–894 (2007)
- [851] G. Marshall, B. Plohr, A random choice method for two-dimensional steady supersonic shock wave diffraction problems. *J. Comput. Phys.* **56**, 410–427 (1992)
- [852] J.M. Martí, E. Müller, The analytical solution of the Riemann problem in relativistic hydrodynamics. *J. Fluid Mech.* **258**, 317–333 (1994)

- [853] J.M. Martí, E. Müller, Extension of the piecewise parabolic method to one-dimensional relativistic hydrodynamics. *J. Comput. Phys.* **123**, 1–14 (1996)
- [854] C. Mascia, R. Natalini, On relaxation hyperbolic systems violating the Shizuta-Kawashima condition. *Arch. Ration. Mech. Anal.* **195**, 729–762 (2010)
- [855] J.-M. Masella, I. Faille, T. Gallouët, On an approximate Godunov scheme. *Int. J. Comput. Fluid Dyn.* **12**, 133–149 (1999)
- [856] H. Mathis, C. Cancès, E. Godlewski, N. Seguin, Dynamic model adaptation for multiscale simulation of hyperbolic systems with relaxation. *J. Sci. Comput.* **63**, 820–861
- [857] D.J. Mavriplis, An advancing front Delaunay triangulation algorithm designed for robustness. *J. Comput. Phys.* **117**, 90–101 (1995)
- [858] P.A. Mazet, On a variational approach to conservative hyperbolic systems. *Rech. Aéronaut.* (1983)
- [859] P.-A. Mazet, F. Bourdel, Multidimensional case of an entropic variational formulation of conservative hyperbolic systems. *Rech. Aéronaut.* 369–378 (1984)
- [860] G. Mehlman, An approximate Riemann solver for fluid systems based on a shock curve decomposition, in *Third International Conference on Hyperbolic Problems, Vol. I, II* (Uppsala, 1990) (Studentlitteratur, Lund, 1991), pp. 727–741
- [861] R. Menikoff, Analogies between Riemann problem for 1-D fluid dynamics and 2-D steady supersonic flow, in *Current progress in Hyperbolic Systems: Riemann Problems and Computations* (Brunswick, ME, 1988). Contemp. Math., vol. 100 (Amer. Math. Soc., Providence, 1989), pp. 225–240
- [862] R. Menikoff, B.J. Plohr, The Riemann problem for fluid flow of real materials. *Rev. Mod. Phys.* **61**, 75–130 (1989)
- [863] R. Menina, R. Saurel, M. Zereg, L. Houas, Modelling gas dynamics in 1d ducts with abrupt area change. *Shock Waves* **21**, 451–466 (2011)
- [864] S. Menne, C. Weiland, D. D'Ambrosio, M. Pandolfi, Comparison of real gas simulations using different numerical methods. *Comput. Fluids* **24**, 189–2008 (1995)
- [865] B. Merlet, J. Vovelle, Error estimate for finite volume scheme. *Numer. Math.* **106**, 129–155 (2007)
- [866] G. Métivier, Interaction de chocs, in *Bony-Sjöstrand-Meyer Seminar, 1984–1985* (École Polytech., Palaiseau, 1985), pp. Exp. No. 6, 18
- [867] G. Métivier, Interaction de deux chocs pour un système de deux lois de conservation, en dimension deux d'espace. *Trans. Am. Math. Soc.* **296**, 431–479 (1986)
- [868] G. Métivier, Stability of multi-dimensional weak shocks. *Commun. Partial Differ. Equ.* **15**, 983–1028 (1990)

- [869] G. Métivier, S. Schochet, The incompressible limit of the non-isentropic Euler equations. *Arch. Ration. Mech. Anal.* **158**, 61–90 (2001)
- [870] S. Mishra, On the convergence of numerical schemes for hyperbolic systems of conservation laws, in *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. IV. Invited Lectures* (World Sci. Publ., Hackensack, 2018), pp. 3641–3668
- [871] S. Mishra, L.V. Spinolo, Accurate numerical schemes for approximating initial-boundary value problems for systems of conservation laws. *J. Hyperbolic Differ. Equ.* **12**, 61–86 (2015)
- [872] J.-L. Montagné, *Noncentered Numerical Schemes in Two-Dimensional Gasdynamics* (Rech. Aérospat., 1984), pp. 323–338
- [873] J.-L. Montagné, H. C. Yee, M. Vinokur, Comparative study of high-resolution shock-capturing schemes for a real gas. *AIAA J* **27**, 1332–1346 (1989)
- [874] P. Montarnal, C.-W. Shu, Real gas computation using an energy relaxation method and high-order WENO schemes. *J. Comput. Phys.* **148**, 59–80 (1999)
- [875] G. Montecinos, C.E. Castro, M. Dumbser, E.F. Toro, Comparison of solvers for the generalized Riemann problem for hyperbolic systems with source terms. *J. Comput. Phys.* **231**, 6472–6494 (2012)
- [876] T. Morales de Luna, M.J. Castro Díaz, C. Parés, Reliability of first order numerical schemes for solving shallow water system over abrupt topography. *Appl. Math. Comput.* **219**, 9012–9032 (2013)
- [877] K. Morton, P. Sweby, A comparison of flux limited difference methods and characteristic Galerkin methods for shock modelling. *J. Comput. Phys.* **73**, 203–229 (1987)
- [878] W. Mulder, S. Osher, J.A. Sethian, Computing interface motion in compressible gas dynamics. *J. Comput. Phys.* **100**, 209–228 (1992)
- [879] W.A. Mulder, B. van Leer, Experiments with implicit upwind methods for the Euler equations. *J. Comput. Phys.* **59**, 232–246 (1985)
- [880] E. Müller, Flux vector splitting for the Euler equations for real gases. *J. Comput. Phys.* **79**, 227–230 (1988)
- [881] S. Müller, A. Voß, The Riemann problem for the Euler equations with nonconvex and nonsmooth equation of state: construction of wave curves. *SIAM J. Sci. Comput.* **28**, 651–681 (2006)
- [882] S.T. Munkejord, A numerical study of two-fluid models with pressure and velocity relaxation. *Adv. Appl. Math. Mech.* **2**, 131–159 (2010)
- [883] C.-D. Munz, On Godunov-type schemes for Lagrangian gas dynamics. *SIAM J. Numer. Anal.* **31**, 17–42 (1994)
- [884] C.-D. Munz, Computational fluid dynamics and aeroacoustics for low Mach number flow, in *Hyperbolic Partial Differential Equations* (Hamburg, 2001) (Friedr. Vieweg, Braunschweig, 2002), pp. 269–320

- [885] C.-D. Munz, M. Dumbser, S. Roller, Linearized acoustic perturbation equations for low Mach number flow with variable density and temperature. *J. Comput. Phys.* **224**, 352–364 (2007)
- [886] C.-D. Munz, L. Schmidt, Numerical simulations of compressible hydrodynamic instabilities with high resolution schemes, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications* (Aachen, 1988). Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 456–465
- [887] F. Murat, Compacité par compensation, *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **5**, 489–507 (1978)
- [888] J. Murillo, P. García-Navarro, Weak solutions for partial differential equations with source terms: application to the shallow water equations. *J. Comput. Phys.* **229**, 4327–4368 (2010)
- [889] J. Murillo, P. García-Navarro, Augmented versions of the HLL and HLLC Riemann solvers including source terms in one and two dimensions for shallow flow applications. *J. Comput. Phys.* **231**, 6861–6906 (2012)
- [890] A. Murrone, H. Guillard, A five equation reduced model for compressible two phase flow problems. *J. Comput. Phys.* **202**, 664–698 (2005)
- [891] G. Naldi, L. Pareschi, Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation. *SIAM J. Numer. Anal.* **37**, 1246–1270 (2000) (electronic)
- [892] G. Narbona-Reina, J.D.D. Zabsonré, E.D. Fernández-Nieto, D. Bresch, Derivation of a bilayer model for shallow water equations with viscosity. Numerical validation. *CMES Comput. Model. Eng. Sci.* **43**, 27–71 (2009)
- [893] R. Natalini, Convergence to equilibrium for the relaxation approximations of conservation laws. *Commun. Pure Appl. Math.* **49**, 795–823 (1996)
- [894] R. Natalini, A discrete kinetic approximation of entropy solutions to multidimensional scalar conservation laws. *J. Differ. Equ.* **148**, 292–317 (1998)
- [895] R. Natalini, Recent results on hyperbolic relaxation problems, in *Analysis of Systems of Conservation Laws* (Aachen, 1997). Chapman & Hall/CRC Monogr. Surv. Pure Appl. Math., vol. 99 (Chapman & Hall/CRC, Boca Raton, 1999), pp. 128–198
- [896] M. Ndjinga, Influence of interfacial pressure on the hyperbolicity of the two-fluid model. *C. R. Math. Acad. Sci. Paris* **344**, 407–412 (2007)
- [897] H. Nessyahu, E. Tadmor, Nonoscillatory central differencing for hyperbolic conservation laws. *J. Comput. Phys.* **87**, 408–463 (1990)
- [898] H. Nessyahu, E. Tadmor, The convergence rate of approximate solutions for nonlinear scalar conservation laws. *SIAM J. Numer. Anal.* **29**, 1505–1519 (1992)

- [899] H. Nessyahu, T. Tassa, Convergence rate of approximate solutions to conservation laws with initial rarefactions. *SIAM J. Numer. Anal.* **31**, 628–654 (1994)
- [900] J. Neusser, V. Schleper, Numerical schemes for the coupling of compressible and incompressible fluids in several space dimensions. *Appl. Math. Comput.* **304**, 65–82 (2017)
- [901] F. Nicoud, T. Poinsot, Boundary conditions for compressible unsteady flows, in *Absorbing Boundaries and Layers, Domain Decomposition Methods* (Nova Sci. Publ., Huntington, NY, 2001), pp. 78–108
- [902] T. Nishida, J. Smoller, A class of convergent finite difference schemes for certain nonlinear parabolic systems. *Commun. Pure Appl. Math.* **36**, 785–808 (1983)
- [903] P. Noble, Roll-waves in general hyperbolic systems with source terms. *SIAM J. Appl. Math.* **67**, 1202–1212 (2007) (electronic)
- [904] S. Noelle, Hyperbolic systems of conservation laws, the Weyl equation, and multidimensional upwinding. *J. Comput. Phys.* **115**, 22–26 (1994)
- [905] S. Noelle, Convergence of higher order finite volume schemes on irregular grids. *Adv. Comput. Math.* **3**, 197–218 (1995)
- [906] S. Noelle, A note on entropy inequalities and error estimates for higher-order accurate finite volume schemes on irregular families of grids. *Math. Comput.* **65**, 1155–1163 (1996)
- [907] S. Noelle, G. Bispfen, K.R. Arun, M. Lukáčová-Medviďová, C.-D. Munz, A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics. *SIAM J. Sci. Comput.* **36**, B989–B1024 (2014)
- [908] S. Noelle, N. Pankratz, G. Puppo, J.R. Natvig, Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *J. Comput. Phys.* **213**, 474–499 (2006)
- [909] S. Noelle, Y. Xing, C.-W. Shu, High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. *J. Comput. Phys.* **226**, 29–58 (2007)
- [910] S. Noelle, Y. Xing, C.-W. Shu, High-order well-balanced schemes, in *Numerical Methods for Balance Laws*. Quad. Mat., vol. 24. Dept. Math., Seconda Univ. Napoli, Caserta (2009), pp. 1–66
- [911] A. Nouri, A. Omrane, J.P. Vila, Boundary conditions for scalar conservation laws from a kinetic point of view. *J. Stat. Phys.* **94**, 779–804 (1999)
- [912] A. Noussair, Riemann problem with nonlinear resonance effects and well-balanced Godunov scheme for shallow fluid flow past an obstacle. *SIAM J. Numer. Anal.* **39**, 52–72 (2001)
- [913] M. Ohlberger, A review of a posteriori error control and adaptivity for approximations of non-linear conservation laws. *Int. J. Numer. Methods Fluids* **59**, 333–354 (2009)
- [914] O.A. Olešník, Discontinuous solutions of non-linear differential equations. *Am. Math. Soc. Transl. (2)* **26**, 95–172 (1963)

