NTNU

Norwegian University of
Science and Technology

Hyperbolic Conservation Laws with Relaxation Terms

A Theoretical and Numerical Study

Peder Kristian Aursand

Master of Science in Physics and Mathematics

Submission date:	June 2011
Supervisor:	Ingve Simonsen, IFY
Co-supervisor:	Tore Flåtten, SINTEF Energi AS
	Svend Tollak Munkejord, SINTEF Energi AS

Norwegian University of Science and Technology
Department of Physics

Abstract

Hyperbolic relaxation systems is an active field of research, with a large number of applications in physical modeling. Examples include models for traffic flow, kinetic theory and fluid mechanics.

This master's thesis is a numerical and theoretical analysis of such systems, and consists of two main parts: The first is a new scheme for the stable numerical solution of hyperbolic relaxation systems using exponential integrators. First and second-order schemes of this type are derived and some desirable stability and accuracy properties are shown. The scheme is also used to solve a granular-gas model in order to demonstrate the practical use of the method.

The second and largest part of this thesis is the analysis of the solutions to 2×2 relaxation systems. In this work, the link between the the sub-characteristic condition and the stability of the solution of the relaxation system is discussed. In this context, the sub-characteristic condition and the dissipativity of the Chapman–Enskog approximation are shown to be equivalent in both 1-D and 2-D.

Also, the dispersive wave dynamics of hyperbolic relaxation systems is analyzed in detail. For 2×2 systems, the wave-speeds of the individual Fourier-components of the solution are shown to fulfill a transitional sub-characteristic condition. Moreover, the transition is monotonic in the variable $\xi = \varepsilon k$, where ε is the relaxation time of the system and k is the wave-number.

A basic 2×2 model is used both as an example-model in the analytical discussions, and as a model for numerical tests in order to demonstrate the implications of the analytical results.

Preface

This master's thesis is the conclusion of a five-year integrated master's programme in applied physics and mathematics at NTNU, Trondheim. The work was carried out at SINTEF Energy Research (the CO_2 Dynamics project) in cooperation with the Department of Physics at NTNU.

The thesis consists of two parts: The first is the completion of the numerical work started in the pre-master's project and the second is the analysis of 2×2 relaxation systems. While I recognize that the completeness and unity of the thesis might have suffered because of this division into two parts, I am happy with the results achieved on both fronts. The work on the completion of the pre-master's project introduced me to the process of scientific publication—a valuable and motivating experience for a master's student. Also, the analytical work on the 2×2 relaxation systems proved to be a very interesting and challenging endeavor. The work was mainly mathematical in nature, but physical considerations were also necessary in order to interpret the results. Being a physics student with an interest in mathematics, this combination made the project very motivating.

I feel privileged for having had Tore Flåtten and Svend Tollak Munkejord at SINTEF as my supervisors during this work. Their continuous help and support, along with giving me the freedom to pursue topics of my interest, has been invaluable. I am grateful to Professor Ingve Simonsen for always having an open door and helping me throughout the work on this thesis. I also want to thank Halvor Lund and Alexandre Morin for fruitful discussions about relaxation systems and numerics.

Peder Aursand

Peder Aursand
Trondheim, June 2011

Contents

1 Introduction

1.1 Hyperbolic Relaxation Systems

We consider systems of conservation laws with relaxation terms. Such a system of N equations in d spatial dimensions can be written in the general relaxation-form

$$\partial_t \boldsymbol{U} + \sum_{j=1}^{d} \partial_{x_j} \boldsymbol{F}_j(\boldsymbol{U}) = \frac{1}{\varepsilon} \boldsymbol{R}(\boldsymbol{U}), \tag{1.1}$$

where $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{x}, t)$ is an N-vector of physical quantities and $\boldsymbol{F}_j(\boldsymbol{U})$ represents the flux in the j'th spatial dimension. $\boldsymbol{R}(\boldsymbol{U})$ is a *relaxation term*, and represents the force driving the perturbed system towards equilibrium. The parameter $\varepsilon > 0$ can be seen as a characteristic time-scale of the relaxation process.

Hyperbolic relaxation systems are useful in describing the transport of a set of physical quantities in a non-equilibrium (perturbed) state. Therefore, these kinds of systems have a large number of applications in the physical modeling of different phenomena. Important examples include traffic flow [1], kinetic theory [6] and gas flow in local thermal non-equilibrium [25, 13]. In certain cases, relaxation models can also be used in the numerical solution of an equilibrium model [23]. These *relaxation schemes* exploit the fact that a relaxation model can be easier to solve numerically than its non-perturbed counterpart.

An important concept for such relaxation models is that of equilibrium. For every non-equilibrium state $\boldsymbol{U}$, there is a corresponding equilibrium approximation characterized by $\boldsymbol{R}(\boldsymbol{U}) = 0$. Furthermore, the dynamics of the equilibrium approximation can be described through a set of $n \leq N$ conservation laws

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} \partial_{x_j} \boldsymbol{f}_j(\boldsymbol{u}) = 0, \tag{1.2}$$

for some reduced variable $\boldsymbol{u}(\boldsymbol{x}, t)$. Also, since ε is a time-scale of the relaxation, the limit $\varepsilon \to 0$ can be seen as the equilibrium-limit of the relaxation model. In other words, the equilibrium model can be seen as the limiting case of the relaxation model where the relaxation speed tends to infinity.

1.2 Previous Work

Hyperbolic relaxation systems has been an active field of research for the last 20 years.

For the stability of the relaxation system, there exists an important constraint that says that the wave-speeds of the local equilibrium approximation (1.2) should be interlaced in the wave-speeds of the homogeneous relaxation system ($\varepsilon \to \infty$). This sub-characteristic condition was introduced in the linear case by Whitham [46] and later for non-linear 2×2 systems by Liu [31]. The topic was further developed for $N \times N$ systems by Chen et al. [8], and shown to be directly related to the convexity of the entropy density of the relaxation system.

Since the pioneering work by Liu [31], the study of 2×2 systems have been an important sandbox for the theoretical analysis of hyperbolic relaxation systems [9, 28, 32, 24]. This can be fruitful because 2×2 systems contain much of the same elements of complexity as a general system, while being less cumbersome to work with. Another important approach is the analysis of linearized relaxation systems. Herein, a notable contribution was made by Yong [47, 48], who derived stability criteria based on the structure of relaxation systems. Also, in a recent work by Barker et al. [2] the dynamics of the solution of the St. Venant equations was investigated by studying the dispersion relation of the corresponding linearized system.

The numerical solution of hyperbolic relaxation systems has also been a popular topic in the recent years. In particular, the stable numerical solution of such systems in the *stiff* limit ($\varepsilon \to 0$) has been the subject of numerous studies [38, 22, 36].

1.3 Scope of Work

It has been established that for well-behaved systems it is expected that the solutions of the relaxation system will approach that of the local equilibrium approximation as $\varepsilon \to 0$ [35, 8]. A characteristic feature of hyperbolic partial differential equations is the wave-nature of the solutions. The homogeneous hyperbolic system, seen as the limit of (1.1) when $\varepsilon \to \infty$, describe N waves; the equilibrium system (1.2), seen as the limit $\varepsilon \to 0$, describe $n \leq N$ waves. Therefore, the relaxation term will in some way influence both the number of waves and the wave-speeds.

The main purpose of this work is to investigate the wave-dynamics of the relaxation model (1.1), and in particular how it relates to the dynamics of the corresponding equilibrium model (1.2). This thesis aims to illuminate the mechanism responsible for changing the wave-dynamics of the relaxation model into the wave-dynamics of the local equilibrium model as ε gets small. This will be done through the linear analysis of 2×2 systems.

The approach used in this thesis is similar to that of Yong [48], who used linear analysis to investigate stability of relaxation systems. Also, some of the present work bears some similarity to a recent work by Barker et at. [2], who analyzed the dispersion relation of the linearized St. Venant equations. However, to the author's knowledge, there has been little or no work along the lines of using linear analysis to do a detailed study of the wave-speeds of the relaxation system, and how they behave for different values of the relaxation time.

The secondary purpose of this thesis is to develop a robust numerical scheme for solving monotonic relaxation systems. This is the continuation of work started by the author during a pre-master's project.

1.4 Structure of the Thesis

This thesis is organized as follows: Chapter 2 gives a brief introduction to hyperbolic relaxation systems, and explains some of the most important general concepts relevant to the discussions in this thesis.

In Chapter 3, 2×2 systems are discussed. Particular attention is given to the relationship between the sub-characteristic condition and the stability and well-posedness of the relaxation system.

Chapter 4 is devoted to the analysis of the wave-dynamics of 2×2 systems. Wave-speeds and amplification factors, and their dependence on the stiffness of the system, is analyzed in detail.

Chapter 5 is a self-contained journal article about the numerical solution of hyperbolic relaxation systems using exponential time-differencing. The article is co-written by Steinar Evje, Tore Flåtten, Knut Erik Giljarhus and Svend Tollak Munkejord.

Chapter 6 contains the results of numerical simulations performed on a basic 2×2 system. The purpose of this chapter is two-fold: Both to test the implications of the analysis done in Chapters 3 and 4, and to demonstrate the practical use of the numerical scheme proposed in Chapter 5.

In Chapter 7 the main conclusions are drawn, and possible topics for further work are outlined.

2 Hyperbolic Relaxation Systems

The purpose of this chapter is to give a brief introduction to the basic concepts related to hyperbolic relaxation systems. Particular emphasis is given on the subjects most relevant for this thesis: The relationship between the relaxation model and the equilibrium approximation.

2.1 Hyperbolic Conservation Laws and Characteristics

Consider a linear system of N conservation laws in one spatial dimension given by

$$\partial_t \boldsymbol{U} + A\, \partial_x \boldsymbol{U} = 0, \tag{2.1}$$

where $\boldsymbol{U} = \boldsymbol{U}(x, t)$ is an N-vector of conserved variables and A is an $N \times N$ matrix.

We say that a system in the form (2.1) is *hyperbolic* if the matrix A is diagonalizable as

$$A = P^{-1} \Lambda P, \tag{2.2}$$

where $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_N\}$ is a diagonal matrix consisting of the *real* eigenvalues of A. By multiplying (2.1) with P from the left, we can write the conservation law as

$$\partial_t \boldsymbol{W} + \Lambda\, \partial_x \boldsymbol{W} = 0 \tag{2.3}$$

in terms of the variable $\boldsymbol{W} = P\boldsymbol{U}$. Since Λ is diagonal, the system (2.3) consists of N decoupled advection equations

$$\partial_t W_i + \lambda_i\, \partial_x W_i = 0 \quad \forall i \in \{1, \ldots, N\} \tag{2.4}$$

with solutions

$$W_i = W_i(x - \lambda_i t). \tag{2.5}$$

The lines in the $x - t$ plane given by $x - \lambda_i t$ are called the *characteristics* of the system; the eigenvalues λ_i are called the characteristic speeds—or wave-speeds.

A *non-linear* system of conservation laws in d spatial dimensions can be written as

$$\partial_t \boldsymbol{U} + \sum_{j=1}^{d} \partial_{x_j} \boldsymbol{F}_j(\boldsymbol{U}) = 0, \tag{2.6}$$

where $\boldsymbol{F}_j(\boldsymbol{U})$ is the flux in the j'th spatial direction.

By applying the chain rule to the divergence term, we can write the system (2.6) in the *quasi-linear* form

$$\partial_t \boldsymbol{U} + \sum_{j=1}^{d} A_j(\boldsymbol{U})\, \partial_{x_j} \boldsymbol{U} = 0, \tag{2.7}$$

where $A_j(\boldsymbol{U})$ is the Jacobian matrix

$$A_j(\boldsymbol{U}) \equiv \frac{\partial \boldsymbol{F}_j(\boldsymbol{U})}{\partial \boldsymbol{U}}. \tag{2.8}$$

The following is then the proper generalization of hyperbolicity to systems in the form (2.6) [30]:

Definition 1 (Hyperbolicity)**.** *A conservation law in the form* (2.6) *is* **hyperbolic** *if for all* $\boldsymbol{k} \in \mathbb{R}^d$ *the* $N \times N$ *matrix*

$$J_N(\boldsymbol{k}) \equiv \sum_{j=1}^{d} k_j \frac{\partial \boldsymbol{F}_j(\boldsymbol{U})}{\partial \boldsymbol{U}} \tag{2.9}$$

is diagonalizable with real eigenvalues.

Definition 1 imposes a strong condition for hyperbolicity, demanding that any linear combination of the individual Jacobian matrices should be real diagonalizable. As discussed by LeVeque [30, Ch. 18], this condition ensures that the system admits well-defined waves in arbitrary directions in the d-dimensional spatial domain—not just along the axes.

2.2 Relaxation Systems

We now consider relaxation systems in d spatial dimensions consisting of N equations, which can be written in the general form

$$\partial_t \boldsymbol{U} + \sum_{j=1}^{d} \partial_{x_j} \boldsymbol{F}_j(\boldsymbol{U}) = \frac{1}{\varepsilon} \boldsymbol{R}(\boldsymbol{U}). \tag{2.10}$$

In the above, $\boldsymbol{R}(\boldsymbol{U})$ is a local relaxation term and represents the driving-force of the relaxation towards an equilibrium, characterized by $\boldsymbol{R}(\boldsymbol{U}) = 0$. The relaxation time ε can be seen as a characteristic time-scale of the relaxation process.

When discussing systems in the form (2.10), a crucial assumption is that of hyperbolicity of the left hand side:

Assumption 2.1 (Hyperbolicity)**.** *The homogeneous conservation law corresponding to the left hand side of the relaxation system* (2.10) *is hyperbolic.*

We can also note that under the assumption of hyperbolicity, the *homogeneous* relaxation system is a hyperbolic conservation law, and thus describes well-defined waves.

Definition 2. *The* **homogeneous relaxation system** *is the hyperbolic conservation laws resulting from the removal of the relaxation term.*

Following the formalism of Chen et al. [8], we let the relaxation term be endowed with an $n \times N$ matrix $\mathcal{Q}$ of rank $n < N$ with the property

$$\mathcal{Q}\boldsymbol{R}(\boldsymbol{U}) = 0. \tag{2.11}$$

Multiplying (2.10) with $\mathcal{Q}$ from the left yields a system of n conservation laws in the *reduced* variable $\boldsymbol{u} = \mathcal{Q}\boldsymbol{U}$, given by

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} \partial_{x_j} \mathcal{Q}\boldsymbol{F}_j(\boldsymbol{U}) = 0. \tag{2.12}$$

The underlying conservation law (2.12) is fulfilled for every solution of the full hyperbolic relaxation system (2.10).

Furthermore, it is assumed that these reduced variables uniquely determine an equilibrium state $\boldsymbol{U} = \mathcal{E}(\boldsymbol{u})$ such that

$$\boldsymbol{R}\left(\mathcal{E}(\boldsymbol{u})\right) = 0. \tag{2.13}$$

In other words, there is always a conserved set of reduced variables $\boldsymbol{u}$ that uniquely determine the equilibrium state of the full set of variables, $\boldsymbol{U}$.

2.2.1 Linear Analysis

Linear analysis of (2.10) is in many cases an important tool for analyzing the properties of hyperbolic relaxation systems [47, 48, 2].