- [915] J. Oliger, A. Sundström, Theoretical and practical aspects of some initial boundary value problems in fluid dynamics. *SIAM J. Appl. Math.* **35**, 419–446 (1978)
- [916] E.S. Oran, J.P. Boris, Numerical simulation of reacting flows, in *Fluid Dynamical Aspects of Combustion Theory*. Pitman Res. Notes Math. Ser., vol. 223 (Longman Sci. Tech., Harlow, 1991), pp. 268–304
- [917] S. Osher, Riemann solvers, the entropy condition, and difference approximations. *SIAM J. Numer. Anal.* **21**, 217–235 (1984)
- [918] S. Osher, Convergence of generalized MUSCL schemes. *SIAM J. Numer. Anal.* **22**, 947–961 (1985)
- [919] S. Osher, S. Chakravarthy, Upwind schemes and boundary conditions with applications to Euler equations in general geometries. *J. Comput. Phys.* **50**, 447–481 (1983)
- [920] S. Osher, S. Chakravarthy, High resolution schemes and the entropy condition. *SIAM J. Numer. Anal.* **21**, 955–984 (1984)
- [921] S. Osher, R. Sanders, Numerical approximations to nonlinear conservation laws with locally varying time and space grids. *Math. Comput.* **41**, 321–336 (1983)
- [922] S. Osher, F. Solomon, Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comput.* **38**, 339–374 (1982)
- [923] S. Osher, P.K. Sweby, Recent developments in the numerical solution of nonlinear conservation laws, in *The State of the Art in Numerical Analysis (Birmingham, 1986)*. Inst. Math. Appl. Conf. Ser. New Ser., vol. 9 (Oxford Univ. Press, New York, 1987), pp. 681–701
- [924] S. Osher, E. Tadmor, On the convergence of difference approximations to scalar conservation laws. *Math. Comput.* **50**, 19–51 (1988)
- [925] M. Pandolfi, D. D'Ambrosio, Numerical instabilities in upwind methods: analysis and cures for the “carbuncle” phenomenon. *J. Comput. Phys.* **166**, 271–301 (2001)
- [926] E.Y. Panov, Uniqueness of the solution of the Cauchy problem for a first-order quasilinear equation with an admissible strictly convex entropy. *Mat. Zametki* **55**, 116–129, 159 (1994)
- [927] E.Y. Panov, Existence of strong traces for generalized solutions of multidimensional scalar conservation laws. *J. Hyperbolic Differ. Equ.* **2**, 885–908 (2005)
- [928] E.Y. Panov, On existence and uniqueness of entropy solutions to the Cauchy problem for a conservation law with discontinuous flux. *J. Hyperbolic Differ. Equ.* **6**, 525–548 (2009)
- [929] C. Parés, M. Castro, On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems. *M2AN Math. Model. Numer. Anal.* **38**, 821–852 (2004)
- [930] M. Parisot, J.-P. Vila, Numerical scheme for multilayer shallow-water model in the low-Froude number regime. *C. R. Math. Acad. Sci. Paris* **352**, 953–957 (2014)

- [931] R.L. Pego, Phase transitions in one-dimensional nonlinear viscoelasticity: admissibility and stability. *Arch. Rational Mech. Anal.* **97**, 353–394 (1987)
- [932] M. Pelanti, F. Bouchut, A. Mangeney, A Roe-type scheme for two-phase shallow granular flows over variable topography. *M2AN Math. Model. Numer. Anal.* **42**, 851–885 (2008)
- [933] M. Pelanti, F. Bouchut, A. Mangeney, A Riemann solver for single-phase and two-phase shallow flow models based on relaxation. Relations with Roe and VFRoe solvers. *J. Comput. Phys.* **230**, 515–550 (2011)
- [934] R.B. Pember, Numerical methods for hyperbolic conservation laws with stiff relaxation. I. Spurious solutions. *SIAM J. Appl. Math.* **53**, 1293–1330 (1993)
- [935] R.B. Pember, Numerical methods for hyperbolic conservation laws with stiff relaxation. II. Higher-order Godunov methods. *SIAM J. Sci. Comput.* **14**, 824–859 (1993)
- [936] R.B. Pember, J.B. Bell, P. Colella, W.Y. Crutchfield, M.L. Welcome, An adaptive Cartesian grid method for unsteady compressible flow in irregular regions. *J. Comput. Phys.* **120**, 278–304 (1995)
- [937] Y.-J. Peng, Solutions faibles globales pour l'équation d'Euler d'un fluide compressible avec de grandes données initiales. *Comm. Partial Differ. Equ.* **17**, 161–187 (1992)
- [938] Y.-J. Peng, Explicit solutions for  $2 \times 2$  linearly degenerate systems. *Appl. Math. Lett.* **11**, 75–78 (1998)
- [939] Y.-J. Peng, Euler-Lagrange change of variables in conservation laws. *Nonlinearity* **20**, 1927–1953 (2007)
- [940] B. Perthame, Global existence to the BGK model of Boltzmann equation. *J. Differ. Equ.* **82**, 191–205 (1989)
- [941] B. Perthame, Boltzmann type schemes for gas dynamics and the entropy property. *SIAM J. Numer. Anal.* **27**, 1405–1421 (1990)
- [942] B. Perthame, Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions. *SIAM J. Numer. Anal.* **29**, 1–19 (1992)
- [943] B. Perthame, Convergence of  $N$ -schemes for linear advection equations, in *Trends in applications of mathematics to mechanics (Lisbon, 1994)*. Pitman Monogr. Surveys Pure Appl. Math., vol. 77 (Longman, Harlow, 1995), pp. 323–333
- [944] B. Perthame, *Kinetic Formulation of Conservation Laws*. Oxford Lecture Series in Mathematics and Its Applications, vol. 21 (Oxford University Press, Oxford, 2002)
- [945] B. Perthame, M. Pulvirenti, Weighted  $L^\infty$  bounds and uniqueness for the Boltzmann BGK model. *Arch. Rational Mech. Anal.* **125**, 289–295 (1993)

- [946] B. Perthame, Y. Qiu, A variant of Van Leer's method for multidimensional systems of conservation laws. *J. Comput. Phys.* **112**, 370–381 (1994)
- [947] B. Perthame, Y. Qiu, B. Stoufflet, Sur la convergence des schémas “fluctuation-splitting” pour l'advection et leur utilisation en dynamique des gaz. *C. R. Acad. Sci. Paris Sér. I Math.* **319**, 283–288 (1994)
- [948] B. Perthame, Y. Qiu, B. Stoufflet, Kinetic discretization of gas dynamics using fluctuation-splitting schemes, in *Hyperbolic Problems: Theory, Numerics, Applications (Stony Brook, NY, 1994)* (World Sci. Publ., River Edge, 1996), pp. 207–216
- [949] B. Perthame, C. Simeoni, A kinetic scheme for the Saint-Venant system with a source term. *Calcolo* **38**, 201–231 (2001)
- [950] B. Perthame, E. Tadmor, A kinetic equation with kinetic entropy functions for scalar conservation laws. *Commun. Math. Phys.* **136**, 501–517 (1991)
- [951] R. Peyret, T.D. Taylor, *Computational Methods for Fluid Flow*. Springer Series in Computational Physics, 2nd edn. (Springer, New York, 1985)
- [952] J. Pike, Grid adaptative algorithms for the solution of the Euler equations on irregular grids. *J. Comput. Phys.* **71**, 194–223 (1987)
- [953] E.B. Pitman, L. Le, A two-fluid model for avalanche and debris flows. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **363**, 1573–1601 (2005)
- [954] G. Poëtte, B. Després, D. Lucor, Uncertainty quantification for systems of conservation laws. *J. Comput. Phys.* **228**, 2443–2467 (2009)
- [955] T. Poinsot, S.M. Candel, The influence of differencing and CFL number on implicit time-dependent nonlinear calculations. *J. Comput. Phys.* **62**, 282–296 (1986)
- [956] T. Poinsot, S.K. Lele, Boundary conditions for direct simulations of compressible viscous flows. *J. Comput. Phys.* **101**, 104–129 (1992)
- [957] K.G. Powell, P.L. Roe, T.J. Linde, T.I. Gombosi, D.L. De Zeeuw, A solution-adaptive upwind scheme for ideal magnetohydrodynamics. *J. Comput. Phys.* **154**, 284–309 (1999)
- [958] K.H. Prendergast, K. Xu, Numerical hydrodynamics from gas-kinetic theory. *J. Comput. Phys.* **109**, 53–66 (1993)
- [959] D. Pullin, Direct simulation methods for compressible inviscid ideal gas flow. *J. Comput. Phys.* **34**, 231–244 (1980)
- [960] G. Puppo, G. Russo (eds.), *Numerical Methods for Balance Laws*. Quaderni di Matematica [Mathematics Series], vol. 24. Department of Mathematics, Seconda Università di Napoli, Caserta, 2009
- [961] J. Qian, J. Li, S. Wang, The generalized Riemann problems for compressible fluid flows: towards high order. *J. Comput. Phys.* **259**, 358–389 (2014)

- [962] J.J. Quirk, A contribution to the great Riemann solver debate, *Internat. J. Numer. Methods Fluids* **18**, 555–574 (1994)
- [963] J.J. Quirk, Godunov-type schemes applied to detonation flows, ICASE Report No 93–15, ICASE NASA Langley Research Center, Hampton, VA (1993)
- [964] R. Radespiel, N. Kroll, Accurate flux vector splitting for shocks and shear layers. *J. Comput. Phys.* **121**, 66–78 (1995)
- [965] V.H. Ransom, D.L. Hicks, Hyperbolic two-pressure models for two-phase flow. *J. Comput. Phys.* **53**, 124–151 (1984)
- [966] M. Rascle, Convergence of approximate solutions to some systems of conservative laws: a conjecture on the product of the Riemann invariants, in *Oscillation Theory, Computation, and Methods of Compensated Compactness (Minneapolis, Minn., 1985)*. IMA Vol. Math. Appl., vol. 2 (Springer, New York, 1986), pp. 275–288
- [967] J. Rauch, BV estimates fail for most quasilinear hyperbolic systems in dimensions greater than one. *Commun. Math. Phys.* **106**, 481–484 (1986)
- [968] P.-A. Raviart, L. Sainsaulieu, Nonconservative hyperbolic systems and two-phase flows, in *International Conference on Differential Equations, Vol. 1, 2 (Barcelona, 1991)* (World Sci. Publ., River Edge, 1993), pp. 225–233.
- [969] P.-A. Raviart, L. Sainsaulieu, A nonconservative hyperbolic system modeling spray dynamics. I. Solution of the Riemann problem. *Math. Models Methods Appl. Sci.* **5**, 297–333 (1995)
- [970] D. Ray, P. Chandrashekhar, U.S. Fjordholm, S. Mishra, Entropy stable scheme on two-dimensional unstructured grids for Euler equations. *Commun. Comput. Phys.* **19**, 1111–1140 (2016)
- [971] G.A. Reigstad, T. Flåtten, N. Erland Haugen, T. Ytrehus, Coupling constants and the generalized Riemann problem for isothermal junction flow. *J. Hyperbolic Differ. Equ.* **12**, 37–59 (2015)
- [972] R. Reitz, One-dimensional compressible gas dynamics calculations using the Boltzmann equation. *J. Comput. Phys.* **42**, 108–123 (1981)
- [973] G.L. Richard, S.L. Gavrilyuk, A new model of roll waves: comparison with Brock’s experiments. *J. Fluid Mech.* **698**, 374–405 (2012)
- [974] R.D. Richtmyer, K.W. Morton, *Difference Methods for Initial-Value Problems*. Interscience Tracts in Pure and Applied Mathematics, No. 4, 2nd edn. (Interscience Publishers/Wiley, New York/London/Sydney, 1967)
- [975] W.J. Rider, A review of approximate Riemann solvers with Godunov’s method in Lagrangian coordinates. *Comput. Fluids* **23**, 397–413 (1994)
- [976] P.J. Roache, *Computational Fluid Dynamics* (Hermosa Publishers, Albuquerque, 1976). With an appendix (“On artificial viscosity”) reprinted from *J. Comput. Phys.* **10**(2), 169–184 (1972). Revised printing

- [977] T.W. Roberts, The behavior of flux difference splitting schemes near slowly moving shock waves. *J. Comput. Phys.* **90**, 141–160 (1990)
- [978] D. Rochette, S. Clain, W. Bussière, Unsteady compressible flow in ducts with varying cross-section: comparison between the nonconservative Euler system and the axisymmetric flow model. *Comput. Fluids* **53**, 53–78 (2012)
- [979] P.L. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.* **43**, 357–372 (1981)
- [980] P.L. Roe, Upwind schemes using various formulations of the Euler equations, in *Proceedings of the INRIA Workshop, Rocquencourt, France (1983)*, ed. by F. Angrand, A. Dervieux, J.A. Desideri, R. Glowinski (SIAM, Philadelphia, 1985)
- [981] P.L. Roe, Some contributions to the modelling of discontinuous flows, in *Large-Scale Computations in Fluid Mechanics (Proceedings of the Fifteenth Summer Seminar on Applied Mathematics, La Jolla, CA, June 27-July 8, 1983. Part 2 (A85-48201 23-34). Providence, RI)* (American Mathematical Society, 1985), pp. 163–193
- [982] P.L. Roe, Discrete models for the numerical analysis of time-dependent multidimensional gas dynamics. *J. Comput. Phys.* **63**, 458–476 (1986)
- [983] P.L. Roe, Upwind differencing schemes for hyperbolic conservation laws with source terms, in *Nonlinear Hyperbolic Problems (St. Etienne, 1986)*. Lecture Notes in Math., vol. 1270 (Springer, Berlin, 1987), pp. 41–51
- [984] P.L. Roe, Discontinuous solutions to hyperbolic systems under operator splitting. *Numer. Methods Partial Differ. Equ.* **7**, 277–297 (1991)
- [985] P.L. Roe, Sonic flux formulae. *SIAM J. Sci. Stat. Comput.* **13**, 611–630 (1992)
- [986] P.L. Roe, Mathematical problems associated with computing flow of real gases, in *Computational Aeronautical Fluid Dynamics (Antibes, 1989)*. Inst. Math. Appl. Conf. Ser. New Ser., , vol. 44 (Oxford Univ. Press, New York, 1994), pp. 3–14
- [987] P.L. Roe, *Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes*, vol. 135 (1997), pp. 249–258. With an introduction by M.J. Baines, Commemoration of the 30th anniversary of *J. Comput. Phys.*
- [988] P.L. Roe, Multidimensional upwinding, in *Handbook of Numerical Methods for Hyperbolic Problems*. Handb. Numer. Anal., vol. 18 (Elsevier/North-Holland, Amsterdam, 2017), pp. 53–80
- [989] P.L. Roe, D.S. Balsara, Notes on the eigensystem of magnetohydrodynamics. *SIAM J. Appl. Math.* **56**, 57–67 (1996)
- [990] P.L. Roe, J. Pike, Efficient construction and utilisation of approximate Riemann solutions, in *Computing Methods in Applied Sciences and Engineering VI (Proceedings of the Sixth International Symposium on Computing Methods in Applied Sciences and Engineering, 1995)*