Let $\hat{\boldsymbol{U}}$ be a constant equilibrium state, i.e. a constant state that satisfies the equilibrium condition

$$\boldsymbol{R}(\hat{\boldsymbol{U}}) = 0. \tag{2.14}$$

The relaxation system (2.10) linearized around $\hat{\boldsymbol{U}}$ can then be written as

$$\partial_t \boldsymbol{U}' + \sum_{j=1}^{d} A^{(j)} \partial_{x_j} \boldsymbol{U}' = \frac{1}{\varepsilon} B \boldsymbol{U}', \tag{2.15}$$

where $\boldsymbol{U}' = \boldsymbol{U} - \hat{\boldsymbol{U}}$ and

$$A^{(j)} = \left.\frac{\partial \boldsymbol{F}_j(\boldsymbol{U})}{\partial \boldsymbol{U}}\right|_{\hat{\boldsymbol{U}}} \quad \text{and} \quad B = \left.\frac{\partial \boldsymbol{R}(\boldsymbol{U})}{\partial \boldsymbol{U}}\right|_{\hat{\boldsymbol{U}}}, \tag{2.16}$$

are both $N \times N$ matrices with constant coefficients.

In order to avoid unnecessary notation when performing linear analysis in this thesis, primes will be omitted and there will be an implicit linearization around a constant equilibrium state.

For linear systems, a solution of the relaxation system can be written in terms of its Fourier components:

Proposition 2.1. *A solution $\boldsymbol{U}(\boldsymbol{x}, t)$ of* (2.15) *is given by*

$$\boldsymbol{U}(\boldsymbol{x}, t) = \sum_{\boldsymbol{k}} \boldsymbol{U}_{\boldsymbol{k}}(\boldsymbol{x}, t) = \sum_{\boldsymbol{k}} \exp\left(H(\boldsymbol{k})\, t\right) \exp\left(i\boldsymbol{k} \cdot \boldsymbol{x}\right) \boldsymbol{a}(\boldsymbol{k}), \tag{2.17}$$

where we sum over all wave-numbers $\boldsymbol{k}$ and

$$H(\boldsymbol{k}) = \frac{1}{\varepsilon}\left(B - i\varepsilon \sum_{j=1}^{d} k_j A^{(j)}\right). \tag{2.18}$$

Proof. By inserting the solution (2.17), we obtain the left hand side of (2.15), given by

$$\begin{aligned}\partial_t \boldsymbol{U}(\boldsymbol{x},t) + \sum_{j=1}^{d} A^{(j)} \partial_{x_j} \boldsymbol{U}(\boldsymbol{x},t) &= \sum_{\boldsymbol{k}} \left(H(\boldsymbol{k}) + i \sum_{j=1}^{d} k_j A^{(j)} \right) \boldsymbol{U}_{\boldsymbol{k}}(\boldsymbol{x},t) \\ &= \frac{1}{\varepsilon} B \boldsymbol{U}(\boldsymbol{x},t). \qquad (2.19)\end{aligned}$$

□

Moreover, as discussed by Yong [48], the solution (2.17) is in fact the *general* solution.

The study of linearized hyperbolic relaxation systems can be fruitful because of this natural splitting of the solution into Fourier-components. This makes the analysis easier, and one then hopes that some of the results derived in the linearized case can be valid also in the general non-linear case.

In this context, a notable contribution was made by Yong [47, 48], who used linear analysis to derive stability criteria for relaxation systems.

2.3 The Local Equilibrium Approximation

The limit where the relaxation time ε tends to zero can be seen as the limit where the relaxation speed tends to infinity. Therefore, in this limit we can expect well-behaved relaxation systems to become equivalent to their corresponding local equilibrium approximation [35]. Local equilibrium is characterized by the equilibrium condition

$$\boldsymbol{U} = \mathcal{E}(\boldsymbol{u}), \qquad (2.20)$$

combined with the $n \times n$ system of conservation laws

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} \partial_{x_j} \boldsymbol{f}_j(\boldsymbol{u}) = 0 \qquad (2.21)$$

for the reduced variable $\boldsymbol{u}(\boldsymbol{x},t)$. In the above, the equilibrium flux $\boldsymbol{f}(\boldsymbol{u})$ is defined as

$$\boldsymbol{f}_j(\boldsymbol{u}) \equiv \mathcal{Q} \boldsymbol{F}_j \left(\mathcal{E}(\boldsymbol{u}) \right). \qquad (2.22)$$

The constraint (2.20) is the assumption of local equilibrium; the conservation law (2.21) governs the equilibrium-dynamics of the remaining $n \leq N$ independent physical variables $\boldsymbol{u}(\boldsymbol{x}, t)$.

The *homogeneous* relaxation system, seen as the limit $\varepsilon \to \infty$, is a hyperbolic conservation law and will in general describe the dynamics of N waves. On the other hand, the $n \times n$ system describing the equilibrium approximation (2.21) will describe n waves. The assumption of local equilibrium therefore has a direct influence on the wave-dynamics of the relaxation system. This influence will change the number of waves, but also the wave-speeds, as will be discussed Chapter 4.

2.4 The Sub-Characteristic Condition

An important concept regarding the relationship between the wave-dynamics of the relaxation system and the local equilibrium approximation is the *sub-characteristic condition.*

2.4.1 Original Formulation

The concept was first introduced by Whitham [46] for linear systems and later by Liu [31] for non-linear 2 × 2 systems. Liu considered relaxation systems in the form

$$\partial_t u + \partial_x f(u, v) = 0, \tag{2.23a}$$

$$\partial_t v + \partial_x g(u, v) = h(u, v). \tag{2.23b}$$

We will assume that for any u there is an unique equilibrium solution $v = v_*(u)$ such that

$$h(u, v_*(u)) = 0, \tag{2.24}$$

giving the local equilibrium approximation

$$\partial_t u + \partial_x f(u, v_*(u)) = 0, \tag{2.25a}$$

$$v = v_*(u). \tag{2.25b}$$

Now, let $\lambda_\pm$ be the eigenvalues—or characteristic speeds—of the Jacobian matrix

$$A = \begin{bmatrix} \partial_u f & \partial_v f \\ \partial_u g & \partial_v g \end{bmatrix}, \tag{2.26}$$

and $\lambda_* = \partial_u f(u, v_*(u))$ the characteristic speed of the local equilibrium approximation. The original sub-characteristic condition, as introduced by Liu, then reads

$$\lambda_- < \lambda_* < \lambda_+. \tag{2.27}$$

As has been pointed out by Natalini [36], the sub-characteristic condition can be interpreted as a causality principle. In the hyperbolic relaxation system, information propagates with the characteristic speeds $\lambda_\pm$. The sub-characteristic conditions then states that the assumption of local equilibrium cannot cause information to propagate faster than in the full system.

2.4.2 General Case

A generalization of the sub-characteristic condition for $N \times N$ systems in the form (2.10) is as follows [46, 8]:

Definition 3 (The Sub-Characteristic Condition)**.** *Let $\{\Lambda_i\}$ be the ordered eigenvalues of the $N \times N$ Jacobian matrix*

$$J_N(\boldsymbol{k}) = \sum_{j=1}^{d} k_j \frac{\partial \boldsymbol{F}_j(\boldsymbol{U})}{\partial \boldsymbol{U}}, \tag{2.28}$$

and $\{\lambda_i\}$ the ordered eigenvalues of the $n \times n$ Jacobian matrix

$$J_n(\boldsymbol{k}) = \sum_{j=1}^{d} k_j \frac{\partial \boldsymbol{f}_j(\boldsymbol{u})}{\partial \boldsymbol{u}}, \tag{2.29}$$

then $\{\Lambda_i\}$ and $\{\lambda_i\}$ are interlaced according to

$$\lambda_i \in [\Lambda_i, \Lambda_{i+N-n}], \tag{2.30}$$

for all wave-numbers $\boldsymbol{k} \in \mathbb{R}^d$.

Ever since this causality principle was first introduced by Liu, it has been shown to be intimately connected to many different aspects of hyperbolic relaxation systems. Chen et al. [8] showed that, for 2×2 systems in one spatial dimension, the diffusion term in the Chapman–Enskog expansion (2.33) is dissipative if an only if the sub-characteristic condition is fulfilled.

Also, Yong [48] showed that, for 2×2 systems in general spatial dimensions, the sub-characteristic condition is equivalent to the condition of linear stability. These important relationships will be discussed in detail for 2×2 systems in Chapter 3

2.5 The Chapman–Enskog Expansion

For many physical relaxation models, the characteristic time-scale ε is small but finite. When this is the case, the local equilibrium approximation

$$\boldsymbol{U} = \mathcal{E}(\boldsymbol{u}) \tag{2.31}$$

cannot be expected to be valid. Instead, one could seek formal corrections to the equilibrium approximation—in orders of the small parameter ε.

This general approach is inspired by the Chapman–Enskog expansion for kinetic theory, in which the diffusion term of the Navier–Stokes equation was derived as a first-order correction to kinetic equilibrium [7]. For this reason, a first-order correction to a relaxation equilibrium is sometimes referred to as a Navier–Stokes-level correction.

Chen et al. [8] gave a generalization of this idea to hyperbolic relaxation systems in the form (2.10). A formal expansion of the solution around the equilibrium can be written as

$$\boldsymbol{U} = \mathcal{E}(\boldsymbol{u}) + \varepsilon \boldsymbol{U}^{(1)} + \varepsilon^2 \boldsymbol{U}^{(2)} + \mathcal{O}(\varepsilon^3). \tag{2.32}$$

As showed rigorously by Chen et al. [8], truncating this expansion at the Navier–Stokes level yields the convection-diffusion equation

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} \partial_{x_j} \boldsymbol{f}_j(\boldsymbol{u}) = \varepsilon \sum_{j,l=1}^{d} \partial_{x_j} D_{jl}(\boldsymbol{u}) \partial_{x_l} \boldsymbol{u}, \tag{2.33}$$

where $D_{jl}(\boldsymbol{u})$ is a diffusion-tensor. In other words, the first-order correction manifests itself as a diffusion-term in the conservation-law for the reduced variables.

3 Linear 2 × 2 Hyperbolic Relaxation Systems

Since the classic paper by Liu [31], the study of 2×2 models has been an essential tool in the study of hyperbolic relaxation systems. The rationale for this is that 2×2 systems contain many of the same elements of complexity as general $N \times N$ systems, while being less cumbersome to work with analytically.

In particular, 2×2 systems are the smallest systems where it makes sense to talk about *equilibrium dynamics*. For scalar relaxation equations, the local equilibrium approximation will be a constant solution; for 2×2 systems on the other hand, the equilibrium system (2.12) can be a scalar conservation law with a well-defined characteristic.

3.1 The General Model

We consider linearized 2×2 relaxation systems in the form

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} A^{(j)} \partial_{x_j} \boldsymbol{u} = \frac{1}{\varepsilon} R \boldsymbol{u}, \tag{3.1}$$

where $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t)$ is a 2-vector of basic physical variables and

$$A^{(j)} = \begin{bmatrix} a_{11}{}^{(j)} & a_{12}{}^{(j)} \\ a_{21}{}^{(j)} & a_{22}{}^{(j)} \end{bmatrix}, \quad j \in \{1, \dots, d\}. \tag{3.2}$$

In the above, R is a 2×2 relaxation matrix, the meaning of which will be clarified later in this chapter.

Hyperbolicity of the Flux Term

For systems in the form (3.1), the eigenvalues of the Jacobian matrix (2.9) of the flux term are given by a direct calculation as

$$\mu(\boldsymbol{k})_\pm = \frac{1}{2}\sum k^{(j)}\left({a_{11}}^{(j)} + {a_{22}}^{(j)}\right) \pm \left(\frac{1}{4}\left(\sum k^{(j)}\left({a_{11}}^{(j)} + {a_{22}}^{(j)}\right)\right)^2 \right.$$
$$\left. - \left(\sum k^{(j)}{a_{11}}^{(j)}\right)\left(\sum k^{(j)}{a_{22}}^{(j)}\right) + \left(\sum k^{(j)}{a_{12}}^{(j)}\right)\left(\sum k^{(j)}{a_{21}}^{(j)}\right)\right)^{1/2}, \tag{3.3}$$

where all sums are taken over the spatial dimensions. The hyperbolicity assumption (Assumption 2.1 on page 7) is then given explicitly as

$$\frac{1}{4}\left(\sum k^{(j)}\left({a_{11}}^{(j)} + {a_{22}}^{(j)}\right)\right)^2 - \left(\sum k^{(j)}{a_{11}}^{(j)}\right)\left(\sum k^{(j)}{a_{22}}^{(j)}\right)$$
$$+ \left(\sum k^{(j)}{a_{12}}^{(j)}\right)\left(\sum k^{(j)}{a_{21}}^{(j)}\right) \geq 0, \quad \forall \boldsymbol{k} \in \mathbb{R}^d. \tag{3.4}$$

3.1.1 Structure of The Relaxation Matrix

Before discussing 2 × 2 systems further, we will make the following assumption regarding the structure of the relaxation matrix R:

Assumption 3.1. *The* 2 × 2 *relaxation matrix R has rank 1.*

The rationale behind Assumption 3.1 becomes clear when considering the other two possible cases: If R has rank 0, then it is the zero matrix and there is no relaxation effect in the system (3.1). Moreover, if R has rank 2, then the local equilibrium assumption

$$R\boldsymbol{u} = 0 \tag{3.5}$$

will impose two linearly independent restrictions on the 2-vector $\boldsymbol{u}$. When this is the case, the local equilibrium approximation will be a constant solution. Thus, for studying the relationship between the dynamics of the relaxation system and that of the equilibrium system, Assumption 3.1 is the only interesting choice.

Any 2 × 2 matrix with rank 1 can, up to a row-swap, be written in the form

$$R = \begin{bmatrix} r_{11} & r_{12} \\ Kr_{11} & Kr_{12} \end{bmatrix}. \tag{3.6}$$

Therefore, there will exist a matrix

$$T = \begin{bmatrix} 1 & 0 \\ -K & 1 \end{bmatrix} \tag{3.7}$$

representing change of variables $\boldsymbol{u} \to T\boldsymbol{u}$ and a corresponding similarity transform $R \to TRT^{-1}$, yielding a relaxation matrix with zeroes in the first row. We can therefore let

$$R = \begin{bmatrix} 0 & 0 \\ r_{21} & r_{22} \end{bmatrix}, \tag{3.8}$$

by simply assuming that this change of variables already has been performed.

We now make another assumption about the matrix R:

Assumption 3.2. *The non-zero eigenvalue of the* 2×2 *relaxation matrix* R *has a negative real part.*

Assumption 3.2 is simply a stability requirement [47, 16]. It is straightforward to verify that Assumption 3.2 for the matrix (3.8) requires $r_{22} < 0$. For the rest of this chapter we will therefore assume, without loss of generality beyond Assumption 3.1 and Assumption 3.2, that R is in the form

$$R = \begin{bmatrix} 0 & 0 \\ r_{21} & -1 \end{bmatrix}. \tag{3.9}$$

In the above, the absolute value of r_{22} has been absorbed into the relaxation time ε.

3.1.2 General Solution

As discussed in Chapter 2, the general solution of the relaxation system (3.1) can be written in the form of plane waves as

$$\boldsymbol{u}(\boldsymbol{x}, t) = \sum_{\boldsymbol{k}} \boldsymbol{u}_{\boldsymbol{k}}(\boldsymbol{x}, t) = \sum_{\boldsymbol{k}} \exp\left(H(\boldsymbol{k})\, t\right) \exp\left(i\boldsymbol{k} \cdot \boldsymbol{x}\right) \boldsymbol{a}(\boldsymbol{k}), \tag{3.10}$$

where H is a 2×2 matrix given by

$$H(\boldsymbol{k}) = \frac{1}{\varepsilon} \left(R - i\varepsilon \sum_{j=1}^{d} k_j A^{(j)} \right). \tag{3.11}$$

If H is diagonalizable, it can be written as

$$H = P\Lambda P^{-1} \quad \text{with} \quad \Lambda = \begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix}, \tag{3.12}$$

where $\lambda_\pm$ are the eigenvalues of H. In general, the eigenvalues are complex, and we can denote $\lambda_\pm = \mathrm{Re}\lambda_\pm + i\,\mathrm{Im}\lambda_\pm$. Now, if we insert the diagonalization of H into (3.10) and rearrange, the general solution takes the form

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{x},t) = \sum_{\boldsymbol{k}} \Big[& \boldsymbol{u}_+(\boldsymbol{k}) \exp\left(i\left(\boldsymbol{k}\cdot\boldsymbol{x} + \mathrm{Im}\lambda_+\, t\right)\right) \exp\left(\mathrm{Re}\lambda_+\, t\right) \\ & + \boldsymbol{u}_-(\boldsymbol{k}) \exp\left(i\left(\boldsymbol{k}\cdot\boldsymbol{x} + \mathrm{Im}\lambda_-\, t\right)\right) \exp\left(\mathrm{Re}\lambda_-\, t\right)\Big], \end{aligned} \tag{3.13}$$

for some unspecified amplitudes $\boldsymbol{u}_+(\boldsymbol{k})$ and $\boldsymbol{u}_-(\boldsymbol{k})$. From (3.13) we see that the solution can be split into plane waves. Moreover, the real and imaginary part of $\lambda_\pm$ can be interpreted as the amplification and frequency of the waves, respectively.

By using (3.2) and (3.9), we can calculate the eigenvalues of the H-matrix as

$$\lambda(\boldsymbol{k})_\pm = \frac{1}{2\varepsilon}\left[-1 - i\varepsilon \sum k^{(j)} \left({a_{11}}^{(j)} + {a_{22}}^{(j)}\right) \pm \sqrt{1 - 4\varepsilon^2\gamma(\boldsymbol{k}) - i4\varepsilon^2\beta(\boldsymbol{k})}\right], \tag{3.14}$$

where we have introduced the shorthands

$$\begin{aligned} \gamma(\boldsymbol{k}) \equiv \frac{1}{4}\left(\sum k^{(j)}\left({a_{11}}^{(j)} + {a_{22}}^{(j)}\right)\right)^2 - \left(\sum k^{(j)}{a_{11}}^{(j)}\right)\left(\sum k^{(j)}{a_{22}}^{(j)}\right) \\ + \left(\sum k^{(j)}{a_{12}}^{(j)}\right)\left(\sum k^{(j)}{a_{21}}^{(j)}\right) \end{aligned} \tag{3.15}$$

and

$$\beta(\boldsymbol{k}) \equiv \sum k^{(j)}\left({a_{11}}^{(j)} + {a_{12}}^{(j)} r_{21} - \frac{1}{2}\left({a_{11}}^{(j)} + {a_{22}}^{(j)}\right)\right). \tag{3.16}$$

Remark 3.1. *Since γ is the discriminant of the eigenvalues of the Jacobian of the relaxation system, the assumption of hyperbolicity (Assumption 2.1) now takes the very simple form*

$$\gamma(\boldsymbol{k}) \geq 0 \quad \forall \boldsymbol{k} \in \mathbb{R}^d. \tag{3.17}$$

In order to calculate the real and complex part of the eigenvalues (3.14), we first need to calculate the square root of a complex number.

Lemma 3.1. *If a and b are real numbers and $b \neq 0$, then*

$$\sqrt{a+ib} = \frac{1}{\sqrt{2}}\sqrt{\sqrt{a^2+b^2}+a} + i\frac{\operatorname{sgn}(b)}{\sqrt{2}}\sqrt{\sqrt{a^2+b^2}-a}. \tag{3.18}$$

Proof. Let

$$\sqrt{a+ib} = c+id. \tag{3.19}$$

Squaring both sides then yields two equations:

$$a = c^2 + d^2 \tag{3.20a}$$

$$b = 2cd. \tag{3.20b}$$

Eliminating d from (3.20a)–(3.20b) gives the quadratic equation in c^2

$$4c^4 - 4ac^2 - b^2 = 0, \tag{3.21}$$

with positive solution

$$c = \frac{1}{\sqrt{2}}\sqrt{\sqrt{a^2+b^2}+a}. \tag{3.22}$$

Finally, equation (3.20b) then gives

$$\begin{aligned} d = \frac{b}{2c} &= \frac{b}{\sqrt{2}\sqrt{\sqrt{a^2+b^2}+a}} \\ &= \frac{b}{\sqrt{2}}\frac{\sqrt{\sqrt{a^2+b^2}-a}}{\sqrt{b^2}} \\ &= \frac{\operatorname{sgn}(b)}{\sqrt{2}}\sqrt{\sqrt{a^2+b^2}-a}. \end{aligned} \tag{3.23}$$

□

Using Lemma 3.1, the real and complex part of the eigenvalues of the H-matrix can be straightforwardly calculated as

$$\begin{aligned}\mathrm{Re}\lambda_{\pm} = \frac{1}{2\varepsilon}\Bigg[&-1\\ &\pm \frac{1}{\sqrt{2}}\left(\left((1-4\varepsilon^2\gamma(\boldsymbol{k}))^2+16\xi^2\beta(\boldsymbol{k})^2\right)^{1/2}+1-4\xi^2\gamma(\boldsymbol{k})\right)^{1/2}\Bigg]\end{aligned} \tag{3.24}$$

and

$$\begin{aligned}\mathrm{Im}\lambda_{\pm} = -\frac{1}{2\varepsilon}\Bigg[&\varepsilon\sum k^{(j)}({a_{11}}^{(j)}+{a_{22}}^{(j)})\\ &\pm \frac{\mathrm{sgn}(\beta(\boldsymbol{k}))}{\sqrt{2}}\left(\left((1-4\varepsilon^2\gamma(\boldsymbol{k})^2+16\varepsilon^2\beta(\boldsymbol{k})^2\right)^{1/2}-1+4\varepsilon^2\gamma(\boldsymbol{k})\right)^{1/2}\Bigg],\end{aligned} \tag{3.25}$$

respectively.

3.2 The Local Equilibrium Approximation

The local equilibrium approximation is characterized by fulfilling the equilibrium condition

$$R\boldsymbol{u} = 0. \tag{3.26}$$

If we denote $\boldsymbol{u} = [u_1, u_2]^T$ and use the assumed form of the relaxation matrix (3.9), the equilibrium condition is given explicitly as

$$u_2 = r_{21}u_1. \tag{3.27}$$

By inserting (3.27) into the system (3.1) we obtain a conservation law in the reduced variable u_1, given by

$$\partial_t u_1 + \sum_{j=1}^{d}\left({a_{11}}^{(j)}+{a_{12}}^{(j)}r_{21}\right)\partial_{x_j}u_1 = 0. \tag{3.28}$$

The Jacobian of the reduced system is in this case a scalar, and its value—and eigenvalue—is given by

$$\mu^{\mathrm{eq}}(\boldsymbol{k}) = \sum_{j=1}^{d} k^{(j)}\left({a_{11}}^{(j)}+{a_{12}}^{(j)}r_{21}\right). \tag{3.29}$$

3.3 The Sub-Characteristic Condition

As discussed in Section 2.4, the sub-characteristic condition requires that the characteristics of the local equilibrium approximation are interlaced in the characteristics of the homogeneous relaxation system. For the 2×2 case where the rank of the relaxation matrix is 1, the equilibrium approximation is a scalar conservation law (3.28) with a single characteristic (3.29). The relaxation model on the other hand, has two characteristics (3.3). The sub-characteristic condition thus takes the form

$$\mu(\boldsymbol{k})_- \leq \mu^{\mathrm{eq}}(\boldsymbol{k}) \leq \mu(\boldsymbol{k})_+ \quad \forall \boldsymbol{k} \in \mathbb{R}^d. \tag{3.30}$$

By rearranging and using the shorthands (3.15) and (3.16), we can rewrite (3.30) into the single inequality

$$\gamma(\boldsymbol{k}) - \beta(\boldsymbol{k})^2 \geq 0 \quad \forall \boldsymbol{k} \in \mathbb{R}^d. \tag{3.31}$$

3.3.1 The Sub-Characteristic Condition and Linear Stability

As previously discussed, the real parts of the eigenvalues of the H-matrix represent the amplification of the plane waves of the general solution. For the linear stability of the solution, we must therefore require

$$\mathrm{Re}\lambda_\pm \leq 0. \tag{3.32}$$

Proposition 3.1. *For the linear 2×2 system the sub-characteristic condition is necessary and sufficient for the linear stability of the general solution.*

Proof. Inserting (3.24) into (3.32) yields

$$\left(\left(1 - 4\varepsilon^2\gamma(\boldsymbol{k})\right)^2 + 16\varepsilon^2\beta(\boldsymbol{k})^2\right)^{1/2} + 1 - 4\varepsilon^2\gamma(\boldsymbol{k}) \leq 2. \tag{3.33}$$

Rearranging and squaring yields

$$\left(1 - 4\varepsilon^2\gamma(\boldsymbol{k})\right)^2 + 16\varepsilon^2\beta(\boldsymbol{k})^2 \leq \left(1 + 4\varepsilon^2\gamma(\boldsymbol{k})\right)^2. \tag{3.34}$$

Furthermore, by canceling terms and rearranging, (3.34) can be simplified to

$$\gamma(\boldsymbol{k}) - \beta(\boldsymbol{k})^2 \geq 0, \tag{3.35}$$

which is the sub-characteristic condition (3.31). □

This result was shown by Yong [48] for linearized systems in the form (3.1), and later commented by Barker [2] in his investigations of the stability of the St. Venant equations.

One reason why this result is particularly interesting is that there seems to be a fundamental connection between the physical principle of causality and the mathematical stability of the general solution.

3.4 A Chapman–Enskog Expansion

We now venture to derive the Chapman–Enskog approximation for the linear 2×2 relaxation system (3.1). The relaxation system consists of the two equations

$$\partial_t u_1 + \sum_{j=1}^{d} \partial_{x_j} \left({a_{11}}^{(j)} u_1 + {a_{12}}^{(j)} u_2 \right) = 0 \tag{3.36a}$$

$$\partial_t u_2 + \sum_{j=1}^{d} \partial_{x_j} \left({a_{21}}^{(j)} u_1 + {a_{22}}^{(j)} u_2 \right) = \frac{1}{\varepsilon}(r_{21} u_1 - u_2). \tag{3.36b}$$

We can now seek an approximation to this system, valid for small but finite relaxation times ε. As described in Section 2.5, the basic idea is to expand the relaxed variable u_2 in powers of ε around its equilibrium value as

$$u_2 = r_{21} u_1 + \varepsilon u_2^{(1)} + \varepsilon^2 u_2^{(2)} + \mathcal{O}(\varepsilon^3). \tag{3.37}$$

Truncating the expansion (3.37) at first order in ε and inserting it into (3.36b) yields

$$\begin{aligned}
u_2^{(1)} &= -\partial_t u_2 - \sum_{j=1}^{d} \partial_{x_j} \left(a_{21}{}^{(j)} u_1 + a_{22}{}^{(j)} u_2 \right) + \mathcal{O}(\varepsilon) \\
&= -r_{21}\partial_t u_1 - \sum_{j=1}^{d} \partial_{x_j} \left(a_{21}{}^{(j)} u_1 + a_{22}{}^{(j)} r_{21} u_1 \right) + \mathcal{O}(\varepsilon) \\
&= r_{21} \sum_{j=1}^{d} \partial_{x_j} \left(a_{11}{}^{(j)} u_1 + a_{12}{}^{(j)} r_{21} u_1 \right) \\
&\qquad\qquad - \sum_{j=1}^{d} \partial_{x_j} \left(a_{21}{}^{(j)} u_1 + a_{22}{}^{(j)} r_{21} u_1 \right) + \mathcal{O}(\varepsilon) \\
&= -\sum_{j=1}^{d} \left(a_{21}{}^{(j)} + a_{22}{}^{(j)} r_{21} - a_{11}{}^{(j)} r_{21} - a_{12}{}^{(j)} r_{21}^2 \right) \partial_{x_j} u_1 + \mathcal{O}(\varepsilon),
\end{aligned} \tag{3.38}$$

where terms of order $\mathcal{O}(\varepsilon)$ have been ignored. With the approximation (3.38), u_2 is given by

$$u_2 = r_{21} u_1 - \varepsilon \sum_{j=1}^{d} \left(a_{21}{}^{(j)} + a_{22}{}^{(j)} r_{21} - a_{11}{}^{(j)} r_{21} - a_{12}{}^{(j)} r_{21}^2 \right) \partial_{x_j} u_1 + \mathcal{O}(\varepsilon^2). \tag{3.39}$$

Finally, inserting (3.39) into (3.36a) yields the advection-diffusion equation

$$\partial_t u_1 + \sum_{j=1}^{d} \left(a_{11}{}^{(j)} + a_{12}{}^{(j)} r_{21} \right) \partial_{x_j} u_1 = \varepsilon \sum_{i,j=1}^{d} D_{ij} \partial_{x_i} \partial_{x_j} u_1, \tag{3.40}$$

where the $d \times d$ tensor D_{ij} is given by

$$D_{ij} = a_{12}{}^{(i)} \left(a_{21}{}^{(j)} + a_{22}{}^{(j)} r_{21} - a_{11}{}^{(j)} r_{21} - a_{12}{}^{(j)} r_{21}^2 \right). \tag{3.41}$$

Note that in the limit $\varepsilon \to 0$, the Chapman–Enskog approximation (3.40) is equivalent to the local equilibrium approximation (3.28), as is to be expected.

3.5 The Sub-Characteristic Condition and Dissipativity

As has been pointed out by Chen et al. [8], it is not immediately obvious that the Chapman–Enskog approximation (3.40) is *dissipative*.

Before the concept of dissipativity of the Chapman–Enskog approximation can be discussed, we need to establish some preliminary results.

Proposition 3.2. *For the symmetric matrix*

$$S = \frac{1}{2}\left(D + D^T\right) \tag{3.42}$$

we have

$$\sum_{i,j=1}^{d} D_{ij}\partial_{x_i}\partial_{x_j}u_1 = \sum_{i,j=1}^{d} S_{ij}\partial_{x_i}\partial_{x_j}u_1 \tag{3.43}$$

Proof. The proof follows directly from the commutative property of the partial derivative:

$$\partial_{x_i}\partial_{x_j} = \partial_{x_j}\partial_{x_i}. \tag{3.44}$$

□

In other words, we can consider the symmetric counterpart of D without changing the partial differential equation. This is particularly useful because of the following standard result in linear algebra:

Lemma 3.2. *A real symmetric matrix S can be diagonalized as*

$$S = Q^T\Lambda Q \tag{3.45}$$

where Q is an orthogonal matrix, i.e. a matrix that fulfills $Q^{-1} = Q^T$.