- France (1983)), ed. by R. Glowinski, J.-L. Lions (Elsevier Science Publishers B.V., North-Holland, Amsterdam, 1984), pp. 499–518
- [991] P.L. Roe, D. Sidilkover, Optimum positive linear schemes for advection in two and three dimensions. SIAM J. Numer. Anal. **29**, 1542–1568 (1992)
- [992] P. Rostand, B. Stoufflet, TVD schemes to compute compressible viscous flows on unstructured meshes, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications (Aachen, 1988)*. Notes Numer. Fluid Mech., vol. 24 (Friedr. Vieweg, Braunschweig, 1989), pp. 510–520
- [993] B. Rubino, Convergence of approximate solutions of the Cauchy problem for a  $2 \times 2$  nonstrictly hyperbolic system of conservation laws, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects (Taormina, 1992)*. Notes Numer. Fluid Mech., vol. 43 (Friedr. Vieweg, Braunschweig, 1993), pp. 487–494
- [994] B. Rubino, On the vanishing viscosity approximation to the Cauchy problem for a  $2 \times 2$  system of conservation laws. Ann. Inst. H. Poincaré Anal. Non Linéaire **10**, 627–656 (1993)
- [995] D.H. Rudy, J.C. Strikwerda, A nonreflecting outflow boundary condition for subsonic Navier-Stokes calculations. J. Comput. Phys. **36**, 55–70 (1980)
- [996] C.L. Rumsey, B. van Leer, P.L. Roe, A multidimensional flux function with applications to the Euler and Navier-Stokes equations. J. Comput. Phys. **105**, 306–323 (1993)
- [997] M. Sablé-Tougeron, Méthode de Glimm et problème mixte. Ann. Inst. H. Poincaré Anal. Non Linéaire **10**, 423–443 (1993)
- [998] M. Sablé-Tougeron, Les  $N$ -ondes de Lax pour le problème mixte. Comm. Partial Differ. Equ. **19**, 1449–1479 (1994)
- [999] L. Sainsaulieu, Travelling wave solutions of convection-diffusion systems and nonconservative hyperbolic systems, in *Nonlinear Hyperbolic Problems: Theoretical, Applied, and Computational Aspects (Taormina, 1992)*. Notes Numer. Fluid Mech., vol. 43 (Friedr. Vieweg, Braunschweig, 1993), pp. 512–519
- [1000] L. Sainsaulieu, Finite volume approximate of two-phase fluid flows based on an approximate Roe-type Riemann solver. J. Comput. Phys. **121**, 1–28 (1995)
- [1001] L. Sainsaulieu, Traveling-wave solutions of convection-diffusion systems in nonconservation form. SIAM J. Math. Anal. **27**, 1286–1310 (1996)
- [1002] J. Sainte-Marie, Vertically averaged models for the free surface non-hydrostatic Euler system: derivation and kinetic interpretation. Math. Models Methods Appl. Sci. **21**, 459–490 (2011)
- [1003] K. Salari, S. Steinberg, Flux-Corrected transport in a moving grid. J. Comput. Phys. **111**, 24–32 (1994)

- [1004] J. Saltzman, An unsplit 3D upwind method for hyperbolic conservation laws. *J. Comput. Phys.* **115**, 153–168 (1994)
- [1005] R. Sanders, On convergence of monotone finite difference schemes with variable spatial differencing. *Math. Comput.* **40**, 91–106 (1983)
- [1006] R. Sanders, The moving grid method for nonlinear hyperbolic conservation laws. *SIAM J. Numer. Anal.* **22**, 713–728 (1985)
- [1007] R. Sanders, A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation laws. *Math. Comput.* **51**, 535–558 (1988)
- [1008] R. Sanders, C.-P. Li, A variation nonexpansive central differencing scheme for nonlinear hyperbolic conservation laws, in *Computing Methods in Applied Sciences and Engineering (Proceedings of the Tenth International Conference on Computing Methods in Applied Sciences and Engineering, France)*, ed. by R. Glowinski (Nova Science Publishers, Inc, New York, 1992), pp. 511–526
- [1009] R. Sanders, A. Weiser, High resolution staggered mesh approach for nonlinear hyperbolic systems of conservation laws. *J. Comput. Phys.* **101**, 314–329 (1992)
- [1010] R. Saurel, R. Abgrall, A multiphase Godunov method for compressible multifluid and multiphase flows. *J. Comput. Phys.* **150**, 425–467 (1999)
- [1011] R. Saurel, R. Abgrall, A simple method for compressible multifluid flows. *SIAM J. Sci. Comput.* **21**, 1115–1145 (1999)
- [1012] R. Saurel, M. Larini, J.-C. Loraud, Exact and approximate Riemann solvers for real gases. *J. Comput. Phys.* **112**, 126–137 (1994)
- [1013] D.G. Schaeffer, M. Shearer, The classification of  $2 \times 2$  systems of non-strictly hyperbolic conservation laws, with application to oil recovery. *Comm. Pure Appl. Math.* **40**, 141–178 (1987)
- [1014] S. Schecter, Undercompressive shock waves and the Dafermos regularization. *Nonlinearity* **15**, 1361–1377 (2002)
- [1015] S. Schecter, D. Marchesin, B.J. Plohr, Structurally stable Riemann solutions. *J. Differ. Equ.* **126**, 303–354 (1996)
- [1016] S. Schecter, B.J. Plohr, D. Marchesin, Computation of Riemann solutions using the Dafermos regularization and continuation. *Discrete Contin. Dyn. Syst.* **10**, 965–986 (2004)
- [1017] V. Schneider, U. Katscher, D.H. Rischke, B. Waldhauser, J.A. Maruhn, C.-D. Munz, New algorithms for ultra-relativistic numerical hydrodynamics. *J. Comput. Phys.* **105**, 92–107 (1993)
- [1018] S. Schochet, The compressible Euler equations in a bounded domain: existence of solutions and the incompressible limit. *Commun. Math. Phys.* **104**, 49–75 (1986)
- [1019] S. Schochet, Sufficient conditions for local existence via Glimm's scheme for large BV data. *J. Differ. Equ.* **89**, 317–354 (1991)
- [1020] S. Schochet, The mathematical theory of low Mach number flows. *M2AN Math. Model. Numer. Anal.* **39**, 441–458 (2005)

- [1021] S. Schochet, E. Tadmor, The regularized Chapman-Enskog expansion for scalar conservation laws. *Arch. Rational Mech. Anal.* **119**, 95–107 (1992)
- [1022] M.E. Schonbek, Second-order conservative schemes and the entropy condition. *Math. Comput.* **44**, 31–38 (1985)
- [1023] H.J. Schroll, A. Tveito, R. Winther, A system of conservation laws with a relaxation term, in *Hyperbolic Problems: Theory, Numerics, Applications (Stony Brook, NY, 1994)* (World Sci. Publ., River Edge, 1996), pp. 431–439.
- [1024] H.J. Schroll, A. Tveito, R. Winther, An  $L^1$ -error bound for a semi-implicit difference scheme applied to a stiff system of conservation laws. *SIAM J. Numer. Anal.* **34**, 1152–1166 (1997)
- [1025] H.J. Schroll, R. Winther, Finite-difference schemes for scalar conservation laws with source terms. *IMA J. Numer. Anal.* **16**, 201–215 (1996)
- [1026] C.W. Schulz-Rinne, Classification of the Riemann problem for two-dimensional gas dynamics. *SIAM J. Math. Anal.* **24**, 76–88 (1993)
- [1027] C.W. Schulz-Rinne, J.P. Collins, H.M. Glaz, Numerical solution of the Riemann problem for two-dimensional gas dynamics. *SIAM J. Sci. Comput.* **14**, 1394–1414 (1993)
- [1028] N. Seguin, J. Vovelle, Analysis and approximation of a scalar conservation law with a flux function with discontinuous coefficients. *Math. Models Methods Appl. Sci.* **13**, 221–257 (2003)
- [1029] V. Selmin, L. Quartapelle, A unified approach to build artificial dissipation operators for finite element and finite volume discretisations, in *Finite Elements in Fluids, Part I, II (Barcelona, 1993)* (Centro Internac. Métodos Numér. Ing., Barcelona, 1993), pp. 1329–1341. Translated in *Comput. Math. Model.* **5**(4), 308–310 (1994)
- [1030] S. Serna, A characteristic-based nonconvex entropy-fix upwind scheme for the ideal magnetohydrodynamic equations. *J. Comput. Phys.* **228**, 4232–4247 (2009)
- [1031] D. Serre, La compacité par compensation pour les systèmes hyperboliques non linéaires de deux équations à une dimension d'espace. *J. Math. Pures Appl.* (9) **65**, 423–468 (1986)
- [1032] D. Serre, Domaines invariants pour les systèmes hyperboliques de lois de conservation. *J. Differ. Equ.* **69**, 46–62 (1987)
- [1033] D. Serre, La stabilité d'une méthode de pas fractionnaires pour la dynamique des gaz, Report UMPA 86, ENS Lyon, France (1992)
- [1034] D. Serre, Solutions à variations bornées pour certains systèmes hyperboliques de lois de conservation. *J. Differ. Equ.* **68**, 137–168 (1987)
- [1035] D. Serre, Problèmes de Riemann singuliers. *Appl. Anal.* **35**, 175–185 (1990)
- [1036] D. Serre, Richness and the classification of quasilinear hyperbolic systems, in *Multidimensional Hyperbolic Problems and Computations (Minneapolis, MN, 1989)*. IMA Vol. Math. Appl., vol. 29 (Springer, New York, 1991), pp. 315–333

- [1037] D. Serre, Remarks about the discrete profiles of shock waves. *Mat. Contemp.* **11**, 153–170 (1996). Fourth Workshop on Partial Differential Equations, Part II (Rio de Janeiro, 1995)
- [1038] D. Serre, *Systèmes de lois de conservation. I, Fondations. [Foundations]*. (Diderot Editeur, Paris, 1996). Hyperbolicité, entropies, ondes de choc. [Hyperbolicity, entropies, shock waves]
- [1039] D. Serre, *Systems of Conservation Laws. 1* (Cambridge University Press, Cambridge, 1999). Hyperbolicity, entropies, shock waves, Translated from the 1996 French original by I. N. Sneddon
- [1040] D. Serre, Relaxations semi-linéaire et cinétique des systèmes de lois de conservation. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **17**, 169–192 (2000)
- [1041] D. Serre, *Systems of Conservation Laws. 2* (Cambridge University Press, Cambridge, 2000). Geometric structures, oscillations, and initial-boundary value problems, Translated from the 1996 French original by I.N. Sneddon
- [1042] D. Serre, Discrete shock profiles: existence and stability, in *Hyperbolic Systems of Balance Laws*. Lecture Notes in Math., vol. 1911 (Springer, Berlin, 2007), pp. 79–158
- [1043] D. Serre, Systems of conservation laws. Theory, numerical approximation and discrete shock profiles, in *Nonlinear Conservation Laws, Fluid Systems and Related Topics*. Ser. Contemp. Appl. Math. CAM, vol. 13 (World Sci. Publishing, Singapore, 2009), pp. 72–125
- [1044] D. Serre, A.F. Vasseur, About the relative entropy method for hyperbolic systems of conservation laws, in *A Panorama of Mathematics: Pure and Applied*. Contemp. Math., vol. 658 (Amer. Math. Soc., Providence, 2016), pp. 237–248
- [1045] D. Serre, L. Xiao, Asymptotic behavior of large weak entropy solutions of the damped  $P$ -system. *J. Partial Differ. Equ.* **10**, 355–368 (1997)
- [1046] V.D. Sharma, *Quasilinear Hyperbolic Systems, Compressible Flows, and Waves*. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, vol. 142 (CRC Press, Boca Raton, 2010)
- [1047] M. Shearer, Nonuniqueness of admissible solutions of Riemann initial value problems for a system of conservation laws of mixed type. *Arch. Rational Mech. Anal.* **93**, 45–59 (1986)
- [1048] M. Shearer, Loss of strict hyperbolicity of the Buckley-Leverett equations for three-phase flow in a porous medium, in *Numerical Simulation in Oil Recovery (Minneapolis, Minn., 1986)*. IMA Vol. Math. Appl., vol. 11 (Springer, New York, 1988), pp. 263–283
- [1049] M. Shearer, D.G. Schaeffer, D. Marchesin, P.L. Paes-Leme, Solution of the Riemann problem for a prototype  $2 \times 2$  system of nonstrictly hyperbolic conservation laws. *Arch. Rational Mech. Anal.* **97**, 299–320 (1987)