The dissipativity of a diffusion term can now be defined in terms of the symmetric matrix S:

Definition 4 (Dissipativity)**.** *The diffusion equation*

$$\partial_t u = \sum_{i,j=1}^{d} S_{ij}\partial_{x_i}\partial_{x_j}u \tag{3.46}$$

is said to be **dissipative** *if the symmetric matrix S can be diagonalized as*

$$S = Q^T \Lambda Q, \tag{3.47}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix} \tag{3.48}$$

with

$$\lambda_i \geq 0 \quad \forall i \in \{1, \dots, d\}. \tag{3.49}$$

The orthogonal matrix Q represents a change of variables to a set of variables corresponding to the *principal axes* of the diffusion. The positivity of the eigenvalues $\{\lambda_i\}$ then ensures that—along these principal axes—the equation describes diffusion rather than *anti-diffusion*. Also, since the matrix S is symmetric, the dissipativity is equivalent with the matrix being *positive-definite*.

3.5.1 The 1-Dimensional Case

The simplest case to consider is that of $d = 1$. In this case, the sub-characteristic condition can be written as

$$\gamma - \beta^2 \geq 0, \tag{3.50}$$

where

$$\gamma \equiv \gamma(1) = \frac{1}{4}(a_{11} + a_{22})^2 - a_{11}a_{22} + a_{12}a_{21}, \tag{3.51}$$

and

$$\beta \equiv \beta(1) = a_{11} + a_{12}r_{21} - \frac{1}{2}(a_{11} + a_{22}). \tag{3.52}$$

Moreover, in the 1-dimensional case, the diffusion-tensor is given by

$$D = a_{12}\left(a_{21} + a_{22}r_{21} - a_{11}r_{21} - a_{12}r_{21}^2\right). \tag{3.53}$$

Since the diffusion-tensor is a scalar, dissipativity simply requires $D \geq 0$. The following result was shown by Chen et al. [8] to be valid also in the non-linear case.

Proposition 3.3. *For linear 2×2 systems in one spatial dimension, the sub-characteristic condition is sufficient and necessary for the dissipativity of the diffusive term of the Chapman–Enskog approximation.*

Proof. The Chapman–Enskog approximation (3.40) is dissipative if and only if

$$a_{12}a_{21} + a_{12}a_{22}r_{21} - a_{12}a_{11}r_{21} - (a_{12}r_{21})^2 \geq 0. \tag{3.54}$$

Rearranging (3.54) and adding $(1/4)(a_{11}+a_{22})^2$ to both sides of the inequality yields

$$2a_{11}a_{12}r_{21} - a_{12}r_{21}(a_{11}+a_{22}) + (a_{12}r_{21})^2 + \left(\frac{1}{2}(a_{11}+a_{22})\right)^2 \leq \frac{1}{4}(a_{11}+a_{22})^2 + a_{12}a_{21}. \tag{3.55}$$

Subtracting $a_{11}a_{22}$ from both sides of the inequality gives

$$a_{11}^2 - a_{12}r_{21}(a_{11}+a_{22}) - a_{11}(a_{11}+a_{22}) + 2a_{11}a_{12}r_{21} + (a_{12}r_{21})^2 + \left(\frac{1}{2}(a_{11}+a_{22})\right)^2 \leq \frac{1}{4}(a_{11}+a_{22})^2 - a_{11}a_{22} + a_{12}a_{21}. \tag{3.56}$$

Finally, by recognizing the square, we can write (3.56) as

$$\left(a_{11} + a_{12}r_{21} - \frac{1}{2}(a_{11}+a_{22})\right)^2 \leq \frac{1}{4}(a_{11}+a_{22})^2 - a_{11}a_{22} + a_{12}a_{21}, \tag{3.57}$$

which is exactly the sub-characteristic condition (3.50). □

3.5.2 The 2-Dimensional Case

We will now attempt to extend this analysis in order to investigate if the connection between dissipativity and the sub-characteristic condition holds beyond the simple 1-D case.

Polar Form of the Sub-Characteristic Condition

In two spatial dimensions, the wave-number $\boldsymbol{k}$ can in general be written in polar coordinates as

$$\boldsymbol{k} = |\boldsymbol{k}|\hat{\boldsymbol{k}} = |\boldsymbol{k}|\left(\cos\theta\,\hat{x} + \sin\theta\,\hat{y}\right), \tag{3.58}$$

where $\hat{x}$ and $\hat{y}$ are unit vectors along their respective spatial dimensions. The radius $|\boldsymbol{k}|$ can then be factored out of the sub-characteristic condition (3.31), yielding a *polar form*

$$\gamma(\theta) - \beta(\theta)^2 \geq 0 \quad \forall\theta \in [0, 2\pi], \tag{3.59}$$

where

$$\begin{aligned}\gamma(\theta) \equiv \frac{1}{|\boldsymbol{k}|^2}\gamma(\boldsymbol{k}) = \frac{1}{4}\left(\cos\theta\left({a_{11}}^{(x)} + {a_{22}}^{(x)}\right) + \sin\theta\left({a_{11}}^{(y)} + {a_{22}}^{(y)}\right)\right)^2 \\ - \left(\cos\theta\, {a_{11}}^{(x)} + \sin\theta\, {a_{11}}^{(y)}\right)\left(\cos\theta\, {a_{22}}^{(x)} + \sin\theta\, {a_{22}}^{(y)}\right) \\ + \left(\cos\theta\, {a_{12}}^{(x)} + \sin\theta\, {a_{12}}^{(y)}\right)\left(\cos\theta\, {a_{21}}^{(x)} + \sin\theta\, {a_{21}}^{(y)}\right)\end{aligned} \tag{3.60}$$

and

$$\begin{aligned}\beta(\theta) \equiv \frac{1}{|\boldsymbol{k}|}\beta(\boldsymbol{k}) = \cos\theta\left({a_{11}}^{(x)} + {a_{12}}^{(x)} r_{21} - \frac{1}{2}\left({a_{11}}^{(x)} + {a_{22}}^{(x)}\right)\right) \\ + \sin\theta\left({a_{11}}^{(y)} + {a_{12}}^{(y)} r_{21} - \frac{1}{2}\left({a_{11}}^{(y)} + {a_{22}}^{(y)}\right)\right).\end{aligned} \tag{3.61}$$

The right-hand side of the sub-characteristic condition can then be written explicitly as

$$\begin{aligned}\gamma(\theta) - \beta(\theta)^2 = \cos^2\theta\left[{a_{12}}^{(x)}\left({a_{21}}^{(x)} + {a_{22}}^{(x)} r_{21} - {a_{11}}^{(x)} r_{21} - {a_{12}}^{(x)} r_{21}^2\right)\right] \\ + \sin^2\theta\left[{a_{12}}^{(y)}\left({a_{21}}^{(y)} + {a_{22}}^{(y)} r_{21} - {a_{11}}^{(y)} r_{21} - {a_{12}}^{(y)} r_{21}^2\right)\right] \\ + \cos\theta\,\sin\theta\Big[{a_{12}}^{(x)}\left({a_{21}}^{(y)} + {a_{22}}^{(y)} r_{21} - {a_{11}}^{(y)} r_{21} - {a_{12}}^{(y)} r_{21}^2\right) \\ + {a_{12}}^{(y)}\left({a_{21}}^{(x)} + {a_{22}}^{(x)} r_{21} - {a_{11}}^{(x)} r_{21} - {a_{12}}^{(x)} r_{21}^2\right)\Big].\end{aligned} \tag{3.62}$$

Furthermore, if we introduce the shorthand

$$\kappa^{(i)} = {a_{21}}^{(i)} + {a_{22}}^{(i)} r_{21} - {a_{11}}^{(i)} r_{21} - {a_{12}}^{(i)} r_{21}^2, \tag{3.63}$$

the expression (3.62) can be simplified, and the polar form of the sub-characteristic condition takes the form

$$\begin{aligned}\gamma(\theta) - \beta(\theta)^2 = \cos^2\theta\left[{a_{12}}^{(x)}\kappa^{(x)}\right] + \sin^2\theta\left[{a_{12}}^{(y)}\kappa^{(y)}\right] \\ + 2\cos\theta\,\sin\theta\left[\frac{1}{2}\left({a_{12}}^{(x)}\kappa^{(y)} + {a_{12}}^{(y)}\kappa^{(x)}\right)\right] \geq 0.\end{aligned} \tag{3.64}$$

Dissipativity

It has been established that a real symmetric matrix can be diagonalized by an orthogonal matrix. In 2 dimensions, an orthogonal matrix Q is a matrix

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}, \tag{3.65}$$

where orthogonality requires

$$q_{11}q_{12} + q_{21}q_{22} = 0, \tag{3.66}$$

with the normalization

$$q_{11}^2 + q_{21}^2 = 1 \quad \text{and} \quad q_{12}^2 + q_{22}^2 = 1. \tag{3.67}$$

Without loss of generality, we can let $q_{11} = \cos\phi$ and $q_{21} = \sin\phi$ for some ϕ. There are now two choices that fulfill the rest of the constraints: Either $q_{12} = -q_{21}$ and $q_{22} = q_{11}$ or $q_{12} = q_{21}$ and $q_{22} = -q_{11}$. The first choice leads to a rotation matrix

$$Q = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \tag{3.68}$$

The second choice leads to a reflection matrix

$$Q = \begin{bmatrix} \cos\phi & \sin\phi \\ \sin\phi & -\cos\phi \end{bmatrix}, \tag{3.69}$$

where the angle ϕ defines the line of reflection.

Proposition 3.4. *A 2×2 real symmetric matrix S can be diagonalized as*

$$S = Q^T \Lambda Q \tag{3.70}$$

where Q is an orthogonal rotation matrix.

Proof. Lemma 3.2 states that S can be diagonalized by an orthogonal matrix. Any orthogonal 2×2 matrix is either a rotation matrix or a reflection matrix. Let Q be a reflection matrix, then it is necessarily its own inverse: $QQ = I$. If S is a non-diagonal real symmetric matrix that is diagonalized with a reflection Q, then

$$S = Q^T \Lambda Q = \Lambda (Q^T)^T Q = \Lambda Q Q = \Lambda, \tag{3.71}$$

which is a contradiction. An orthogonal rotation matrix Q is then the only remaining choice for diagonalization. □

The symmetric analog (3.42) of the diffusion matrix (3.41) is given explicitly as

$$S = \begin{bmatrix} {a_{12}}^{(x)} \kappa^{(x)} & \frac{1}{2}\left({a_{12}}^{(x)} \kappa^{(y)} + {a_{12}}^{(y)} \kappa^{(x)}\right) \\ \frac{1}{2}\left({a_{12}}^{(x)} \kappa^{(y)} + {a_{12}}^{(y)} \kappa^{(x)}\right) & {a_{12}}^{(y)} \kappa^{(y)} \end{bmatrix}. \tag{3.72}$$

Proposition 3.5. *For linear 2×2 systems in two spatial dimensions, the sub-characteristic condition is sufficient and necessary for the dissipativity of the diffusion term of the Chapman–Enskog approximation.*

Proof. For any orthogonal change of variables $\boldsymbol{x}' = Q\boldsymbol{x}$, the symmetric diffusion-matrix S undergoes the similarity transform $S' = Q^T S Q$. By a direct calculation, using (3.68) and (3.72), the diagonal elements of S' are given by

$$S'_{11} = \cos^2\phi \left[{a_{12}}^{(x)} \kappa^{(x)}\right] + \sin^2\phi \left[{a_{12}}^{(y)} \kappa^{(y)}\right] + 2\,\cos\phi \sin\phi \left[\frac{1}{2}\left({a_{12}}^{(x)} \kappa^{(y)} + {a_{12}}^{(y)} \kappa^{(x)}\right)\right] \tag{3.73}$$

$$S'_{22} = \sin^2\phi \left[{a_{12}}^{(x)} \kappa^{(x)}\right] + \cos^2\phi \left[{a_{12}}^{(y)} \kappa^{(y)}\right] - 2\,\cos\phi \sin\phi \left[\frac{1}{2}\left({a_{12}}^{(x)} \kappa^{(y)} + {a_{12}}^{(y)} \kappa^{(x)}\right)\right]. \tag{3.74}$$

By comparing these to the sub-characteristic condition (3.64), we see that the case $\theta = \phi$ ensures the positivity of (3.73). Furthermore, by using the identities

$$\sin\left(\phi + \frac{\pi}{2}\right) = \cos\phi, \quad \text{and} \quad \cos\left(\phi + \frac{\pi}{2}\right) = -\sin\phi, \tag{3.75}$$

it is clear that the case $\theta = \phi + \pi/2$ ensures the positivity of (3.74). Since this holds for all rotation-matrices $R(\phi)$, then it must also hold for the rotation corresponding to the diagonalization of S; hence we have shown the sufficiency part of the proposition.

Furthermore, since positive diagonal elements are a necessary condition for a positive-definite matrix, and any similarity-transform conserves the positive-definite property of a matrix, the sub-characteristic condition is also necessary for dissipativity. □

3.6 Summary

This chapter has been devoted to linear 2×2 systems in d spatial dimensions. The content of this chapter has been a mix of well-known results and present contributions.

In Section 3.1 the structure of the general 2×2 system was discussed given some conditions on the relaxation term. Moreover, the general solution was shown to consist of plane waves, with wave-speeds and amplifications depending on the eigenvalues of the H-matrix (3.11).

The relationship between the sub-characteristic condition and the stability of the relaxation system was discussed in detail. In Proposition 3.1, the linear stability of the general linear 2×2 systems was shown to be equivalent to the sub-characteristic condition—a result known from literature [8, 48, 2].

Section 3.5 was devoted to the relationship between the sub-characteristic condition and the dissipativity of the Chapman–Enskog approximation. As shown by Chen et al. [8] in the non-linear case, these are equivalent in 1-D (Proposition 3.3). This was then shown to be true also in the 2-D case (Proposition 3.5) by constructing a polar form of the Chapman–Enskog approximation and looking at the diffusion tensor in a rotated coordinate system.

4 Wave-Dynamics for Linear 2 × 2 Hyperbolic Relaxation Systems

The plane wave decomposition of the general solution (3.10) of the linearized 2×2 system motivates a more detailed study of the wave-dynamics of such relaxation systems. Specifically, there are at least two questions that warrants further investigation:

- How does the relaxation term influence the wave-dynamics of the relaxation system?
- How does the wave-dynamics of the relaxation system relate to that of the equilibrium system?

By a direct analysis of the eigenvalues of the H-matrix (3.11), these questions and others will be addressed in this chapter.

4.1 The 1-Dimensional Case

For simplicity, only the 1-dimensional case will be considered. The 2×2 system then becomes

$$\partial_t \boldsymbol{u} + A\,\partial_x \boldsymbol{u} = \frac{1}{\varepsilon} R\boldsymbol{u}, \tag{4.1}$$

where $\boldsymbol{u} = \boldsymbol{u}(x,t)$ is a 2-vector and

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \text{and} \quad R = \begin{bmatrix} 0 & 0 \\ r_{21} & -1 \end{bmatrix}. \tag{4.2}$$

The eigenvalues—or characteristic speeds—of the A-matrix are given by

$$\mu_\pm = \frac{1}{2}(a_{11} + a_{22}) \pm \sqrt{\frac{1}{4}(a_{11} + a_{22})^2 - a_{11}a_{22} + a_{12}a_{21}}. \tag{4.3}$$

As discussed in Section 2.1, the eigenvalues (4.3) are the wave-speeds of the homogeneous relaxation system.

4.1.1 The Local Equilibrium Approximation

In the 1-dimensional case, the local equilibrium approximation (3.27)–(3.28) simplifies to

$$u_2 = u_1 r_{21} \tag{4.4}$$

and the corresponding advection equation

$$\partial_t u_1 + (a_{11} + a_{12} r_{21}) \partial_x u_1 = 0. \tag{4.5}$$

The characteristic speed of the equilibrium model is therefore given by $a_{11} + a_{12} r_{21}$.

4.1.2 Eigenvalues of the H-matrix

As discussed in Section 3.1.2, the eigenvalues of the H-matrix (3.11) are of special significance. In the 1-D case, the eigenvalues of the H-matrix are most conveniently given by

$$\lambda_\pm(k) = \frac{k}{2\xi} \left[-1 - i\xi \left(a_{11} + a_{22} \right) \pm \left(1 - 4\xi^2 \gamma - i4\xi^2 \beta \right)^{1/2} \right], \tag{4.6}$$

where $\xi = k\varepsilon$ and

$$\gamma \equiv \gamma(1) = \frac{1}{4} \left(a_{11} + a_{22} \right)^2 - a_{11} a_{22} + a_{12} a_{21} \tag{4.7}$$

and

$$\beta \equiv \beta(1) = a_{11} + a_{12} r_{21} - \frac{1}{2} \left(a_{11} + a_{22} \right). \tag{4.8}$$

4.1.3 Interpretation of γ and β

One of the benefits of doing analysis in the 1-dimensional case, is that the numbers γ and β both have a simple interpretation. First, we make the observation

$$\begin{aligned} \beta &= a_{11} + a_{12} r_{21} - \frac{1}{2} \left(a_{11} + a_{22} \right) \\ &= - \left(a_{22} - a_{12} r_{21} - \frac{1}{2} \left(a_{11} + a_{22} \right) \right). \end{aligned} \tag{4.9}$$

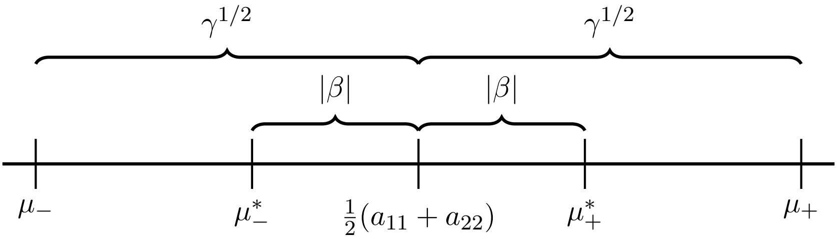


Figure 4.1: In the 1-D case the numbers γ and β have a simple interpretation: $\gamma^{1/2}$ is the radius of the hyperbolic speeds and $|\beta|$ can be seen as the radius of the equilibrium speed, both with regard to the mean value $(1/2)(a_{11} + a_{22})$.

Now, by defining

$$\mu_+^* \equiv \max\{a_{11} + a_{12}r_{21}, a_{22} - a_{12}r_{21}\} \tag{4.10a}$$
$$\mu_-^* \equiv \min\{a_{11} + a_{12}r_{21}, a_{22} - a_{12}r_{21}\}, \tag{4.10b}$$

the value of $|\beta|$ can be seen as the radius, with respect to the root center $(1/2)(a_{11}+a_{22})$, of the pair composed of the equilibrium speed $a_{11}+a_{12}r_{21}$ and a *mirror-speed* $a_{22} - a_{12}r_{21}$, see Figure 4.1. The number γ is the discriminant of the wave-speeds of the homogeneous relaxation system (4.3), thus $\gamma^{1/2}$ is the radius of these speeds with respect to the same root center.

The concept of the mirror-image of the equilibrium speed will also have significance later in this chapter.

4.2 An Example Model

The analysis of the wave-dynamics performed in this chapter is valid for general 2×2 systems in the form (4.1)–(4.2). However, in order to visualize and interpret various analytical results, a basic model will be used as an example.

The basic example will be the model

$$\partial_t u + \partial_x v = 0 \tag{4.11a}$$

$$\partial_t v + \lambda_R^2 \partial_x u = \frac{1}{\varepsilon}(\lambda_E u - v). \tag{4.11b}$$

The equations (4.11a)–(4.11b) represent a classical 2 × 2 example model used by Natalini [36]. This model is useful because it is a simple 2 × 2 model, but still with a non-trivial coupling between the equations. The idea is that this simple model will contain much of the same complexity as the general model.

In the context of the general 1-D model, the A-matrix of the example-model is given by

$$A = \begin{bmatrix} 0 & 1 \\ \lambda_R^2 & 0 \end{bmatrix} \tag{4.12}$$

with eigenvalues

$$\mu_\pm = \pm \lambda_R. \tag{4.13}$$

The local equilibrium approximation is given by

$$v = \lambda_E u \tag{4.14}$$

together with the advection equation

$$\partial_t u + \lambda_E \partial_x u = 0. \tag{4.15}$$

The two fixed parameters of the model, λ_R and λ_E, are the wave-speeds of the homogeneous relaxation model and the equilibrium approximation, respectively. The sub-characteristic condition is therefore in this case given by

$$\lambda_R^2 \geq \lambda_E^2. \tag{4.16}$$

Unless otherwise specified, the parameters used in examples will be $\lambda_R = 1.0$ and $\lambda_E = 0.2$.

4.3 Amplification and Wave-Speeds

As pointed out in Section 3.1.2, the real and imaginary part of the H-matrix can be interpreted as the amplification and frequency of the Fourier-component with wave-number k, respectively. These eigenvalues therefore contain a lot of information regarding the wave-dynamics of the general solution. In particular, given the frequencies $\omega_\pm(k)$ of the Fourier-component with wave-number k, the wave-speeds $v_\pm(k)$ can be calculated using the standard relation

$$v_\pm(k) = \frac{1}{k}\omega_\pm(k) = -\frac{1}{k}\mathrm{Im}\lambda_\pm. \tag{4.17}$$

The application of Lemma 3.1 on page 17 yields the real and imaginary part of the eigenvalues (4.6) of the H-matrix as

$$\mathrm{Re}\lambda_\pm = \frac{k}{2\xi}\left[-1 \pm \frac{1}{\sqrt{2}}\left(\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2} + 1 - 4\xi^2\gamma\right)^{1/2}\right] \tag{4.18}$$

and

$$\begin{aligned}\mathrm{Im}\lambda_\pm = -\frac{k}{2\xi}\Bigg[\xi(a_{11}+a_{22}) \pm \frac{\mathrm{sgn}(\beta)}{\sqrt{2}}\Bigg(&\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2} \\ &- 1 + 4\xi^2\gamma\Bigg)^{1/2}\Bigg],\end{aligned} \tag{4.19}$$

respectively.

By comparing (4.17) to the imaginary part of the eigenvalues (4.19), it becomes clear that the wave-speeds are given by

$$v_\pm = v_\pm(\xi). \tag{4.20}$$

In other words, the wave-speed of the k'th Fourier component only depends on the variable $\xi = k\varepsilon$. For this reason, there will always be a duality in the interpretation of the functional behavior of the wave-speeds. For instance, the *stiff* limit ($\xi \to 0$), can be seen as both the equilibrium limit ($\varepsilon \to 0$) and low wave-number limit ($k \to 0$); as far as the wave-speeds are concerned, these limits are the same.

4.3.1 The Stiff Limit

Proposition 4.1. *In the* **stiff** *limit, given by*

$$\xi \to 0, \tag{4.21}$$

the two eigenvalues of the H-matrix are:

λ_+ *Purely imaginary corresponding to the undamped equilibrium characteristic.*

λ_- *Complex with imaginary part equal to the mirror-speed of the equilibrium speed and with* $\mathrm{Re}\lambda_- \to -\infty$ *as* $\xi \to 0$.

Proof. By introducing the shorthand

$$\Theta(\xi) \equiv \left(\left(1 - 4\xi^2\gamma\right)^2 + 16\xi^2\beta^2\right)^{1/2} + 4\xi^2\gamma - 1, \tag{4.22}$$

we can write the amplification (4.18) as

$$\mathrm{Re}\lambda_\pm = \frac{k}{2\xi}\left[-1 \pm \left(\frac{\Theta(\xi)}{2} + 1 - 4\xi^2\gamma\right)^{1/2}\right] \tag{4.23}$$

and the dispersion relation (4.19) as

$$\mathrm{Im}\lambda_\pm = -\frac{k}{2}\left[(a_{11} + a_{22}) \pm \mathrm{sgn}(\beta)\left(\frac{\Theta(\xi)}{2\xi^2}\right)^{1/2}\right]. \tag{4.24}$$

Simple inspection of (4.22) reveals that

$$\Theta(\xi) \geq 0, \tag{4.25}$$

with equality only for $\xi = 0$. That is

$$\lim_{\xi\to 0} \Theta(\xi) = 0, \tag{4.26}$$

and, as a consequence, the amplification (4.23) is in the stiff limit given by

$$\lim_{\xi\to 0} \mathrm{Re}\lambda_\pm = \lim_{\xi\to 0} \frac{k}{2\xi}\left(-1 \pm 1\right). \tag{4.27}$$

Thus, in this limit, one of the waves is suppressed and the other is undamped. In order to evaluate (4.24) in the stiff limit, we first calculate

$$\lim_{\xi^2\to 0} \frac{\Theta(\xi)}{2\xi^2} = \lim_{\xi^2\to 0} \frac{1}{2}\frac{\partial^2\Theta(\xi)}{\partial\xi^2} = 4\beta^2, \tag{4.28}$$

where L'Hôpital's rule is applied in the first step. By inserting (4.28) into (4.24) we can now obtain

$$\begin{aligned}\lim_{\xi\to 0}\mathrm{Im}\lambda_{\pm} &= -\,k\left[\frac{1}{2}(a_{11}+a_{22}) \pm \frac{1}{2}\mathrm{sgn}(\beta)|\beta|\right] \\ &= -\,k\left[a_{11}\left(\frac{1}{2}\pm\frac{1}{2}\right) + a_{22}\left(\frac{1}{2}\mp\frac{1}{2}\right) \pm a_{12}r_{21}\right], \qquad (4.29)\end{aligned}$$

and finally the wave-speed

$$\lim_{\xi\to 0} v_{\pm}(\xi) = -\frac{1}{k}\lim_{\xi\to 0}\mathrm{Im}\lambda_{\pm} = a_{11}\left(\frac{1}{2}\pm\frac{1}{2}\right)+a_{22}\left(\frac{1}{2}\mp\frac{1}{2}\right)\pm a_{12}r_{21}. \quad (4.30)$$

Thus, the undamped wave will have the wave-speed of the equilibrium model; the suppressed wave on the other hand will have the mirror-speed $a_{22} - a_{12}r_{21}$. □

Remark 4.1. *The suppressed wave will in the stiff limit have wave-speed* $a_{22} - a_{12}r_{21}$. *As discussed in Section 4.1.3, this wave-speed is the mirror-speed of the equilibrium-speed* $a_{11} + a_{12}r_{21}$ *relative to the root center* $(1/2)(a_{11} + a_{22})$*, see Figure 4.1 on page 31. Therefore, if the sub-characteristic condition*

$$\gamma - \beta^2 \geq 0 \qquad (4.31)$$

is fulfilled for the equilibrium-speed, then it is automatically fulfilled for the mirror-speed.

4.3.2 The Non-Stiff Limit

Proposition 4.2. *In the* **non-stiff** *limit, given by*

$$\xi \to \infty, \qquad (4.32)$$

the eigenvalues $\lambda_{\pm}$ *of the* H*-matrix are purely imaginary with wave-speeds*

$$v_{\pm}(\xi) = \frac{1}{2}(a_{11}+a_{22}) \pm \mathrm{sgn}(\beta)\gamma^{1/2}. \qquad (4.33)$$

Thus, the eigenvalues of H *will correspond to the characteristics* (4.3) *of the homogeneous relaxation system.*

Proof. By slightly rewriting (4.19), we get

$$\mathrm{Im}\lambda_{\pm} = -\frac{k}{2}\left[(a_{11}+a_{22}) \pm \frac{\mathrm{sgn}(\beta)}{\sqrt{2}}\left(\left(\left(\frac{1}{\xi^2}-4\gamma\right)^2 + \frac{16}{\xi^2}\beta^2\right)^{1/2} - \frac{1}{\xi^2}+4\gamma\right)^{1/2}\right]. \quad (4.34)$$

In the non-stiff limit this becomes

$$\lim_{\xi\to\infty} \mathrm{I}m\lambda_{\pm} = -k\left[\frac{1}{2}(a_{11}+a_{22}) \pm \mathrm{sgn}(\beta)\gamma^{1/2}\right]. \quad (4.35)$$

Dividing by $(-k)$ gives the wave-speeds

$$\lim_{\xi\to\infty} v_{\pm}(\xi) = -\frac{1}{k}\lim_{\xi\to\infty}\mathrm{Im}\lambda_{\pm} = \frac{1}{2}(a_{11}+a_{22}) \pm \mathrm{sgn}(\beta)\gamma^{1/2}, \quad (4.36)$$

which are the characteristics of the homogeneous relaxation system. By rewriting the amplification (4.18) in the same way, we get

$$\mathrm{Re}\lambda_{\pm} = \frac{k}{2}\left[-\frac{1}{\xi} \pm \frac{1}{\sqrt{2}}\left(\left(\left(\frac{1}{\xi^2}-4\gamma\right)^2 + \frac{16}{\xi^2}\beta^2\right)^{1/2} + \frac{1}{\xi^2}-4\gamma\right)^{1/2}\right]. \quad (4.37)$$

In the non-stiff limit this becomes

$$\lim_{\xi\to\infty}\mathrm{Re}\lambda_{\pm} = 0. \quad (4.38)$$

□

Thus, in the non-stiff limit, the solutions are the undamped characteristics of the homogeneous relaxation system.

4.3.3 Transitional Behavior

As shown in the previous sections, the limit behavior of the wave-dynamics of the relaxation system is as expected: In the stiff limit, which can be seen as the limit of infinite relaxation speed, the wave-dynamics of the relaxation system coincides with the dynamics of the local equilibrium approximation.

Moreover, in the non-stiff limit, which can be seen as the limit where the relaxation term is negligible, the wave-dynamics coincides with the dynamics of the homogeneous relaxation system.

However, for any practical relaxation system in the form (4.1)–(4.2), the relaxation time ε has a finite value. In this case, the wave-speeds of the different Fourier-components are dependent on the wave-number k. This is a phenomenon often referred to in wave physics as *dispersion*.