- [1050] M. Shearer, S. Schecter, Undercompressive shocks in systems of conservation laws, in *Nonlinear Evolution Equations That Change Type*. IMA Vol. Math. Appl., vol. 27 (Springer, New York, 1990), pp. 218–231
- [1051] W. Sheng, D.C. Tan, Weak deflagration solutions to the simplest combustion model. J. Differ. Equ. **107**, 207–230 (1994)
- [1052] W. Sheng, T. Zhang, The Riemann problem for the transportation equations in gas dynamics. Mem. Amer. Math. Soc. **137**, viii+77 (1999)
- [1053] Y. Shizuta, S. Kawashima, Systems of equations of hyperbolic-parabolic type with applications to the discrete Boltzmann equation. Hokkaido Math. J. **14**, 249–275 (1985)
- [1054] C.-W. Shu, TVB uniformly high-order schemes for conservation laws. Math. Comput. **49**, 105–121 (1987)
- [1055] C.-W. Shu, Total-variation-diminishing time discretizations. SIAM J. Sci. Stat. Comput. **9**, 1073–1084 (1988)
- [1056] C.-W. Shu, A numerical method for systems of conservation laws of mixed type admitting hyperbolic flux splitting. J. Comput. Phys. **100**, 424–429 (1992)
- [1057] C.-W. Shu, Discontinuous Galerkin methods: general approach and stability, in *Numerical Solutions of Partial Differential Equations*. Adv. Courses Math. CRM Barcelona (Birkhäuser, Basel, 2009), pp. 149–201
- [1058] C.-W. Shu, S. Osher, Efficient implementation of essentially nonoscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)
- [1059] C.-W. Shu, S. Osher, Efficient implementation of essentially nonoscillatory shock-capturing schemes. II, J. Comput. Phys. **83**, 32–78 (1989)
- [1060] C.-W. Shu, T. Zang, G. Erlebacher, D. Whitaker, S. Osher, High-order ENO schemes applied to two- and three-dimensional compressible flow. Appl. Numer. Math. **9**, 45–71 (1992)
- [1061] D. Sidilkover, A genuinely multidimensional upwind scheme for the compressible Euler equations, in *Hyperbolic problems: theory, numerics, applications (Stony Brook, NY, 1994)* (World Science Publication, River Edge, 1996), pp. 447–455
- [1062] C. Simeoni, Remarks on the consistency of upwind source at interface schemes on nonuniform grids. J. Sci. Comput. **48**, 333–338 (2011)
- [1063] M. Slemrod, A.E. Tzavaras, Shock profiles and self-similar fluid dynamic limits, in *Proceedings of the Second International Workshop on Nonlinear Kinetic Theories and Mathematical Aspects of Hyperbolic Systems (Sanremo, 1994)*, vol. 25 (1996), pp. 531–541
- [1064] R.G. Smith, The Riemann problem in gas dynamics. Trans. Am. Math. Soc. **249**, 1–50 (1980)
- [1065] J.A. Smoller, ed. *Nonlinear partial differential equations*. Contemporary Mathematics, vol. 17 (American Mathematical Society, Providence, 1983). Papers from the Conference held at the University of New Hampshire, Durham, N.H., June 20–26 (1982)

- [1066] J.A. Smoller, *Shock Waves and Reaction-Diffusion Equations*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 258, 2nd edn. (Springer, New York, 1994)
- [1067] J.A. Smoller, J.B. Temple, Z.P. Xin, Instability of rarefaction shocks in systems of conservation laws. *Arch. Rational Mech. Anal.* **112**, 63–81 (1990)
- [1068] G.A. Sod, A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *J. Comput. Phys.* **27**, 1–31 (1978)
- [1069] G.A. Sod, *Numerical Methods in Fluid Dynamics* (Cambridge University, Cambridge, 1985). Initial and initial-boundary value problems
- [1070] Y. Song, T. Tang, Dispersion and group velocity in numerical schemes for three-dimensional hydrodynamic equations. *J. Comput. Phys.* **105**, 72–82 (1993)
- [1071] S. Spekreijse, Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws. *Math. Comput.* **49**, 135–155 (1987)
- [1072] J.L. Steger, R.F. Warming, Flux vector splitting of the inviscid gas-dynamic equations with application to finite-difference methods. *J. Comput. Phys.* **40**, 263–293 (1981)
- [1073] H.B. Stewart, B. Wendroff, Two-phase flow: models and methods. *J. Comput. Phys.* **56**, 363–409 (1984)
- [1074] Y. Stiriba, R. Donat, A numerical study of postshock oscillations in slowly moving shock waves. *Comput. Math. Appl.* **46**, 719–739 (2003)
- [1075] B. Stoufflet, Implicit finite element methods for the Euler equations, in *Numerical methods for the Euler equations of fluid dynamics (Rocquencourt, 1983)* (SIAM, Philadelphia, 1985), pp. 409–434
- [1076] I. Suliciu, On the thermodynamics of rate-type fluids and phase transitions. I. Rate-type fluids. *Internat. J. Eng. Sci.* **36**, 921–947 (1998)
- [1077] M.T. Sun, S.T. Wu, M. Dryer, On the time-dependent numerical boundary conditions of magnetohydrodynamic flows. *J. Comput. Phys.* **116**, 330–342 (1995)
- [1078] R.C. Swanson, E. Turkel, On central-difference and upwind schemes. *J. Comput. Phys.* **101**, 292–306 (1992)
- [1079] P.K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.* **21**, 995–1011 (1984)
- [1080] P.K. Sweby, “TVD” schemes for inhomogeneous conservation laws, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications (Aachen, 1988)*. Notes Numerical Fluid Mechanical, Friedr. (Vieweg, Braunschweig, 1989), pp. 599–607
- [1081] P.K. Sweby, M.J. Baines, On convergence of Roe’s scheme for the general nonlinear scalar wave equation. *J. Comput. Phys.* **56**, 135–148 (1984)

- [1082] A. Szepessy, An existence result for scalar conservation laws using measure valued solutions. *Commun. Partial Differ. Equ.* **14**, 1329–1350 (1989)
- [1083] A. Szepessy, Measure-valued solutions of scalar conservation laws with boundary conditions. *Arch. Rational Mech. Anal.* **107**, 181–193 (1989)
- [1084] A. Szepessy, Convergence of a streamline diffusion finite element method for scalar conservation laws with boundary conditions. *RAIRO Modél. Math. Anal. Numér.* **25**, 749–782 (1991)
- [1085] A. Szepessy, Lectures on stability of nonlinear waves in viscous media and numerics, in *Analysis of Systems of Conservation Laws (Aachen, 1997)*. Chapman & Hall/CRC Monographs Surveys Pure Application Mathematical, vol. 99 (Chapman & Hall/CRC, Boca Raton, 1999), pp. 199–261
- [1086] E. Tadmor, The equivalence of  $L_2$ -stability, the resolvent condition, and strict  $H$ -stability. *Linear Algebra Appl.* **41**, 151–159 (1981)
- [1087] E. Tadmor, The unconditional instability of inflow-dependent boundary conditions in difference approximations to hyperbolic systems. *Math. Comput.* **41**, 309–319 (1983)
- [1088] E. Tadmor, The large-time behavior of the scalar, genuinely nonlinear Lax-Friedrichs scheme. *Math. Comput.* **43**, 353–368 (1984)
- [1089] E. Tadmor, Numerical viscosity and the entropy condition for conservative difference schemes. *Math. Comput.* **43**, 369–381 (1984)
- [1090] E. Tadmor, A minimum entropy principle in the gas dynamics equations. *Appl. Numer. Math.* **2**, 211–219 (1986)
- [1091] E. Tadmor, Entropy functions for symmetry systems of conservation laws. *J. Math. Anal. Appl.* **122**, 355–359 (1987)
- [1092] E. Tadmor, The numerical viscosity of entropy stable schemes for systems of conservation laws. I, *Math. Comput.* **49**, 91–103 (1987)
- [1093] E. Tadmor, Entropy stable schemes, in *Handbook of Numerical Methods for Hyperbolic Problems*. Handbook Numerical Analysis, vol. 17 (Elsevier/North-Holland, Amsterdam, 2016), pp. 467–493
- [1094] E. Tadmor, T. Tang, Pointwise error estimates for relaxation approximations to conservation laws. *SIAM J. Math. Anal.* **32**, 870–886 (2000)
- [1095] E. Tadmor, T. Tassa, On the piecewise smoothness of entropy solutions to scalar conservation laws. *Commun. Partial Differ. Equ.* **18**, 1631–1652 (1993)
- [1096] Y. Tamura, K. Fujii, A multi-dimensional upwind scheme for the euler equations on structured grid. *Comput. Fluids* **22**, 125–138 (1993)
- [1097] D.C. Tan, T. Zhang, Riemann problem for the self-similar ZND-model in gas dynamical combustion. *J. Differ. Equ.* **95**, 331–369 (1992)
- [1098] D.C. Tan, T. Zhang, Two-dimensional Riemann problem for a hyperbolic system of nonlinear conservation laws. I. Four- $J$  cases. *J. Differ. Equ.* **111**, 203–254 (1994)

- [1099] D.C. Tan, T. Zhang, Y.X. Zheng, Delta-shock waves as limits of vanishing viscosity for hyperbolic systems of conservation laws. *J. Differ. Equ.* **112**, 1–32 (1994)
- [1100] S. Tan, C.-W. Shu, Inverse Lax-Wendroff procedure for numerical boundary conditions of conservation laws. *J. Comput. Phys.* **229**, 8144–8166 (2010)
- [1101] T. Tanaka, Finite volume TVD scheme on an unstructured grid system for three-dimensional M.H.D. simulation of inhomogeneous systems including strong background potential fields. *J. Comput. Phys.* **111**, 381–389 (1994)
- [1102] T. Tang, Convergence analysis for operator-splitting methods applied to conservation laws with stiff source terms. *SIAM J. Numer. Anal.* **35**, 1939–1968 (1998)
- [1103] T. Tang, Z.H. Teng, Time-splitting methods for nonhomogeneous conservation laws, in *Mathematics of Computation 1943–1993: A Half-century of Computational Mathematics (Vancouver, BC, 1993)*. Proceedings of the Symposium Application Mathematical, vol. 48 (American Mathematical Society, Providence, 1994), pp. 389–393
- [1104] T. Tang, Z.H. Teng, Error bounds for fractional step methods for conservation laws with source terms. *SIAM J. Numer. Anal.* **32**, 110–127 (1995)
- [1105] T. Taniuti, K. Nishihara, *Nonlinear waves*, Monographs and Studies in Mathematics, Pitman (Advanced Publishing Program), Boston, MA, 1983, vol. 15. Translated from the Japanese by Taniuti and Alan Jeffrey
- [1106] L. Tartar, Compensated compactness and applications to partial differential equations, in *Nonlinear Analysis and Mechanics: Heriot-Watt Symposium*, vol. IV. Research Notes in Mathematical, vol. 39 (Pitman, Boston, Mass.-London, 1979), pp. 136–212
- [1107] L. Tartar, The compensated compactness method applied to systems of conservation laws, in *Systems of Nonlinear Partial Differential Equations (Oxford, 1982)*. NATO Advantage of Science Institution Series C Mathematical of Physics Science, vol. 111 (Reidel, Dordrecht, 1983), pp. 263–285
- [1108] L. Tartar, *Une Introduction à la théorie mathématique des systèmes hyperboliques de lois de conservation*, Report 682 (Istituto di Analisi del CNR 27100 Pavia, Italy, 1989)
- [1109] L. Tartar, From hyperbolic systems to kinetic theory, *Lecture Notes of the Unione Matematica Italiana*, vol. 6 (Springer, Berlin, 2008). A personalized quest
- [1110] M.E. Taylor, *Partial differential equations III. Nonlinear equations*, Applied Mathematical Sciences, vol. 117, 2nd edn. (Springer, New York, 2011)