Proposition 4.3. *For a* 2×2 *relaxation system of the form* (4.1) – (4.2)*, the transitional Fourier wave-speeds* $v(\xi)_\pm$ *will be* **monotonic** *functions of* ξ*.*

Proof. Using (4.24) we can write

$$\frac{\partial v_\pm(\xi)}{\partial \xi} = \pm \frac{\operatorname{sgn}(\beta)}{4\sqrt{2}} \frac{1}{\xi \Theta^{1/2}} \left[\frac{\partial \Theta}{\partial \xi} - 2\frac{\Theta}{\xi} \right]. \tag{4.39}$$

Furthermore, we can calculate

$$\frac{\partial \Theta}{\partial \xi} = \frac{-16\xi\gamma(1-4\xi^2\gamma) + 32\xi\beta^2}{2\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2}} + 8\xi\gamma, \tag{4.40}$$

which gives

$$\begin{aligned} \frac{\partial \Theta}{\partial \xi} - 2\frac{\Theta}{\xi} =& \frac{-16\xi\gamma(1-4\xi^2\gamma) + 32\xi\beta^2}{2\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2}} + \frac{2}{\xi} - \frac{2}{\xi}\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2} \\ =& \frac{2}{\xi} + \frac{-8\xi^2\gamma(1-4\xi^2\gamma) + 16\xi^2\beta^2 - 2(1-4\xi^2\gamma)^2 - 32\xi^2\beta^2}{\xi\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2}} \\ =& \frac{2}{\xi}\left(1 - \frac{(1-4\xi^2\gamma) + 8\xi^2\beta^2}{\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2}}\right). \end{aligned} \tag{4.41}$$

The absolute value of the second term in (4.41) can be written as

$$\left| \frac{(1-4\xi^2\gamma) + 8\xi^2\beta^2}{\left((1-4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2}} \right| = \sqrt{1 - \frac{64\xi^4\beta^2}{(1-4\xi^2\gamma)^2 + 16\xi^2\beta^2}(\gamma - \beta^2)}. \tag{4.42}$$

Thus, for $\gamma-\beta^2 > 0$, (4.41) will be a positive function. Furthermore, since the second term in (4.41) is positive for $\gamma - \beta < 0$, we can conclude that

$$\operatorname{sgn}\left(\frac{\partial v_\pm(\xi)}{\partial \xi}\right) = \begin{cases} 0 & \text{if} \quad \gamma - \beta^2 = 0 \\ \pm\operatorname{sgn}(\beta) & \text{if} \quad \gamma - \beta^2 > 0 \\ \mp\operatorname{sgn}(\beta) & \text{if} \quad \gamma - \beta^2 < 0 \end{cases} . \tag{4.43}$$

□

As previously discussed, the sub-characteristic condition is a causality-principle restricting the wave-speeds of the local equilibrium approximation to within the characteristics of the homogeneous relaxation system. The wave-dynamics of the relaxation model for a finite ε is dispersive, i.e. the solution does not have a well-defined characteristic speed. However, the same causality-argument should apply for the wave-speeds of the individual Fourier-components:

Proposition 4.4. *For a linear* 2×2 *relaxation system where the sub-characteristic condition is fulfilled, the transitional wave-speeds* $v(\xi)_\pm$ *will satisfy a* **transitional sub-characteristic condition**

$$\mu_- \leq v_\pm(\xi) \leq \mu_+, \tag{4.44a}$$

where $\mu_\pm$ *are the wave-speeds of the homogeneous relaxation-system* (4.3).

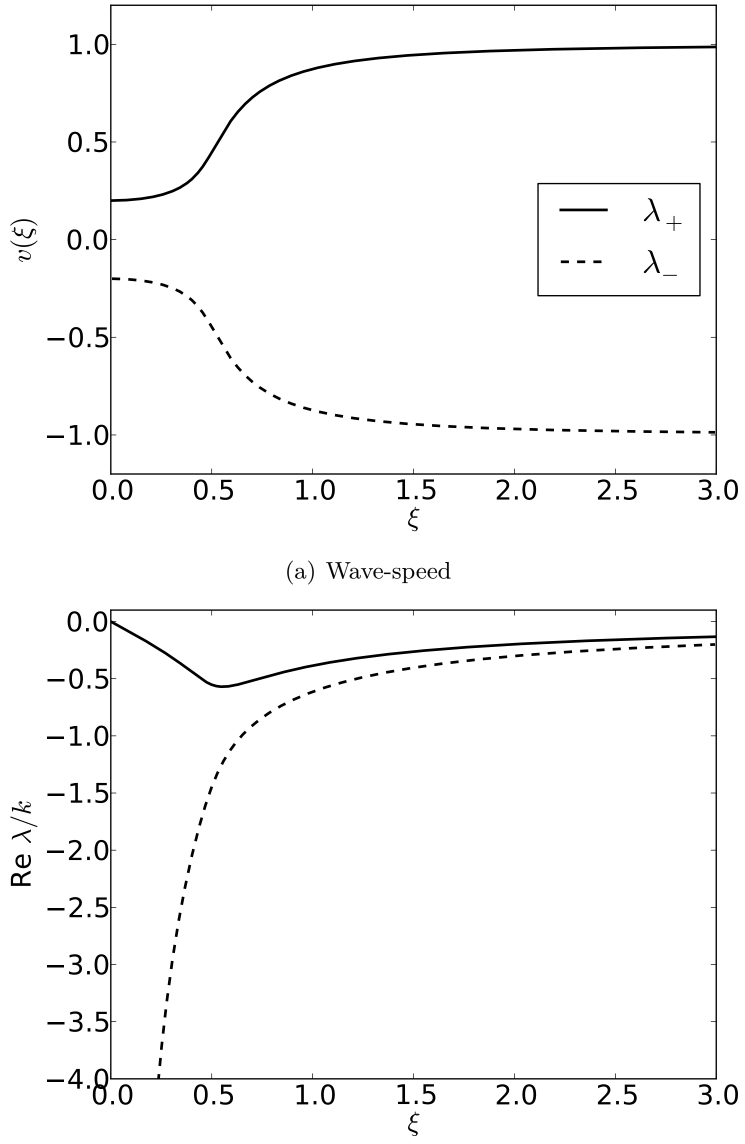
Proof. The proof follows directly from the limit behavior from Proposition 4.1 and Proposition 4.2, and the monotonicity from Proposition 4.3. □

The Example Model

The results of this section can be better illustrated using the example-model introduced in Section 4.2.

Figure 4.2 shows the wave-speeds and amplifications, calculated using (4.19) and (4.18), for the example-model using $\lambda_R = 1.0$ and $\lambda_E = 0.2$. As ξ gets smaller, the wave-speeds change into the equilibrium-speed and the mirror-speed in a monotonic manner, as shown in Proposition 4.3.

Figure 4.2 also indicate that in the non-stiff limit both modes are undamped. In the stiff limit however, the mode corresponding to the equilibrium-characteristic is undamped, while the mode corresponding to the mirror-characteristic is suppressed.



(a) Wave-speed

(b) Amplification

Figure 4.2: The wave-speed and amplification for the two modes of the example model, using $\lambda_R = 1$ and $\lambda_L = 0.2$.

4.4 Validity of the Chapman–Enskog Approximation

In Section 3.4 the Chapman–Enskog expansion was performed for linear 2 × 2 systems, yielding a convection-diffusion-equation (3.40) as an approximation to the full relaxation system—to first order in ε.

In the 1-dimensional case, the Chapman–Enskog approximation takes the form

$$\partial_t u_1 + (a_{11} + a_{12} r_{21})\, \partial_x u_1 = \varepsilon(\gamma - \beta^2) \partial_{xx} u_1, \tag{4.45}$$

and is thus dissipative under the sub-characteristic condition.

In order to put the Chapman–Enskog approximation in context with the wave-analysis of this chapter, we must first derive the general solution to (4.45) in terms of Fourier-components.

Proposition 4.5. *A solution of* (4.45) *can be written in terms of Fourier components*

$$u_1(x,t) = \sum_k \hat{a}(k) \exp\left(ik\left(x - (a_{11} + a_{12} r_{21})t\right)\right) \exp\left(-\varepsilon(\gamma - \beta^2) k^2 t\right). \tag{4.46}$$

Proof. By inserting (4.46) into (4.45) we obtain

$$\begin{aligned} &\partial_t u_1 + (a_{11} + a_{12} r_{21}) \partial_x u_1 - \varepsilon(\gamma - \beta^2) \partial_{xx} u_1 \\ &\quad = \sum_k \hat{a}(k) \Big[-ik(a_{11} + a_{12} r_{21}) - \varepsilon(\gamma - \beta^2) k^2 + ik(a_{11} + a_{12} r_{21}) \\ &\quad + \varepsilon(\gamma - \beta^2) k^2 \Big] \exp\left(ik[x - (a_{11} + a_{12} r_{21})t]\right) \exp\left(-\varepsilon(\gamma - \beta^2) k^2 t\right) = 0. \end{aligned} \tag{4.47}$$

□

As shown in Proposition 4.1, in the stiff limit the plane-wave corresponding to the eigenvalue λ_+ is undamped while the one corresponding to λ_- is suppressed. From the general solution (4.46) we see that the dampening-factor has a k^2-dependence. Thus, in order to be consistent with the wave-analysis, the real part of the eigenvalue λ_+ should exhibit the same dependence—to first order in ε.

Proposition 4.6. *The amplification of the λ_+-wave* (4.18) *can be written as*

$$\mathrm{Re}\lambda_+ = -(\gamma - \beta^2)\varepsilon k^2 + \mathcal{O}(\varepsilon^2), \tag{4.48}$$

consistent with the Chapman–Enskog approximation.

Proof. The formal expansion of (4.18) around $\varepsilon = 0$ can be written as

$$\begin{aligned}\mathrm{Re}\lambda_+ =& \mathrm{Re}\lambda_+(0) + \left.\frac{\partial \mathrm{Re}\lambda_+}{\partial \varepsilon}\right|_{\varepsilon=0} \varepsilon + \mathcal{O}(\varepsilon^2) \\ =& 0 + \left.\frac{\partial \mathrm{Re}\lambda_+}{\partial \xi}\right|_{\xi=0} \frac{\partial \xi}{\partial \varepsilon}\varepsilon + \mathcal{O}(\varepsilon^2).\end{aligned} \tag{4.49}$$

Inserting for (4.23) we obtain

$$\mathrm{Re}\lambda_+ = Ck^2\varepsilon + \mathcal{O}(\varepsilon^2), \tag{4.50}$$

where

$$C = -\left.\frac{\partial}{\partial \xi}\right|_{\xi=0} \frac{1}{2\xi}\left[-1 \pm \left(\frac{\Theta(\xi)}{2} + 1 - 4\xi^2\gamma\right)^{1/2}\right]. \tag{4.51}$$

We can calculate

$$\begin{aligned}&\frac{\partial}{\partial \xi}\frac{1}{2\xi}\left[-1 \pm \left(\frac{\Theta(\xi)}{2} + 1 - 4\xi^2\gamma\right)^{1/2}\right] \\ &= \frac{1}{2}\left[\frac{1}{\xi^2} - \frac{1}{\xi^2}\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{1/2} + \frac{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{-1/2}}{4\xi}\left(\frac{\partial \Theta}{\partial \xi} - 16\xi\gamma\right)\right] \\ &\qquad = \frac{1}{2\xi^2}\left[1 + \left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{-1/2}\left(\frac{\xi}{4}\frac{\partial \Theta}{\partial \xi} - \frac{\Theta}{2} - 1\right)\right].\end{aligned} \tag{4.52}$$

In order to evaluate (4.52) in $\xi = 0$, we need apply L'Hôpital's rule two times. Let

$$f(\xi) = 1 + \left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{-1/2}\left(\frac{\xi}{4}\frac{\partial \Theta}{\partial \xi} - \frac{\Theta}{2} - 1\right), \tag{4.53}$$

with the derivatives

$$f'(\xi) = -\frac{1}{2}\frac{\frac{1}{2}\frac{\partial\Theta}{\partial\xi} - 8\xi\gamma}{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{3/2}}\left(\frac{\xi}{4}\frac{\partial\Theta}{\partial\xi} - \frac{\Theta}{2} - 1\right) + \frac{1}{4}\frac{\xi\frac{\partial^2\Theta}{\partial\xi^2} - \frac{\partial\Theta}{\partial\xi}}{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{1/2}} \tag{4.54}$$

and

$$\begin{aligned} f''(\xi) = &- \frac{1}{2}\frac{\frac{1}{2}\frac{\partial^2\Theta}{\partial\xi^2} - 8\gamma}{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{3/2}}\left(\frac{\xi}{4}\frac{\partial\Theta}{\partial\xi} - \frac{\Theta}{2} - 1\right) \\ &+ \frac{3}{4}\frac{\left(\frac{1}{2}\frac{\partial\Theta}{\partial\xi} - 8\xi\gamma\right)^2}{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{5/2}}\left(\frac{\xi}{4}\frac{\partial\Theta}{\partial\xi} - \frac{\Theta}{2} - 1\right) \\ &+ \frac{1}{2}\frac{\frac{1}{2}\frac{\partial\Theta}{\partial\xi} - 8\xi\gamma}{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{3/2}}\left(\frac{\xi}{4}\frac{\partial^2\Theta}{\partial\xi^2} - \frac{1}{2}\frac{\partial\Theta}{\partial\xi}\right) \\ &- \frac{1}{2}\frac{\frac{\xi}{4}\frac{\partial^2\Theta}{\partial\xi^2} - \frac{1}{4}\frac{\partial\Theta}{\partial\xi}}{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{1/2}}\left(\frac{1}{2}\frac{\partial\Theta}{\partial\xi} - 8\xi\gamma\right) - \frac{\xi}{4}\frac{\frac{\partial^3\Theta}{\partial\xi^3}}{\left(\frac{\Theta}{2} + 1 - 4\xi^2\gamma\right)^{1/2}}. \end{aligned} \tag{4.55}$$

Recall that the function $\Theta(\xi)$ is defined as

$$\Theta(\xi) \equiv \left(\left(1 - 4\xi^2\gamma\right)^2 + 16\xi^2\beta^2\right)^{1/2} + 4\xi^2\gamma - 1, \tag{4.56}$$

giving the derivatives

$$\frac{\partial\Theta}{\partial\xi} = \frac{-16\xi\gamma(1 - 4\xi^2\gamma) + 32\xi\beta^2}{2\left((1 - 4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2}} + 8\xi\gamma, \tag{4.57}$$

and

$$\frac{\partial^2\Theta}{\partial\xi^2} = \frac{-16\gamma + 192\xi^2\gamma^2 + 32\beta^2}{2\left((1 - 4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{1/2}} - \frac{\left(-16\xi\gamma(1 - 4\xi^2\gamma) + 32\xi\beta^2\right)^2}{4\left((1 - 4\xi^2\gamma)^2 + 16\xi^2\beta^2\right)^{3/2}} + 8\gamma. \tag{4.58}$$

By inspecting (4.56)–(4.58), we obtain

$$\Theta(0) = 0, \quad \left.\frac{\partial \Theta}{\partial \xi}\right|_{\xi=0} = 0 \quad \text{and} \quad \left.\frac{\partial^2 \Theta}{\partial \xi^2}\right|_{\xi=0} = 16\beta^2. \tag{4.59}$$

Two successive applications of L'Hôpital's rule yields

$$C = \frac{\lim_{\xi\to 0} f'(\xi)}{\lim_{\xi\to 0} 4\xi} = \frac{1}{4}\lim_{\xi\to 0} f''(\xi), \tag{4.60}$$

and finally, by using (4.59) with (4.55), we see that

$$C = -(\gamma - \beta^2). \tag{4.61}$$

□

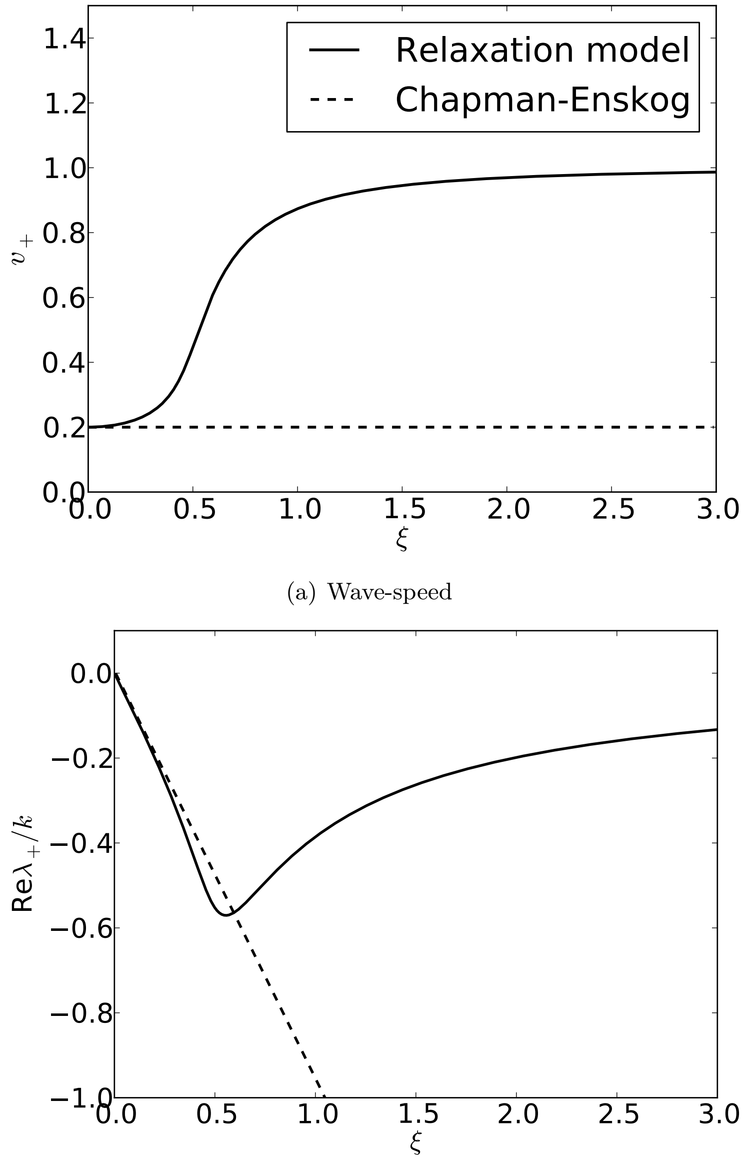
Proposition 4.6 shows that the Chapman–Enskog approximation can be validated in the context of the wave-analysis. Moreover, because of the dual interpretation of the parameter ξ, the Chapman–Enskog approximation can be seen as valid in both the equilibrium ($\varepsilon \to 0$) and the low wave-number ($k \to 0$) limit.

Figure 4.3 shows—for the example-model—the wave-speed and amplification of the convection-diffusion-equation of the Chapman–Enskog approximation, compared to the λ_+-wave of the relaxation system. As shown in Proposition 4.5, a diffusive term gives a k^2-dependence in the amplification. For the example-model, one can clearly see that this approximation is valid in the stiff limit.

4.5 Critical Point

The region of validity of the Chapman–Enskog approximation, as showed in Figure 4.3, seems to motivate the definition of a critical point $\xi_c > 0$, where:

- For $\xi \ll \xi_c$, the dynamics of the relaxation system is like that of the equilibrium system.
- For $\xi \gg \xi_c$, the dynamics of the relaxation system is like that of the homogeneous relaxation-system.



(a) Wave-speed

(b) Amplification

Figure 4.3: The wave-speed and amplification of the λ_+-wave compared to the Chapman–Enskog approximation for the example-model, using $\lambda_R = 1$ and $\lambda_L = 0.2$.

The extremum of the amplification factor of the λ_+-wave, if unique, could be a suitable candidate for such a critical point. It is defined as

$$\xi_\mathrm{c} := \left\{ \xi > 0 : \frac{\partial \mathrm{Re} \lambda_+}{\partial \xi} = 0 \right\}, \tag{4.62}$$

and from Figure 4.3 it seems to be the point where, at least for the example-model, the diffusive approximation stops being valid.

Because of the dual interpretation of ξ, there is also a dual interpretation of such a critical point:

Given a relaxation time ε, a critical point ξ_c will define a critical wave-number k_c. Fourier components with $k \ll k_\mathrm{c}$ will then resemble the equilibrium system; conversely, components with $k \gg k_\mathrm{c}$ will resemble the homogeneous relaxation system.

Also, for a given wave-number k, there will exist a critical relaxation-time ε_c.

4.6 Summary

The topic of this chapter has been the linear analysis of the wave-dynamics of 2 × 2 systems. The main lesson is that the wave-dynamics depend on the parameter $\xi = k\varepsilon$ as follows:

Stiff region For $\xi \to 0$, the dynamics of the relaxation system is *equilibrium-like* and consists of one wave.

Non-stiff region For $\xi \to \infty$, the dynamics is similar to that of the homogeneous relaxation system, and consists of two waves.

Intermediate region For finite ξ, the dynamics consists of two waves with amplification or dampening—depending on the sub-characteristic condition. The wave-speeds of the system will have values between the wave-speeds of the homogeneous relaxation system (a *transitional* sub-characteristic condition). The wave-dynamics in this region is dispersive, and the wave-speeds are monotonic functions of ξ.

The dispersion effect in the intermediate region is in agreement with numerical observations made by Munkejord [34] for a more complicated relaxation system. This indicate that the results from this chapter might also be valid in a more general sense. To the author's knowledge, a detailed study of the wave-speeds in the intermediate region has not previously been performed.

Because of the way ξ is defined, all these results have a dual interpretation. For instance, the stiff region can either be seen as the high-frequency region ($k \to 0$) or as the short-relaxation-time region ($\varepsilon \to 0$).

Also, in Proposition 4.6 it was showed that the diffusive term of the Chapman–Enskog approximation is valid to first order in ξ (and ε). In other words, because of the symmetric scaling of the variable ξ, the Chapman–Enskog approximation can be seen as valid also in the high-frequency limit. This is a result that—to the author's knowledge—has not been commented on in the literature.

Finally, in Section 4.5, the notion of a *critical point* ξ_c was introduced. It was argued that for the example model, the critical point coincides with the point where the diffusive Chapman–Enskog approximation no longer can be thought of as valid.

5 An Exponential Time-Differencing Method for Monotonic Relaxation Systems*

We consider stiff relaxation processes, emphasizing the application to hyperbolic conservation laws. We present first and second-order accurate exponential time-differencing methods for systems of monotonic relaxation ODEs. Some desirable accuracy and robustness properties of these methods are established.

Through operator splitting, we show how the methods may be applied to hyperbolic conservation laws with relaxation terms. In particular, global second-order accuracy for smooth solutions may be achieved through Strang splitting and MUSCL interpolation. An application to granular-gas flow is presented.

5.1 Introduction

We are interested in numerical methods for hyperbolic relaxation systems in the form

$$\partial_t \boldsymbol{U} + \partial_x \boldsymbol{F}(\boldsymbol{U}) = \frac{1}{\varepsilon}\boldsymbol{R}(\boldsymbol{U}), \tag{5.1}$$

to be solved for the unknown M-vector $\boldsymbol{U}$. Herein, $\boldsymbol{R}(\boldsymbol{U})$ is a *relaxation source term*, the effect of which is to drive the system towards some local equilibrium value $\boldsymbol{U}^{\mathrm{eq}}$. The parameter ε represents a characteristic *relaxation time* towards equilibrium. This relaxation time is typically small, imposing a high degree of stiffness in the system (5.1).

*The content of this chapter has been submitted for publication as an article. It is co-written by Steinar Evje, Tore Flåtten, Knut Erik Teigen Giljarhus and Svend Tollak Munkejord.

Such systems were extensively analysed by Chen et al. [8], with a particular focus on the stiff limit $\varepsilon \to 0$. In this paper, we investigate numerical methods suitable for systems in the form (5.1) for nonzero, yet small values of ε. In particular, we will use *fractional-step* methods, based on splitting the system (5.1) into two parts:

(i) The conservation law

$$\partial_t \boldsymbol{U} + \partial_x \boldsymbol{F}(\boldsymbol{U}) = 0; \tag{5.2a}$$

(ii) The ordinary differential equation

$$\partial_t \boldsymbol{U} = \frac{1}{\varepsilon} \boldsymbol{R}(\boldsymbol{U}). \tag{5.2b}$$

This allows for applying methods that are particularly tailored to such problems individually. In particular, we here focus on methods particularly suited for relaxation models in the form (5.2b).

Recently, a popular approach towards solving stiff systems in the form (5.2b) has been the use of *exponential integrators* [11, 19, 27]. Such methods are motivated mainly by computational efficiency considerations [18]; without sacrificing high-order accuracy, one gets rid of the severe restriction on the time step commonly associated with explicit methods for stiff problems. The main idea behind such methods consists of splitting the source term into a linear and a nonlinear part as follows:

$$\frac{1}{\varepsilon} \boldsymbol{R}(\boldsymbol{U}) = \boldsymbol{L}\boldsymbol{U} + \boldsymbol{N}(\boldsymbol{U}), \tag{5.3}$$

where $\boldsymbol{L}$ is a constant $M \times M$ matrix. One then attempts to associate the stiffness of the system (5.2b) with the linear part, which may be solved exactly through the matrix exponential. Coupled to this, the non-linear part $\boldsymbol{N}(\boldsymbol{U})$ is solved by standard Runge–Kutta methods.

In this paper, we wish to emphasize another aspect of exponential time-differencing methods; the potential for strong robustness in the sense that the numerical solution is bounded with no restriction on the time step. In particular, one may use such methods to ensure that the relaxation step does not introduce unphysical solutions such as vacuum or negative-density states.

To achieve this, we here present what seems to us a slightly original twist to the idea of exponential integrators. Instead of viewing the exponential integration step as the *exact* solution to a linear sub-problem as given by the splitting (5.3), we interpret the exponential integration as a *numerical approximation* to the original nonlinear problem, and this approximation is nevertheless accurate to a certain order in the time step. This change of perspective leads to a slightly different formulation, and allows us to construct consistent methods that *by design* guarantee that the equilibrium solution cannot be exceeded.

For consistency, the property that the numerical solution is bounded by the equilibrium value must be shared by the mathematical solution. Therefore, we will limit our investigations in this paper to what we denote as *monotonic* equations in the relaxation step (5.2b), as defined more precisely in Section 5.3. This restricts the class of systems where our methods are applicable, but in particular includes many relaxation processes of interest within the context of (5.1).

Furthermore, as the solution of such hyperbolic relaxation systems tend to remain close to the equilibrium state, we are interested in deriving methods that exhibit a particularly high level of accuracy near equilibrium. In these respects, the methods investigated in this paper may be particularly well suited for systems in the form (5.1).

However, the investigations in this paper are in many ways preliminary. In particular, our analysis is limited to the relaxation step (5.2b). We do not formally address the convergence of our splitting method when applied to the full system (5.1). Hence, the purpose of this paper may be summarized as follows.

- We wish to emphasize the potential robustness properties of exponential methods. Towards this aim, we explicitly present first and second-order methods possessing a strong form of stability, which we will denote as *monotonic asymptotic* stability.

- We wish to demonstrate the practical feasibility of such methods by applying them to a benchmark case previously investigated in the literature.

By this, we hope to pave the way for further work.

This paper is organized as follows. In Section 5.2, we briefly review hyperbolic relaxation systems in the form (5.1), and some existing numerical methods to solve such systems. In Section 5.3, we present the exponential integration technique which is the topic of this paper. First and second-order versions are provided. We also prove the following.

(i) The methods are stable in the strong sense that no numerical overshoots of the equilibrium value are possible.

(ii) The methods are accurate in the sense that they correspond to the exact solution to first-order deviations from the equilibrium.

In Section 5.4, we describe a granular-gas model investigated by Serna and Marquina [42]. In Section 5.5, we present some numerical examples. Herein, Section 5.5.2 details our numerical method as applied to the granular-gas model. The simulations indicate that our proposed method compares satisfactorily to results previously reported in the literature.

Finally, in Section 5.6 we summarize our results and discuss some directions for further work.

5.2 Hyperbolic Relaxation Systems

A hyperbolic relaxation system can be written in general quasilinear form as follows [36]:

$$\partial_t \boldsymbol{U} + \boldsymbol{A}(\boldsymbol{U})\partial_x \boldsymbol{U} = \frac{1}{\varepsilon}\boldsymbol{R}(\boldsymbol{U}), \tag{5.4}$$

where the matrix $\boldsymbol{A}$ is assumed to be diagonalizable with real eigenvalues in the domain of interest. In the context of (5.1), $\boldsymbol{A}$ is given by

$$\boldsymbol{A}(\boldsymbol{U}) = \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{U}}. \tag{5.5}$$

Such systems model many relevant physical problems, such as two-phase flows which are locally not in thermodynamic equilibrium [13, 14, 41, 49].

The limiting process $\varepsilon \to 0$ in systems in the form (5.4) was extensively analysed by Liu [31] and Chen et al. [8], with a particular focus on the relationship between stability and wave propagation. In this paper, we are interested in *numerical methods* for systems in the form (5.4) when the relaxation source term is stiff; i.e. the parameter ε is so small that the time scales associated with the homogeneous system (5.2a) are significantly larger than the time scales associated with the relaxation terms (5.2b).

Several approaches have been proposed in the literature. These may be roughly divided into *splitting* and *unsplit* methods [38].

5.2.1 Fractional-Step Methods

We assume a uniform computational grid, and let $\boldsymbol{U}_j^n$ denote the cell averages of $\boldsymbol{U}$ in the cell $[x_{j-1/2}, x_{j+1/2}]$ at time t^n. Let $\mathcal{H}(t)$ be the operator that advances the system (5.2a) forward in time, and let $\mathcal{S}(t)$ be the corresponding stiff ODE operator for the system (5.2b). Then we may consider two main classes of splitting methods [21]:

- *Godunov splitting*:

$$\boldsymbol{U}^{n+1} = \mathcal{S}\left(\Delta t\right) \circ \mathcal{H}\left(\Delta t\right) \boldsymbol{U}^n, \tag{5.6}$$

- *Strang splitting* [43]:

$$\boldsymbol{U}^{n+1} = \mathcal{H}\left(\frac{1}{2}\Delta t\right) \circ \mathcal{S}\left(\Delta t\right) \circ \mathcal{H}\left(\frac{1}{2}\Delta t\right) \boldsymbol{U}^n. \tag{5.7}$$

Godunov splitting is first-order accurate, whereas Strang splitting is second-order accurate provided that both $\mathcal{H}$ and $\mathcal{S}$ are second-order accurate operators. In particular, Strang splitting applied to (5.2a)–(5.2b) is second-order accurate for any fixed ε. However, as emphasized by Pareschi and Russo [38], and proved by Jin [22], the method in general degenerates to first order in the limit $\varepsilon \to 0$. Although this limit may never be fully realized in practical applications, this is nevertheless an undesirable property. Following the terminology of [38], we will denote schemes that retain their order of accuracy also in the limit $\varepsilon \to 0$ as *asymptotically accurate*.

Jin [22] proposed an asymptotically second-order accurate splitting method based on two-stage Runge–Kutta time integration. This paved the way for a currently popular class of methods; implicit-explicit (IMEX) Runge–Kutta methods [4, 5, 38] where an explicit discretization is applied to the flux terms and an implicit one to the source terms. This provides a general framework for achieving high-order asymptotic accuracy.