- [1111] B. Temple, Global solution of the Cauchy problem for a class of  $2 \times 2$  nonstrictly hyperbolic conservation laws. *Adv. Appl. Math.* **3**, 335–375 (1982)
- [1112] B. Temple, Degenerate systems of conservation laws, in *Nonstrictly hyperbolic Conservation Laws (Anaheim, California, 1985)*. Contemporary Mathematical, vol. 60 (American Mathematical Society, Providence, 1987), pp. 125–133
- [1113] B. Temple, R. Young, Solutions to the Euler equations with large data, in *Hyperbolic Problems: Theory, Numerics, Applications (Stony Brook, NY, 1994)* (World Science Publication, River Edge, 1996), pp. 258–267
- [1114] Z.H. Teng, On the accuracy of fractional step methods for conservation laws in two dimensions. *SIAM J. Numer. Anal.* **31**, 43–63 (1994)
- [1115] Z.H. Teng, First-order  $L^1$ -convergence for relaxation approximations to conservation laws. *Commun. Pure Appl. Math.* **51**, 857–895 (1998)
- [1116] Z.H. Teng, A.J. Chorin, T.P. Liu, Riemann problems for reacting gas, with applications to transition. *SIAM J. Appl. Math.* **42**, 964–981 (1982)
- [1117] M.D. Thanh, Numerical treatment in resonant regime for shallow water equations with discontinuous topography. *Commun. Nonlinear Sci. Numer. Simul.* **18**, 417–433 (2013)
- [1118] M.D. Thanh, D.H. Cuong, Existence of solutions to the Riemann problem for a model of two-phase flows. *Electron. J. Differ. Equ.* **32**, 18 (2015)
- [1119] M.D. Thanh, D. Kröner, Numerical treatment of nonconservative terms in resonant regime for fluid flows in a nozzle with variable cross-section. *Comput. Fluids* **66**, 130–139 (2012)
- [1120] J.W. Thomas, *Numerical partial differential equations*, Texts in Applied Mathematics, vol. 33 (Springer, New York, 1999). Conservation laws and elliptic equations
- [1121] K.W. Thompson, Time-dependent boundary conditions for hyperbolic systems. *J. Comput. Phys.* **68**, 1–24 (1987)
- [1122] J. Thouvenin, *Les mécanismes élémentaires de la détonique* (unpublished)
- [1123] V.A. Titarev, E.F. Toro, ADER: arbitrary high order Godunov approach, in *Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala)*, vol. 17 (2002), pp. 609–618
- [1124] E.F. Toro, A weighted average flux method for hyperbolic conservation laws. *Proc. Roy. Soc. London Ser. A* **423**, 401–418 (1989)
- [1125] E.F. Toro, A linearized Riemann solver for the time-dependent Euler equations of gas dynamics. *Proc. Roy. Soc. London Ser. A* **434**, 683–693 (1991)

- [1126] E.F. Toro, Riemann problems and the WAF method for solving the two-dimensional shallow water equations. *Philos. Trans. R. Soc. London Ser. A* **338**, 43–68 (1992)
- [1127] E.F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 3rd edn. (Springer, Berlin, 2009). A practical introduction
- [1128] M. Torrilhon, D.S. Balsara, High order WENO schemes: investigations on non-uniform convergence for MHD Riemann problems. *J. Comput. Phys.* **201**, 586–600 (2004)
- [1129] I. Toumi, A weak formulation of Roe’s approximate Riemann solver. *J. Comput. Phys.* **102**, 360–373 (1992)
- [1130] I. Toumi, A. Kumbaro, An approximate linearized Riemann solver for a two-fluid model. *J. Comput. Phys.* **124**, 286–300 (1996)
- [1131] J.A. Trangenstein, *Numerical Solution of Hyperbolic Partial Differential Equations* (Cambridge University, Cambridge, 2009). With 1 CD-ROM (Intel compatible PC or Mac)
- [1132] L.N. Trefethen, Group velocity in finite difference schemes. *SIAM Rev.* **24**, 113–136 (1982)
- [1133] L.N. Trefethen Instability of difference models for hyperbolic initial-boundary value problems. *Commun. Pure Appl. Math.* **37**, 329–367 (1984)
- [1134] L.N. Trefethen, Stability of hyperbolic finite-difference models with one or two boundaries, in *Large-scale Computations in Fluid Mechanics, Part 2 (La Jolla, California, 1983)*. Lectures in Application of Mathematical (American Mathematical Society, Providence, 1985), pp. 311–326
- [1135] J. Tryoen, O. Le Maître, M. Ndjinga, A. Ern, Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems. *J. Comput. Phys.* **229**, 6485–6511 (2010)
- [1136] J. Tryoen, O. Le Maître, M. Ndjinga, A. Ern, Roe solver with entropy corrector for uncertain hyperbolic systems. *J. Comput. Appl. Math.* **235**, 491–506 (2010)
- [1137] E. Turkel, Acceleration to a steady state for the Euler equations, in *Numerical Methods for the Euler Equations of Fluid Dynamics (Rocquencourt, 1983)* (SIAM, Philadelphia, 1985), pp. 281–311
- [1138] E. Turkel, Accuracy of schemes with nonuniform meshes for compressible fluid flows. *Appl. Numer. Math.* **2**, 529–550 (1986)
- [1139] E. Turkel, Review of preconditioning methods for fluid dynamics. *Appl. Numer. Math.* **12**, 257–284 (1993). Special issue to honor Professor Saul Abarbanel on his sixtieth birthday (Neveh, 1992)
- [1140] A. Tveito, R. Winther, Existence, uniqueness, and continuous dependence for a system of hyperbolic conservation laws modeling polymer flooding. *SIAM J. Math. Anal.* **22**, 905–933 (1991)
- [1141] A. Tveito, R. Winther, An error estimate for a finite difference scheme approximating a hyperbolic system of conservation laws. *SIAM J. Numer. Anal.* **30**, 401–424 (1993)

- [1142] A. Tveito, R. Winther, On the rate of convergence to equilibrium for a system of conservation laws with a relaxation term. SIAM J. Math. Anal. **28**, 136–161 (1997)
- [1143] A.E. Tzavaras, Wave interactions and variation estimates for self-similar zero-viscosity limits in systems of conservation laws. Arch. Rational Mech. Anal. **135**, 1–60 (1996)
- [1144] A.E. Tzavaras, Materials with internal variables and relaxation to conservation laws. Arch. Ration. Mech. Anal. **146**, 129–155 (1999)
- [1145] A.E. Tzavaras, Viscosity and relaxation approximation for hyperbolic systems of conservation laws, in *An introduction to recent developments in theory and numerics for conservation laws (Freiburg/Littenweiler, 1997)*. Lectures of Notes Computer Science Engineers, vol. 5 (Springer, Berlin, 1999), pp. 73–122
- [1146] A.E. Tzavaras, Relative entropy in hyperbolic relaxation, Commun. Math. Sci. **3**, 119–132 (2005)
- [1147] C.J. van Duijn, L.A. Peletier, I.S. Pop, A new class of entropy solutions of the Buckley-Leverett equation. SIAM J. Math. Anal. **39**, 507–536 (2007) (electronic)
- [1148] B. van Leer, Towards the ultimate conservative difference scheme: I. The quest of monotonicity, in *Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics (1972)*. Lecture Notes in Physics, vol. 18 (Springer, Berlin, 1973), pp. 163–168
- [1149] B. van Leer, II. Monotonicity and conservation combined in a second-order scheme. J. Comput. Phys. **14**, 361–370 (1974)
- [1150] B. van Leer, III. Upstream-centered finite-difference schemes for ideal compressible flow. J. Comput. Phys. **23**, 263–275 (1977)
- [1151] B. van Leer, IV. A new approach to numerical convection. J. Comput. Phys. **23**, 276–279 (1977)
- [1152] B. van Leer, Flux-vector splitting for the Euler equations, in *Lecture Notes in Physics*, vol. 170 (Springer, Berlin, 1982), pp. 507–512
- [1153] B. van Leer, Multidimensional explicit difference schemes for hyperbolic conservation laws, in *Computing Methods in Applied Sciences and Engineering, VI (Versailles, 1983)* (North-Holland, Amsterdam, 1984), pp. 493–497
- [1154] B. van Leer, On the relation between the upwind-differencing schemes of Godunov, Engquist-Osher and Roe. SIAM J. Sci. Statist. Comput. **5**, 1–20 (1984)
- [1155] B. van Leer, Upwind-difference methods for aerodynamic problems governed by the Euler equations, in *Large-scale Computations in Fluid Mechanics, Part 2 (La Jolla, California, 1983)*. Lectures in Application Mathematical, , vol. 22 (American Mathematical Society, Providence, 1985), pp. 327–336
- [1156] B. van Leer, On numerical dispersion by upwind differencing. Appl. Numer. Math. **2**, 379–384 (1986)

- [1157] B. van Leer, Progress in multi-dimensional upwind differencing, in *ICASE Report 92-43, ICASE NASA* (Langley Research Center, Hampton, 1992)
- [1158] B. van Leer, Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method [J. Comput. Phys. **32** (1979), no. 1, 101–136] **135**, 227–248 (1997). With an introduction by Ch. Hirsch, Commemoration of the 30th anniversary {of J. Comput. Phys.}
- [1159] B. van Leer, An introduction to the article “Reminiscences about difference schemes” [J. Comput. Phys. **153**(1), 6–25 (1999)] by S.K. Godunov. J. Comput. Phys. **153**, 1–5 (1999)
- [1160] B. van Leer, W.A. Mulder, Relaxation methods for hyperbolic conservation laws, in *Numerical Methods for the Euler equations of Fluid Dynamics (Rocquencourt, 1983)* (SIAM, Philadelphia, 1985), pp. 312–333
- [1161] T.C. Vanajakshi, K.W. Thompson, D. Black, Boundary value problems in magnetohydrodynamics (and fluid dynamics). i. radiation boundary condition. J. Comput. Phys. **84**, 343–359 (1989)
- [1162] A.F. Vasseur, Well-posedness of scalar conservation laws with singular sources. Methods Appl. Anal. **9**, 291–312 (2002)
- [1163] A.F. Vasseur, Recent results on hydrodynamic limits, in *Handbook of Differential Equations: Evolutionary Equations*. Handbook Differential Equation, vol. IV (Elsevier/North-Holland, Amsterdam, 2008), pp. 323–376
- [1164] A.F. Vasseur, A rigorous derivation of the coupling of a kinetic equation and Burgers' equation. Arch. Ration. Mech. Anal. **206**, 1–30 (2012)
- [1165] S. Vater, R. Klein, Stability of a Cartesian grid projection method for zero Froude number shallow water flows. Numer. Math. **113**, 123–161 (2009)
- [1166] M.E. Vázquez-Cendón, Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry. J. Comput. Phys. **148**, 497–526 (1999)
- [1167] V. Venkatakrishnan, Convergence to steady state solutions of the Euler equations on unstructured grids with limiters. J. Comput. Phys. **118**, 120–130 (1995)
- [1168] J.-P. Vieuillet, L. Cambier, A subdomain approach for the computation of compressible inviscid flows, in *Numerical Methods for the Euler Equations of Fluid Dynamics (Rocquencourt, 1983)* (SIAM, Philadelphia, 1985), pp. 470–489
- [1169] G. Vijayasundaram, Transonic flow simulations using an upstream centered scheme of Godunov in finite elements. J. Comput. Phys. **63**, 416–433 (1986)

- [1170] J.-P. Vila, Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws. I. Explicit monotone schemes. *RAIRO Modél. Math. Anal. Numér.* **28**, 267–295 (1994)
- [1171] P. Villedieu, P.A. Mazet, Schémas cinétiques pour les équations d'Euler hors équilibre thermochimique. *Rech. Aéronaut. spat.* **2**, 85–102 (1995)
- [1172] M. Vinokur, An analysis of finite-difference and finite-volume formulations of conservation laws. *J. Comput. Phys.* **81**, 1–52 (1989)
- [1173] M. Vinokur, J.-L. Montagné, Generalized flux-vector splitting and Roe average for an equilibrium real gas. *J. Comput. Phys.* **89**, 276–300 (1990)
- [1174] D.H. Wagner, The Riemann problem in two space dimensions for a single conservation law. *SIAM J. Math. Anal.* **14**, 534–559 (1983)
- [1175] D.H. Wagner, Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions. *J. Differ. Equ.* **68**, 118–136 (1987)
- [1176] D.H. Wagner, Conservation laws, coordinate transformations, and differential forms, in *Hyperbolic Problems: Theory, Numerics, Applications (Stony Brook, NY, 1994)* (World Science Publication, River Edge, 1996), pp. 471–477
- [1177] J.H. Wang, G. Warnecke, On entropy consistency of large time step schemes. I. The Godunov and Glimm schemes. *SIAM J. Numer. Anal.* **30**, 1229–1251 (1993)
- [1178] J.H. Wang, G. Warnecke, On entropy consistency of large time step schemes. II. Approximate Riemann solvers. *SIAM J. Numer. Anal.* **30**, 1252–1267 (1993)
- [1179] R.F. Warming, R.M. Beam, On the construction and application of implicit factored schemes for conservation laws, in *Computational Fluid Dynamics (Proceedings of SIAM-AMS Symposium Application of Mathematical, New York, 1977)* (1978), pp. 85–129. SIAM-AMS Proceeding, vol. XI
- [1180] R.F. Warming, R.M. Beam, B.J. Hyett, Diagonalization and simultaneous symmetrization of the gas-dynamic matrices. *Math. Comput.* **29**, 1037–1045 (1975)
- [1181] R.F. Warming, B.J. Hyett, The modified equation approach to the stability and accuracy analysis of finite-difference methods. *J. Comput. Phys.* **14**, 159–179 (1974)
- [1182] G. Warnecke, Admissibility of solutions to the Riemann problem for systems of mixed type—transonic small disturbance theory, in *Nonlinear Evolution Equations that Change Type*. IMA Volume Mathematical of Application, vol. 27 (Springer, New York, 1990), pp. 258–284
- [1183] G. Warnecke, ed., *Analysis and Numerics for Conservation Laws* (Springer, Berlin, 2005)

- [1184] B. Wendroff, The Riemann problem for materials with nonconvex equations of state. I. Isentropic flow. *J. Math. Anal. Appl.* **38**, 454–466 (1972)
- [1185] B. Wendroff, The Riemann problem for materials with nonconvex equations of state. II. General flow. *J. Math. Anal. Appl.* **38**, 640–658 (1972)
- [1186] B. Wendroff, A.B. White, Jr., Some supraconvergent schemes for hyperbolic equations on irregular grids, in *Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications (Aachen, 1988)*. Notes Numerical Fluid Mechanical, vol. 24 (Friedrich, Braunschweig, 1989), pp. 671–677
- [1187] M. Westdickenberg, S. Noelle, A new convergence proof for finite volume schemes using the kinetic formulation of conservation laws. *SIAM J. Numer. Anal.* **37**, 742–757 (2000)
- [1188] G.B. Whitham, *Linear and nonlinear waves*, Pure and Applied Mathematics (New York) (Wiley, New York, 1999). Reprint of the 1974 original, A Wiley-Interscience Publication
- [1189] F. Williams, *Combustion Theory* (Menlo Park Publishing company, Benjamin/Cummings, 1985)
- [1190] P. Woodward, P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks. *J. Comput. Phys.* **54**, 115–173 (1984)
- [1191] G. Wu, Z. Tan, J. Huang, Global existence and large time behavior for the system of compressible adiabatic flow through porous media in  $\mathbb{R}^3$ . *J. Differ. Equ.* **255**, 865–880 (2013)
- [1192] Z.-N. Wu, *Conditions Aux Limites Pour un Problème Hyperbolique*, Internal Report (unpublished), Sinumef, ENSAM, 75013 Paris (France)
- [1193] Z.-N. Wu, Steady and unsteady shock waves on overlapping grids. *SIAM J. Sci. Comput.* **20**, 1851–1874 (1999)
- [1194] Y. Xing, C.-W. Shu, High order finite difference WENO schemes with the exact conservation property for the shallow water equations. *J. Comput. Phys.* **208**, 206–227 (2005)
- [1195] Y. Xing, C.-W. Shu, High order well-balanced WENO scheme for the gas dynamics equations under gravitational fields. *J. Sci. Comput.* **54**, 645–662 (2013)
- [1196] Y. Xing, C.-W. Shu, A survey of high order schemes for the shallow water equations. *J. Math. Study* **47**, 221–249 (2014)
- [1197] Y. Xing, C.-W. Shu, S. Noelle, On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations. *J. Sci. Comput.* **48**, 339–349 (2011)
- [1198] K. Xu, L. Martinelli, A. Jameson, Gas-kinetic finite volume methods, flux-vector splitting, and artificial diffusion. *J. Comput. Phys.* **120**, 48–65 (1995)
- [1199] K. Xu, K.H. Prendergast, Numerical Navier-Stokes solutions from gas kinetic theory. *J. Comput. Phys.* **114**, 9–17 (1994)

- [1200] H. Yang, An artificial compression method for ENO schemes: the slope modification method. *J. Comput. Phys.* **89**, 125–160 (1990)
- [1201] H.Q. Yang, A.J. Przekwas, A comparative study of advanced shock-capturing schemes applied to Burgers' equation. *J. Comput. Phys.* **102**, 139–159 (1992)
- [1202] J.Y. Yang, J.C. Huang, Rarefied flow computations using nonlinear model Boltzmann equations. *J. Comput. Phys.* **120**, 323–339 (1995)
- [1203] J.-Y. Yang, C.-A. Hsu, Numerical experiments with nonoscillatory schemes using Eulerian and new Lagrangian formulations. *Comput. Fluids* **22**(2), 163–177 (1993)
- [1204] H.C. Yee, *Numerical Approximation of Boundary Conditions with Applications to Inviscid Equations of Gas Dynamics* (1981). NASA Technical Memorandum 81285
- [1205] H.C. Yee, Construction of a class of symmetric TVD schemes, in *Oscillation Theory, Computation, and Methods of Compensated Compactness (Minneapolis, Minn., 1985)*. IMA Volume Mathematical Application, vol. 2 (Springer, New York, 1986), pp. 381–395
- [1206] H.C. Yee, Construction of explicit and implicit symmetric tvd schemes and their applications. *J. Comput. Phys.* **68**, 151–179 (1987)
- [1207] H.C. Yee, G.H. Klopfer, J.-L. Montagné, High-resolution shock-capturing schemes for inviscid and viscous hypersonic flows. *J. Comput. Phys.* **88**, 31–61 (1990)
- [1208] H.C. Yee, R.F. Warming, A. Harten, Application of TVD schemes for the Euler equations of gas dynamics, in *Large-scale Computations in Fluid Mechanics, Part 2 (La Jolla, California, 1983)*. Lectures in Application of the Mathematical (American Mathematical Society, Providence, 1985), pp. 357–377
- [1209] H.C. Yee, R.F. Warming, A. Harten, Implicit total variation diminishing (TVD) schemes for steady-state calculations. *J. Comput. Phys.* **57**, 327–360 (1985)
- [1210] W.-A. Yong, Singular perturbations of first-order hyperbolic systems with stiff source terms. *J. Differ. Equ.* **155**, 89–132 (1999)
- [1211] W.-A. Yong, Entropy and global existence for hyperbolic balance laws. *Arch. Ration. Mech. Anal.* **172**, 247–266 (2004)
- [1212] W.-A. Yong, A note on the zero Mach number limit of compressible Euler equations. *Proc. Am. Math. Soc.* **133**, 3079–3085 (2005) (electronic)
- [1213] R. Young, The  $p$ -system. I. The Riemann problem, in *The Legacy of the Inverse Scattering Transform in Applied Mathematics (South Hadley, MA, 2001)*. Contemporary Mathematical, vol. 301 (American Mathematical Society, Providence, 2002), pp. 219–234
- [1214] A.L. Zachary, P. Colella, A higher-order Godunov method for the equations of ideal magnetohydrodynamics. *J. Comput. Phys.* **99**, 341–347 (1992)

- [1215] A.L. Zachary, A. Malagoli, P. Colella, A higher-order Godunov method for multidimensional ideal magnetohydrodynamics. *SIAM J. Sci. Comput.* **15**, 263–284 (1994)
- [1216] S.T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* **31**, 335–362 (1979)
- [1217] A. Zein, M. Hantke, G. Warnecke, Modeling phase transition for compressible two-phase flows applied to metastable liquids. *J. Comput. Phys.* **229**, 2964–2998 (2010)
- [1218] Y. Zeng, Gas dynamics in thermal nonequilibrium and general hyperbolic systems with relaxation. *Arch. Ration. Mech. Anal.* **150**, 225–279 (1999)
- [1219] T. Zhang, Y.X. Zheng, Riemann problem for gasdynamic combustion. *J. Differ. Equ.* **77**, 203–230 (1989)
- [1220] T. Zhang, Y.X. Zheng, Two-dimensional Riemann problem for a single conservation law. *Trans. Am. Math. Soc.* **312**, 589–619 (1989)
- [1221] T. Zhang, Y.X. Zheng, Conjecture on the structure of solutions of the Riemann problem for two-dimensional gas dynamics systems. *SIAM J. Math. Anal.* **21**, 593–630 (1990)
- [1222] Y. Zheng, Systems of conservation laws. Two-dimensional Riemann problems, in *Progress in Nonlinear Differential Equations and their Applications*, vol. 38 (Birkhäuser Boston Inc., Boston, 2001). Two-dimensional Riemann problems
- [1223] X. Zhong, T.Y. Hou, P.G. LeFloch, Computational methods for propagating phase boundaries. *J. Comput. Phys.* **124**, 192–216 (1996)

# Index

## A

adiabatic exponent, 17, 19, 150, 203, 358, 741  
amplification factor, 235, 475, 476  
matrix, 231, 232, 474, 715, 718  
anisotropy, 474, 475  
Arrhenius law, 19  
asymptotic behavior, 401, 629, 708  
behavior (large friction), 721  
behavior (large time), 706  
behavior (low Mach), 550, 552  
expansion (GRP), 330, 333  
expansion (low Mach), 547, 557  
asymptotic preserving, 401, 547, 561, 629, 706, 729  
strongly, 720  
weakly, 714, 726  
average Roe state, 283  
state, 222, 228, 269  
averaging operator, 291, 296  
procedure, 20, 22, 24, 274, 639  
section-, 642  
step, 237, 250  
vertical-, 641

## B

Baer-Nunziato system, 26, 662  
BGK model, 363–365, 380  
kernel, 364, 389, 390, 392, 395, 629  
bicharacteristic, 457, 459, 596  
Boltzmann equation, 354

## boundary

artificial, 606, 610  
fluid, 623  
inflow/outflow, see inflow/outflow 597  
non characteristic, 590  
numerical flux, 621  
rigid wall, 621  
solid wall, 610  
transparent, 610  
value (admissible), 602, 603  
boundary condition, 108  
absorbing, 610, 614  
coupling, 734  
exact, 610  
interface, 735, 736  
mirror state, 621  
nonreflecting, 610, 617  
numerical, 618  
numerical (compatibility relations), 619  
numerical (extrapolation), 618, 620, 623  
numerical (translatory), 619  
residual, 602  
Silver-Müller, 613  
slip, 550, 610, 621  
Sommerfeld, 611  
Buckley-Leverett equation, 4, 52  
Burgers equation, 3, 29, 35, 90, 275, 436  
IBVP, 601, 604  
with damping, 630  
burnt/unburnt gas, 19, 188, 208  
BV (bounded variation), 31, 137, 138, 515

## C

carbuncle, 283  
Carleman model, 709

- Cauchy problem, 3, 29, 30, 56, 236  
   coupling, 736  
 cell, see also finite volume 215  
   average, 329  
   center, 494  
   dual, 495  
   entropy inequality, 217  
   rectangular, 468  
   vertex, 495  
 centroid, 494, 495, 500, 502, 520  
 CFL condition, 222, 223, 232, 235, 237, 254,  
   264, 317, 375, 382, 404, 471, 476, 506,  
   637, 686, 719, 720, 730  
 change  
   of frame, 62, 64, 66, 88, 134, 213, 256, 294,  
   685, 688, 705, 722  
   of variable, 42, 60, 61, 284, 406, 736  
 Chapman-Enskog expansion, 386, 396, 398,  
   399, 406, 414, 711  
   corrector, 396  
 Chapman-Jouguet, 188, 204, 209  
   deflagration, 194, 196  
   detonation, 194, 196, 212, 213  
 characteristic  
   boundary, 584  
   curve, 28, 90, 103, 105, 106, 108, 185, 429,  
   630  
   equation, 103–106, 429, 431, 458, 465  
   field, 59–61, 64, 65, 68, 70, 108, 644  
   form, 617  
   incoming, 585, 587, 590  
   interface, 736  
   line, 28, 30, 35, 90, 91, 185, 630  
   manifold, 593, 595  
   method of -s, 28–30, 35, 432  
   non-, 429, 439, 591, 614  
   speed, 96  
   surface, 84, 428, 455, 593  
   variable, see variable 73  
 chemical potential, 26  
 Cole-Hopf transform, 4  
 collision  
   invariant, 355  
   operator, 355  
 combustion, 213  
   front, 201  
   velocity, 192  
   wave, 189, 193  
 compatibility relations, 619  
 compensated compactness, 50, 137  
 compression wave, 91, 275, 528  
 conjugate function, 43, 442  
 conservation  
   of energy, 6, 21, 136  
   of entropy, 67  
   of mass, 6, 10, 18, 20, 66, 256  
   of momentum, 6, 22  
 conservative  
   form, 1, 6  
   variable, 10, 18, 23, 154, 186, 224  
 consistency  
   with asymptotic system, 714  
   with conservation law, 217, 252, 483  
   with entropy condition, 217, 252  
   with integral form, 251, 269, 680, 688  
   with Jacobian, 260  
 contact discontinuity, 98, 102, 117, 155, 421,  
   438  
   nonconservative case, 648  
   smearing, 274, 388, 533  
   stationary, 321, 533  
 control volume, see also cell or mesh 488  
 convergence result, 393, 395, 396, 401, 423,  
   499, 512, 516  
 convex combination, 220, 223, 290, 310,  
   479, 513  
 corrector, see Chapman-Enskog 396  
 coupled Riemann problem, 738  
 coupling, see interface 731  
 Courant–Isaacson–Rees scheme, 223  
 Courant number, 223, 637  
 critical flow (super/sub), 137, 663  
 critical time, 29, 432  
 Crussard curve, 190, 191, 194, 195, 210
- D**
- Dalton law, 151, 185  
 Darcy law, 401, 711, 723  
 Davis scheme, 225, 546, 578  
 deflagration, 192  
   speed, 197  
   strong, 194  
   weak, 194, 206  
 density, 6, 10  
 detonation, 192, 204  
   speed, 197  
   strong, 194, 210  
   weak, 194, 210  
 DG-method, 495  
 diffusive limit, 401, 708, 710, 722  
 discontinuity, 31  
   admissible, 32  
   line, 32  
   speed, 34  
   stationary, 265, 737, 738  
   surface, 32  
 dispersion relation, 234, 235, 475, 596  
 dissipative

- matrix, 49
- scheme, 233
- scheme (Kreiss), 233
- second order approximation, 396, 400
- source term, 395
  
- E**
- E-scheme, 479
- elasticity, 52, 113, 389, 397, 399, 406
- energy
  - internal, 6, 66, 357, 362
  - internal (barotropic case), 134, 411, 657
  - microscopic variable, 361
  - Saint-Venant, 136, 391
  - total, 6, 66, 70, 357
- Engquist-Osher scheme, see Osher scheme 274
- ENO, 421, 473
- enthalpy, 149, 285, 287, 288, 290, 328, 441, 454, 518, 623, 657
- entropy, 38, 39, 43, 110, 382
  - consistency, 253
  - extension, 395, 412
  - fix, 266
  - flux, 38, 39, 43, 383
  - flux (numerical), 217
  - for Euler, 368
  - inequality (discrete), 382, 416, 480, 674, 682, 729
  - jump inequality, 46, 47, 154, 165
  - kinetic, 356, 364
  - Kruzhkov, 738
  - mathematical, 21, 37, 134, 154, 657
  - minimization principle, 413
  - pair, 65, 67, 69, 70, 76, 126, 254, 319, 395, 644, 649
  - pair (linear system), 605
  - production, 480
  - relative, 710
  - solution, 46, 50, 153
  - specific, 67, 71, 75, 141, 440
  - variable, 43, 261
  - variable (Euler), 441
- entropy condition, 45, 170, 385, 396, 628, 644, 663
- interface, 652
- Lax, 108, 125, 126, 170, 438, 463
- Liu, 112, 136
  - one sided or Oleinik, 52, 421
- equal area rule, 52
- equation of state, 6, 19, 25, 26, 47, 66, 68, 70, 72, 75, 149, 153, 210, 366, 440
- $\gamma$ -law, 150
  
- F**
- finite difference scheme, 216
  - 2D, 468
  - consistent, 220
- finite volume
  - approach, 215, 237
  - control cell, 488, 494, 495
  - dual cell, 495, 496
  - method (one-dimensional), 215
  - method (two-dimensional), 488
  - scheme (cell center), 494, 535
  - scheme (cell vertex), 495, 524, 543
  - staggered, 496
  - weighted, 497
- flow in a duct, see nozzle 21
- fluid system, 6, 44, 147, 255, 420, 686, 742
- flux
  - discontinuous, 189
  - discontinuous, 655, 734
  - homogeneous, 320, 322
  - interface, 216, 237, 669
  - numerical, 216
- barotropic, 7, 15, 24, 133, 134, 566, 571, 574
- Grüneisen, 153, 158, 174, 283, 295, 337, 548, 568
- incomplete, 151
- incremental form, 293, 310
- non convex, 132
- van der Waals, 132, 151, 213
- equilibrium, 357, 364, 389–391, 395, 629, 637
- coupled Riemann problem, 738
- discrete, 638, 667, 669, 671, 695
- lake at rest, 664, 672, 675
- Saint-Venant, 663
- solution of simple scheme, 682
- strong, 684, 695
- system, 743
- equivalent equation/system, 233, 234
- Lax-Wendroff scheme, 234
- Euler equations, 6, 359, 366, 526, 548
  - 2D steady, 467
  - barotropic, 134, 641, 643, 657
  - isentropic, 411, 593
  - Lagrangian, 741
  - with gravity and friction, 16, 640, 688, 721
- Eulerian coordinates, 6, 70, 240, 692
- Euler identity, 320, 322
- evolution step, 236, 329, 375, 415
- exothermic reaction, 19, 188, 191

- flux (*cont.*)
  - numerical entropy, 217, 253
  - strictly convex, 36, 51
 flux difference splitting method, 280
 flux vector splitting method, 320
  - 2D, 531
  - kinetic, 378
  - Steger-Warming, 321, 322
  - Steger-Warming (2D), 532
  - van Leer, 326, 379
  - van Leer (2D), 532
 Fourier
  - component, 234
  - discrete mode, 235
  - Laplace–transform, 591, 598
  - mode, 232, 234, 235, 474, 475
  - transform, 230, 232, 473, 591, 715
 friction, 16, 21, 401, 640, 707
  - Darcy-Weisbach, 641, 642
  - linear, 641
 Froude number, see shallow water 136
- G**
- Galilean, see invariance 6
- gas dynamics equations, see also Euler
  - equations 5
  - 2D, 439
  - 2D isentropic, 426, 440
  - Eulerian coordinates, 6, 43, 70, 88, 102, 106, 153, 180, 240
  - incompressible, 551
  - isentropic, 5, 134, 389, 390, 407, 422
  - Lagrangian coordinates, 14, 66, 87, 100, 104, 156, 241, 309, 417
  - moving frame, 240
  - non dimensional, 548, 549
 Gay-Lussac (or Boyle) law, 149, 357
 generalized functions, 115
 genuinely nonlinear, 59, 61, 65, 69, 73, 80, 82, 90, 96, 108, 426, 645
 Gibbs potential, 442
 GKS theory, 620
 Godunov-Mock theorem, 2, 40, 426
 Godunov-type scheme, 250, 253, 259, 260, 275
 entropy satisfying, 254, 257, 313
 HLL, see HLL 269
 Godunov method, 236, 280, 382
  - for gas dynamics (Eulerian coordinates), 245
  - for gas dynamics (Lagrangian coordinates), 242
 Lagrange-projection, 246
 linear case, 239, 264
 moving grid, 249
 gravity, 16, 21, 640, 641
 grid
  - Cartesian, 468, 480, 499
  - orientation effects, 487, 494
  - uniform, 216
 group velocity, see velocity 234
 GRP, 330, 333, 579, 746
  - gas dynamics (Lagrangian), 336
  - modified, 353

**H**

Harten’s criteria, 221, 222, 542

Hessian matrix, 39, 141

HLL method, 269
 
  - all Mach, 574
  - HLLE version, 274
  - numerical flux, 270
  - semi-discrete, 568
  - source term, 679
 Hugoniot
 
  - curve, 157–159, 172, 190, 195, 210, 302, 463
  - equation, 157

hydrostatic
 
  - approximation, 21
  - pressure, 21, 678
  - reconstruction, 676
 hyperbolic system, 2
 
  - 2 equations, 103, 120, 122, 603
  - 2 equations (linear), 121, 402
  - convex, 51, 119, 138
  - linear, 55, 120, 223, 230, 238
  - linearly degenerate, 121
  - strictly, 2, 5, 7, 55, 58, 65, 426
  - Temple class, 137, 602

**I**

IBVP, 3, 739
 
  - Burgers equation, 599
  - linear 1D scalar, 582
  - linear 1D system, 587, 590
  - linear 2D scalar, 583
  - linear 2D system, 590
  - nonlinear scalar, 599
  - weak entropy solution, 600
  - well posed nonlinear 1D case, 605
 ideal gas, 149
 
  - polytropic, 7, 17, 19, 43, 68, 150, 157, 160, 163, 171, 174, 177, 182, 203, 282, 286, 290, 323, 328, 348, 440, 442, 451, 549, 558, 564, 567, 573
 incompressible
 
  - constraint, 547, 560, 568, 572

- fluid equations, 547
- incremental coefficient, 220, 221, 478, 542
- inflow/outflow
  - subsonic, 609, 616, 618, 624
  - supersonic, 606, 607, 609, 623, 624
- initial boundary value problem, see IBVP 581
- integral curve, 81, 83, 84, 97, 277, 431, 432, 648
- interface
  - characteristic, 741, 742
  - coupling, 731
  - coupling condition, 655, 731, 732, 734
  - material, 742
  - non characteristic, 741
  - numerical flux, 744
- invariance
  - Galilean, 6, 443
  - rotational, 445, 454, 464, 491, 519, 532, 554
- invariant region, 138
- irreversible reaction, 19, 207, 211
- isentrope, 197, 203
- isentropic, see gas dynamics equations 5
- isothermal flow, 135
  
- J**
- Jacobian matrix, 2, 58, 62, 68, 222, 283, 440, 527
  - eigenvector, 288
  - gas dynamics (Eulerian), 288, 699
  - gas dynamics (ideal gas), 290
- Jin-Xin, see relaxation 121
- jump, 32, 33, 38
  
- K**
- kinetic
  - discrete approximation, 391
  - equation, 390
  - formulation, 388
  - representation, 388, 390
  - scheme, 328, 368, 398
  - scheme (Euler), 354
  - scheme (numerical flux), 378
  - scheme (Saint-Venant), 678
- Kruzhkov
  - entropy, 51
  - theorem, 50, 579, 599
  
- L**
- Lagrangian coordinates, 5, 11, 62, 66, 241, 256, 685
  - 2D case, 13
- Lax-Friedrichs scheme, 221, 229, 234, 270, 382, 503, 578
  - 2D, 471
  - LLF (local), 266, 267, 525
- Lax-Wendroff
  - scheme, 221, 222, 225, 235
  - scheme (2D), 470
  - scheme (two-step Richtmyer), 222, 472, 546
  - theorem, 218, 503
  - theorem (2D), 504
- Lax entropy condition, see entropy condition 107, 109
- Leibniz integral rule, 20
- limiter
  - flux, 226
  - gradient, 534, 543
  - minmod, 226, 535, 539, 540
  - slope, 332, 348, 353, 386, 538
  - superbee , 226
  - van Albada, 546
- linear degeneracy, 426
- linear hyperbolic system, see hyperbolic system 55
- linearization
  - Euler equations, 593, 607
  - Roe-type, 259, 261, 263
- linearization, GodlewskaOlazabalRaviart1999, 139
- linearly degenerate, 59, 61, 65, 69, 73, 80, 84, 91, 97, 109, 119, 645, 703
  
- M**
- Mach
  - all - scheme, 552
  - line, 468
  - low - limit, 547
  - number, 324, 451, 468, 547, 548
- magnetohydrodynamics, see MHD 8
- mass, 357
  - flux, 155
  - fraction, 17, 19, 185, 188, 207, 741
  - variable, 12, 66, 156, 241, 701
  - variable increment, 243
- maximum principle, 381, 382, 386, 514, 517, 538
- Maxwell equations, 594, 610, 612
- Maxwellian, 355, 362, 394, 629
- mean-free path, 354
- mesh, see also finite-volume 487
  - control cell, 488
  - dual, 496, 498, 524
  - regular, 506
  - staggered, 472

- mesh (*cont.*)
  - structured, 487
  - unstructured, 487, 499, 504
- method of lines, 216, 480
- MHD equations, 8–10, 15, 75, 80, 147, 420, 423, 579
- minimum entropy principle, 357, 361, 364, 390, 391, 395, 412
- mixture model, 132, 151, 185, 297, 323, 741
  - equilibrium, 26
- moment, 357
- monoatomic perfect gas, 150, 354, 357
- monotone, see numerical scheme 220
- moving frame, 240
- multicomponent, 17, 119, 185, 213, 420, 648, 732, 741, 742
- multiple eigenvalue, 119, 452
- multiscale resolution, 707
- multivalued solution, 275, 277, 278
- MUSCL, 386, 421, 533
  - cell center, 535
  - cell vertex, 543
- N**
- Navier–Stokes equations, 3, 20, 47, 48, 283, 398, 496, 533, 744
- nonconservative, 25, 74
  - form, 27, 60, 61, 71, 74, 76, 87, 102, 104, 186, 288, 303
  - product, 24, 27, 115, 186, 309, 646, 662, 664
  - product (family of paths), 647
  - product (kinetic relation), 648
  - product (Volpert), 647
  - Riemann problem, 647
  - system, 24, 115, 644, 737
  - variables, 60, 99, 100, 337
- normal mode, 591, 597
  - analysis, 610, 612
- nozzle, 22, 642
  - barotropic, 646, 656, 665
- numerical flux, 217
  - conservation property, 489
  - consistent, 217, 219, 220, 253, 469, 483, 490, 633
  - interface, 673
  - Lax–Wendroff, 225
  - modified, 228
  - upwind, 225
- numerical scheme
  - (2k+1)-point, 217
  - 3-point, 217, 221
  - conservation form, 216, 469
  - conservative, 217
  - consistent, 260, 415
  - dispersive, 234
  - dissipative (Kreiss), 233
  - entropy satisfying, 254, 258, 259, 271, 317, 319, 408, 681, 704, 730
  - essentially 3-point, 217, 220
  - incremental form, 220, 221, 478, 484, 485, 542
  - linear, 230
  - linear (2D), 473
  - monotone, 220, 471, 477, 483, 512
  - monotonicity preserving, 220
  - positively conservative, 268, 317, 319, 320
  - relaxation, see relaxation 317
  - Rusanov, 404
  - TVD, *see* TVD
  - upstream, 321
  - upwind, 238, 239, 475
  - upwinding, 280
  - viscous form, 220, 221, 229, 232, 265
  - with flux limiter, 225
  - with modified flux, 228
- O**
- one-pressure model, 25
- order of accuracy, 219, 469, 499
  - second-order, 222, 225, 228, 234, 329, 386, 533, 637
- Osher scheme, 274
  - 2D, 528
  - gas dynamics equations, 281
  - numerical flux, 277
  - scalar case, 274, 320
- P**
- p-system, 5, 15, 39, 60, 87, 104, 122, 134, 138, 397, 399, 405, 422
  - coupling, 739
  - with friction, 707
  - with gravity, 729
- parameter vector, *see* Roe method 283
- particle
  - derivative, 455
  - path, 455, 465
- path of integration, 277, 303
- Peclet number, 707
- phase
  - error, 474
  - surface, 475
  - velocity, *see* velocity 233
- phase transition, 26, 132
- piecewise
  - $C^1$ , 32, 38, 80, 435

- constant, 92, 218, 230, 236, 374, 506, 714  
 linear, 329, 497, 533
- polar function, see conjugate 43
- polytropic, see ideal gas 5
- porous medium, 4, 579, 643, 662, 709, 711, 732
- positivity preserving, 303, 378, 408, 675, 676
- power boundedness condition, 231
- prediction-correction, 332, 533
- pressure, 6, 66, 357, 362
- pressureless gas, 133, 136
- production rate, 17, 19
- projection, 352, 353  
 $L^2$ , 236, 329, 497  
 Eulerian grid, 246  
 on equilibrium manifold, 415  
 step, 246, 329, 375
- R**
- radiative transfer, 710
- Rankine-Hugoniot  
 condition, 32, 34, 35, 38, 46, 51, 92, 110, 153, 208, 237, 256, 437, 628  
 condition (2D Euler), 461  
 set, 92, 437
- rarefaction, 107  
 curve, 83, 117, 122, 175, 434  
 fan, 36, 83, 91, 132, 436  
 wave, 82, 88, 434
- Rayleigh line, 162, 168, 192, 194, 209
- reacting gas flow, 17, 151, 185, 188  
 Z.N.D. model, 207
- real gas, see equation of state 151
- reconstruction, 236, 329, 332, 375, 415
- regularization, 647, 733  
 Dafermos, 139, 648  
 parabolic, 115  
 quadratic, 265
- relaxation, 121, 397  
 approximation, 274  
 diffusive, 401, 708  
 Euler equations, 407  
 Euler equations (barotropic), 407  
 Euler equations (Lagrangian), 417  
 instantaneous, 403, 416  
 Jin-Xin, 121, 393, 397, 401, 405, 731, 743
- $p$ -system, 405  
 scheme, 317, 394, 697, 726  
 scheme (Euler equations with gravity and friction), 702  
 scheme (isentropic Euler equations), 415
- source term, 26, 629, 640
- stiff, 707
- Suliciu, 397, 406
- system, 743
- time, 364
- to a scalar law, 399
- residual, 488
- resonance, 120, 137, 646, 650, 655, 737, 742  
 nozzle, 661
- Saint-Venant system, 664
- scalar equation, 652
- transition curve, 653
- transition manifold, 650
- Riemann invariant, 83, 85, 88, 96, 97, 121, 175, 402, 426, 645, 673, 699  
 for Euler system, 87, 88, 151  
 for Euler system (2D), 452  
 for the  $p$ -system, 87
- Riemann problem, 3  
 $p$ -system, 122, 127  
 2D, 3, 139, 464  
 Burgers, 35  
 coupled, 738  
 Euler system, 180  
 gas dynamics, 171  
 general case, 116, 236, 333  
 generalized, see GRP 329  
 LD relaxation system, 410  
 linear, 57, 238, 239, 265  
 linearized, 260  
 nonconservative, 646  
 nonuniqueness, 655  
 partial, 624  
 plane (Euler 2D), 464  
 reacting gas, 205, 213  
 scalar, 51  
 two-component flow, 186
- Riemann solver, 329  
 approximate, 250, 251, 268, 275, 278, 299, 311, 313, 316, 403, 689  
 exact, 278, 317, 403, 415  
 simple, 254, 269, 319, 420, 679, 680
- Roe-type linearization, 259, 303, 310  
 gas dynamics (real gas), 303, 305
- Roe method, 259, 280  
 2D extension, 518  
 all-Mach, 561  
 average state, 284, 290, 307  
 entropy correction, 265  
 fully 2D, 520  
 gas dynamics (Eulerian), 283  
 gas dynamics (real gas), 294  
 gas dynamics (Lagrangian), 309, 310  
 linearization, 268, 273, 284  
 linearization (2D), 520  
 matrix, 222, 228, 259

- Roe method (*cont.*)  
 numerical flux, 263  
 parameter vector, 284, 304, 518, 526  
 semi discrete scheme, 552  
 VF Roe, 268  
 with LLF, 266  
 Runge-Kutta scheme, 216, 387, 634
- S**
- Saint-Venant system, 20, 136, 641  
 kinetic equation, 391  
 scalar equation, 27  
 advection with diffusive term, 234  
 advection with dispersive term, 234  
 convex flux, 3  
 linear advection, 223, 234  
 linear advection (2D), 583  
 linear advection (geometric source), 631, 637, 668  
 linear advection (IBVP), 582  
 linear advection-reaction, 629, 635  
 linear transport, 355  
 second law of thermodynamics, 67, 134, 141, 287, 441, 688  
 self-similar  
 Riemann solver, 254, 680  
 solution, 36, 51, 57, 80, 82, 90, 333, 460, 464, 644, 646  
 solution (Euler 2D), 466  
 set of states, 1, 4, 7, 9, 319  
 shallow water  
 equations, 20, 269, 388, 391, 641, 662, 675  
 equilibrium, 675  
 Froude number, 136, 578, 663  
 lake at rest, see equilibrium 675  
 model, 398, 677  
 relaxation, 408  
 shock, 4, 30  
 -tube problem (Sod), 183  
 admissible, 51, 97, 112, 113, 163, 165, 166  
 back/front side, 164  
 compressive, 165  
 curve, 96, 117, 123, 172  
 delta-shock wave, 120, 139  
 front (multidimensional), 439  
 nonclassical, 114, 138  
 plane shock wave, 437, 461  
 profile, 113  
 slowly moving, 280, 283, 422  
 strength, 97  
 simple wave, 89, 91  
 plane-wave, 433, 435, 460  
 slab symmetry, 7, 16, 18, 66, 153
- slip line, 155  
 solution  
 classical, 27, 31, 33, 38, 46  
 distribution, 31  
 entropy, see entropy solution 46  
 equilibrium, 667  
 measure-valued, 50, 120, 136, 139, 478, 480, 515  
 nonuniqueness, 35, 36, 120, 664, 737  
 operator (exact), 637  
 plane wave, 460, 596  
 stationary, 629, 693  
 weak, 27, 31, 35, 36, 38, 45, 66, 489, 503  
 solution operator  
 discrete, 217–219, 234, 469, 634  
 exact, 221, 483, 635, 636  
 sonic  
 curve, 646  
 flow, 212  
 locus, 210  
 point, 265, 266, 268, 279, 281, 320, 321, 653  
 state, 650, 666  
 subsonic, 134, 166, 206, 210, 606  
 supersonic, 134, 166, 210, 324, 467, 606  
 transonic, 468, 661  
 sound speed, 68, 134, 142, 152, 287, 290, 451, 547, 643  
 Lagrangian, 68, 689  
 source term, 10, 16, 47, 51  
 geometric, 23, 639, 641, 643  
 measure, 731, 737  
 relaxation, 629  
 stiff, 636, 705  
 specific  
 energy, 6, 9  
 enthalpy, 149, 191, 285, 287, 293, 441, 549  
 entropy, see entropy (specific) 43  
 heat, 18, 43, 68, 149, 150, 358  
 volume, 12, 66  
 speed of propagation, 34, 35, 50–52, 57, 92, 96, 154, 254, 628  
 splitting  
 dimensional, 480  
 fast/slow wave, 707  
 flux -, see flux splitting 280  
 method, 633  
 operator, 369, 415, 720, 729  
 Strang, 481, 635  
 Trotter formula, 482, 634  
 stability  
 $L^1$ , 378, 379, 517  
 $L^2$ , 230–232, 273, 473, 714, 719  
 $L^\infty$ , 220, 268, 478, 479, 512, 540, 633

- BV, 220, 515  
 BV (weak), 516, 517  
 CFL condition, see CFL 222  
 condition, 216, 396, 439  
 condition (structural), 396  
 GKS theory, 620  
 of a numerical scheme, 220  
 sub-characteristic condition (Whitham), 393, 396, 400, 402, 405, 408, 703, 729  
 Steger-Warming, see flux vector splitting 321  
 stiffened gas, 153, 283, 568  
 streamline, 455, 456, 467  
 sub-characteristic, see stability condition 396  
 symbol of differential operator, 616  
 symmetric system, 39  
 Symmetrizable, 426  
 symmetrizable, 2, 7, 14, 441
- T**  
 Taylor wave, 212  
 telegraph equations, 708  
 temperature, 47, 67, 149, 151, 362  
     kinetic, 357  
 thermodynamics, 67, see also second law 68, 141  
 time scheme  
     explicit, 216  
     implicit, 216, 403, 421, 634, 635, 707, 719, 720, 728, 729  
     time step, 216  
 topography, 20  
 total variation  
     2D, 476  
     bounded-, 478, 485, 487  
     diminishing, see TVD 220  
     TV, 220, 515  
 traffic model, 52, 139, 423  
     junction, 733  
     LWR, 4  
 transport-collapse, 52  
 travelling wave, 113, 115, 208  
     velocity, 208  
 Trotter formula, see splitting 482  
 truncation error, 219, 233, 499  
 TVB, 487  
 TVD scheme, 220, 229, 268, 478, 484  
 two-phase flow, 24, 27, 388, 420, 639, 640, 643, 648, 662, 706, 741, 743  
     Baer-Nunziato, 662  
     HEM/HRM, 662, 732
- U**  
 umbilic point, 120, 138  
 uniform Kreiss condition, 591, 592, 598  
 uniqueness result, 137, 516, 652, 740  
     Cauchy problem, 50  
     IBVP, 600, 606  
     scalar Cauchy problem, 50, 628
- upwind  
     flux, 745  
     scheme, 320, 391, 393, 402, 547  
     scheme (characteristic variable), 402  
     scheme (scalar linear), 223, 635  
 upwinding (source term), 637, 668, 670
- V**  
 vacuum, 183, 411  
     dry zone, 642, 675  
      $p$ -system, 131  
 vanishing viscosity, 44, 46, 49, 50, 52, 137, 599, 647  
 van Leer method, see also flux vector splitting 326  
     Lagrange+projection, 348  
     predictor-corrector, 331, 534  
     second order, 329  
     with GRP, 345
- variable  
     characteristic, 73, 120, 224, 264, 322, 588, 605  
     conservative, see conservative 8, 736  
     dimensionless, 548  
     entropy, see entropy 43  
     incoming/outgoing, 587, 598, 607, 609  
     nonconservative, 268, 450  
     primitive, 74, 171, 269, 348, 426, 453, 742  
     space, 1, 7  
     thermodynamic, 47, 68, 72, 141, 157, 175, 307
- velocity, 6, 10, 66, 357  
     frame, 240  
     group, 234, 235, 475, 595, 596  
     group (tangential), 598  
     normal, 475, 618  
     peculiar, 357  
     phase, 233, 234, 475, 596  
     relative, 154, 164, 166, 189, 462  
     tangential component, 461, 462, 529
- viscosity  
     coefficient, 221, 222, 267  
     matrix, 49, 113, 115, 265, 602, 648  
     matrix (numerical), 229, 230, 232, 265, 273  
     term, 44, 321

- viscous
  - form, 220
  - perturbation, 44, 47, 48, 600
  - profile, 49, 113–116
- volume fraction, 24
- von Neumann
  - condition, 231, 474
  - state, 212
- W**
- wave
  - acoustic, 451, 452, 578, 596, 657
  - amplitude, 234
  - composite, 119, 132, 136, 138, 204, 205, 651, 652
  - curve, 740
  - dispersive, 234
  - entropy, 451
- Y**
- Young measure, 479, 480, 515, 516
- Z**
- ZND model, see reacting gas 196