In this paper however, we are interested in exploring robust *explicit* methods for the relaxation source terms. For simplicity, we will remain in the framework of the Godunov and Strang splittings described above. For the hyperbolic operator $\mathcal{H}$, we will use the MUSTA method of Toro [44], augmented with the MUSCL approach of van Leer [45].

Our stiff operator $\mathcal{S}$ will be described in the following section.

5.3 Monotonically Asymptotic Exponential Integration

In general, relaxation processes in the form (5.2b) only affect some of the variables of the full system. Furthermore, the relaxation processes often represent an exchange of a conserved property between two variables, for which the relaxation source term will differ only in sign.

This situation allows us to fully express the solution vector $\boldsymbol{U}$ through the dynamics of a *reduced* variable $\boldsymbol{V}(\boldsymbol{U})$, with rank $N < M$. For the purposes of this paper, we make the following definition.

Definition 5. *Consider the equation*

$$\partial_t \boldsymbol{V} = \frac{1}{\varepsilon}\boldsymbol{S}(\boldsymbol{V}), \qquad \boldsymbol{V} \in \mathcal{D} \subseteq \mathbb{R}^N \tag{5.8}$$

where $\boldsymbol{S}(\boldsymbol{V})$ is a smooth function. The system is said to be a **relaxation ODE** *provided there exists a unique point $\boldsymbol{V}^{\text{eq}} \in \mathcal{D}$ such that $\boldsymbol{S}(\boldsymbol{V}^{\text{eq}}) = 0$, and the solution satisfies*

$$\lim_{t\to\infty} \boldsymbol{V}(t) = \boldsymbol{V}^{\text{eq}}. \tag{5.9}$$

Herein, the initial condition $\boldsymbol{U}_0$ of (5.2b) determines an invertible function $\boldsymbol{V}(\boldsymbol{U})$, as well as the function $\boldsymbol{S}(\boldsymbol{V})$ and the point $\boldsymbol{V}^{\text{eq}}$. This will be illustrated by an explicit example in Section 5.4.2.

5.3.1 Monotonic Relaxation ODEs

One way of solving relaxation ODEs is by using *exponential integrators*, an idea that dates back at least several decades [12, 29]. A common starting point for such methods is a splitting of the source term into a linear and a nonlinear part as follows [3, 11, 18, 20]:

$$\frac{1}{\varepsilon}\boldsymbol{S}(\boldsymbol{V}) = \boldsymbol{L}\boldsymbol{V} + \boldsymbol{N}(\boldsymbol{V}), \tag{5.10}$$

where $\boldsymbol{L}$ is a constant $N \times N$ matrix. The linear part may then be solved exactly through the matrix exponential of $\boldsymbol{L}$; this solution is then coupled to the nonlinear part $\boldsymbol{N}(\boldsymbol{V})$ through standard Runge–Kutta methods.

For stiff problems, exponential integrators allow for larger time steps and improved stability compared to straightforward Runge–Kutta methods. Berland et al. [3] presented a general theory for constructing higher-order versions of such exponential integrators.

Much of the literature focuses on computational *accuracy* and *efficiency*. In our current paper, we wish to shift the focus more strongly towards numerical *robustness*. Towards this end, we first define a subclass of relaxation ODEs.

Definition 6. *A relaxation ODE in the form* (5.8) *is said to be a* **monotonic relaxation ODE** *if*

$$V_i'(t)\,(V_i^{\mathrm{eq}} - V_i) > 0 \quad \forall V_i \neq V_i^{\mathrm{eq}} \tag{5.11}$$

for all $i \in \{1, \ldots, N\}$.

In other words, we denote the system as monotonic if all the components of the solution vector are monotonic functions of time. From (5.8) and (5.11) we immediately see that a *necessary* condition for a system in the form (5.8) to be a monotonic relaxation ODE is that the source term must satisfy

$$S_i(\boldsymbol{V})\,(V_i^{\mathrm{eq}} - V_i) > 0 \quad \forall V_i \neq V_i^{\mathrm{eq}} \tag{5.12}$$

for all $i \in \{1, \ldots, N\}$.

Within the framework of hyperbolic relaxation systems in the form (5.1), monotonicity seems to be a rather inclusive restriction. For instance, it is an essential property of scalar relaxation ODEs.

Proposition 5.1. *All scalar relaxation ODEs are monotonic, and a scalar ODE in the form* (5.8) *is a relaxation ODE if and only if there is a point* V^{eq} *such that*

$$[\min(V_0, V^{\mathrm{eq}}), \max(V_0, V^{\mathrm{eq}})] \subseteq \mathcal{D} \tag{5.13}$$

and

$$\begin{gathered} S(V)\,(V^{\mathrm{eq}} - V) > 0 \quad \forall V \neq V^{\mathrm{eq}}, \\ S(V^{\mathrm{eq}}) = 0. \end{gathered} \tag{5.14}$$

Proof. Either all scalar relaxation ODEs are monotonic, or some orbit of V exists where $V'(t)$ changes sign. However, given that $S(V)$ is a smooth function, such a point could only be the equilibrium point V^{eq} which would remain constant in time. Hence all scalar relaxation ODEs are monotonic, and (5.14), being a special case of (5.12), is a *necessary* condition for (5.8) to be a relaxation ODE.

Now if the initial condition V_0 is given as $V_0 = V^{\text{eq}}$, then $S(V_0) = 0$ and (5.9) holds trivially. Otherwise, for any δ satisfying

$$0 < \delta < |V^{\text{eq}} - V_0|, \tag{5.15}$$

we define

$$\mathcal{W} = \left[\min(V_0, V^{\text{eq}} + \delta), \max(V^{\text{eq}} - \delta, V_0)\right], \tag{5.16}$$

$$C = \min_{V \in \mathcal{W}} |S(V)|. \tag{5.17}$$

It follows from (5.14) that C^{-1} is a finite number, and that

$$|V(t > T) - V^{\text{eq}}| < \delta \tag{5.18}$$

where T is given by

$$T = \frac{|V^{\text{eq}} - V_0|}{C} \varepsilon. \tag{5.19}$$

Hence (5.9) holds, and (5.14) is also a *sufficient* condition for (5.8) to be a relaxation ODE in the scalar case. □

If the relaxation processes are fully independent, this property will carry directly over to systems. For instance, the relaxation part of the five-equation two-phase flow model investigated by Munkejord [34], describing simultaneous volume and momentum transfer, consists of independent relaxation processes and is monotonic in the sense of Definition 6.

Remark 5.1. *A simple example of a coupled, nonlinear and globally monotonic relaxation system can be constructed as follows:*

$$\boldsymbol{V} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \qquad \boldsymbol{S}(\boldsymbol{V}) = \begin{bmatrix} \left(\alpha_1 + \beta_1 v_2^2\right) v_1 \\ \left(\alpha_2 + \beta_2 v_1^2\right) v_2 \end{bmatrix}, \tag{5.20}$$

where

$$\boldsymbol{V}^{\mathrm{eq}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{5.21}$$

and

$$\alpha_i, \beta_i < 0 \quad \forall i \in \{1, 2\}. \tag{5.22}$$

This is however a theoretical example, and monotonicity may easily be lost for strongly coupled relaxation systems of practical interest. Consequently, one should be aware that the methods presented in this paper are fully general only for the scalar case, yet also applicable to a limited class of coupled systems.

5.3.2 A Strong Stability Requirement

An essential property of monotonic relaxation systems is that the solution vector remains bounded by the equilibrium value at all times. To avoid unphysical solutions and numerical oscillations, we want our numerical method to possess an analogous property.

Definition 7. *Consider a monotonic relaxation ODE with initial conditions $\boldsymbol{V}^n$ and equilibrium point $\boldsymbol{V}^{\mathrm{eq}}$. Let the numerical solution be given through some operator $\mathcal{S}(\Delta t)$ as*

$$\boldsymbol{V}^{n+1} = \mathcal{S}(\Delta t)\boldsymbol{V}^n. \tag{5.23}$$

The operator $\mathcal{S}$ will be denoted as **monotonically asymptotically stable** *if it satisfies the following properties.*

MA1: The operator is **consistent** *with the relaxation system to be solved, i.e. the local truncation error is of at least second order in Δt.*

MA2: The solution is unconditionally **bounded** *by the equilibrium value, i.e.*

$$\begin{aligned} V_i^{n+1} &\in (V_i^n, V_i^{\mathrm{eq}}) && \textit{for} \quad V_i^n < V_i^{\mathrm{eq}}, \\ V_i^{n+1} &= V_i^n && \textit{for} \quad V_i^n = V_i^{\mathrm{eq}}, \\ V_i^{n+1} &\in (V_i^{\mathrm{eq}}, V_i^n) && \textit{for} \quad V_i^n > V_i^{\mathrm{eq}} \end{aligned} \tag{5.24}$$

for all $i \in \{1, \ldots, N\}$ and for all Δt.

Common explicit methods typically do not possess this form of stability. For instance, the forward Euler method satisfies the property MA2 only conditionally, with a strong restriction on the time step:

$$\frac{\Delta t}{\varepsilon} < \min_i \left(\frac{V_i^{\mathrm{eq}} - V_i^n}{S_i(\boldsymbol{V}^n)} \right). \tag{5.25}$$

Implicit methods may however possess such strong stability, as exemplified as follows.

Proposition 5.2. *The backward Euler method, defined by*

$$\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + \frac{\Delta t}{\varepsilon} \boldsymbol{S}(\boldsymbol{V}^{n+1}), \tag{5.26}$$

is monotonically asymptotically stable in the sense of Definition 7.

Proof. It is well known and easy to check that the backward Euler method is consistent; i.e. the property MA1 is satisfied. We now prove the property MA2 by showing that we otherwise get contradictions. First, we note that the backward Euler method preserves the equilibrium point. We now consider the case $V_i^{\mathrm{eq}} > V_i^n$. Assume that the solution $\boldsymbol{V}^{n+1}$ of (5.26) satisfies

$$V_i^{n+1} < V_i^n. \tag{5.27}$$

From (5.12), we then have $S_i(\boldsymbol{V}^{n+1}) > 0$ which inserted into (5.26) yields $V_i^{n+1} > V_i^n$, in contradiction to (5.27).

Similarly, assume that the solution $\boldsymbol{V}^{n+1}$ of (5.26) satisfies

$$V_i^{n+1} > V_i^n. \tag{5.28}$$

From (5.12), we then have $S_i(\boldsymbol{V}^{n+1}) < 0$ which inserted into (5.26) yields $V_i^{n+1} < V_i^n$, in contradiction to (5.28).

The same steps will prove the remaining case $V_i^{\mathrm{eq}} < V_i^n$. □

Implicit methods generally require the solution of a system of nonlinear equations, which raises its own computational efficiency and robustness issues. This motivates the *explicit* monotonically asymptotically stable method presented in the following.

Definition 8. *The numerical method given by*

$$V_i^{n+1} = V_i^n + (V_i^{\text{eq}} - V_i^n)\left(1 - \exp\left(-\frac{\Delta t}{\tau_i}\right)\right), \tag{5.29}$$

where

$$\tau_i = \varepsilon \frac{V_i^{\text{eq}} - V_i^n}{S_i(\boldsymbol{V}^n)}, \tag{5.30}$$

will be denoted as the **ASY1** *method.*

Proposition 5.3. *The ASY1 method is monotonically asymptotically stable in the sense of Definition 7.*

Proof. Assume first that $V_i^{\text{eq}} \neq V_i^n$. Taylor-expanding (5.29) shows that the method is consistent to first order with (5.8). Note also that (5.29) satisfies

$$\lim_{V_i^n \to V_i^{\text{eq}}} V_i^{n+1} = V_i^{\text{eq}}, \tag{5.31}$$

hence the property MA1 is satisfied. From (5.12) and (5.30) it also follows that the exponential function is bounded by the interval $(0, 1]$. Hence the property MA2 is satisfied. □

Note that the ASY1 method (5.29) inserts a numerical "barrier" at the point $V_i = V_i^{\text{eq}}$ through which the solution can never pass. Hence the method cannot be consistent unless this barrier is also present in the underlying mathematical equation, as is the case for monotonic relaxation ODEs.

Otherwise, we will formally lose first-order accuracy at the barrier, as described in the following.

Proposition 5.4. *When applied to a general ODE*

$$\partial_t \boldsymbol{V} = \boldsymbol{Q}(\boldsymbol{V}), \tag{5.32}$$

where $\boldsymbol{Q}(\boldsymbol{V})$ is a smooth function, the method (5.29)–(5.30) *is consistent in the limit $V_i^n \to V_i^{\text{eq}}$ only if*

$$Q_i(\boldsymbol{V}) = 0 \quad \textit{for} \quad V_i = V_i^{\text{eq}}. \tag{5.33}$$

Proof. The local truncation error of the method for the component V_i can be written as

$$T_i(\boldsymbol{V}^n) = \frac{1}{2}(\Delta t)^2 \left(Q_i(\boldsymbol{V}^n) \left(\frac{\partial Q_i}{\partial V_i} + \frac{Q_i(\boldsymbol{V}^n)}{V_i^{\mathrm{eq}} - V_i^n} \right) + \sum_{k \neq i} \frac{\partial Q_i}{\partial V_k} Q_k(\boldsymbol{V}^n) \right) + \mathcal{O}(\Delta t^3). \tag{5.34}$$

Now if (5.33) holds, we obtain

$$\lim_{V_i^n \to V_i^{\mathrm{eq}}} \frac{\partial Q_i}{\partial V_k} = 0 \quad \forall k \neq i, \tag{5.35}$$

and also

$$\lim_{V_i^n \to V_i^{\mathrm{eq}}} T_i(\boldsymbol{V}^n) = 0. \tag{5.36}$$

However, if (5.33) does *not* hold, the second-order coefficient diverges and the local truncation error degenerates to

$$T_i(\boldsymbol{V}^n) \sim \mathcal{O}(\Delta t). \tag{5.37}$$

□

The notion of *monotonic asymptotic* stability may be interpreted as a *dual* consistency principle; consistency in the large (MA2) and the small (MA1), or the stiff and non-stiff limit of the time step.

5.3.3 Accuracy Near Equilibrium

The exponential function employed in (5.29) is of course only one of many functions that asymptotically approaches a limit value. However, it becomes the natural choice as it corresponds to the *exact* solution for linear monotonic relaxation problems. In this respect, it is worth noting that solutions to relaxation systems in the form (5.1) tend to remain close to equilibrium. We have the following proposition.

Proposition 5.5. *When applied to a monotonic relaxation ODE, the ASY1 method is exact to first-order perturbations to the equilibrium state. More precisely, if we write*

$$\boldsymbol{V}^n = \boldsymbol{V}^{\mathrm{eq}} + \delta \tilde{\boldsymbol{V}}, \tag{5.38}$$

then for all $\Delta t \geq 0$ *the numerical solution* (5.29) *satisfies*

$$V_i(t^n + \Delta t) - V_i^{n+1} = \mathcal{O}(\delta^2) \quad \forall i, \tag{5.39}$$

where $\boldsymbol{V}(t)$ *is the exact solution.*

Proof. It follows from monotonicity that

$$V_i(t) - V_i^{\mathrm{eq}} \sim \mathcal{O}(\delta) \quad \forall i. \tag{5.40}$$

Consequently, we may expand the source term as

$$S_i(\boldsymbol{V}(t)) = S_i(\boldsymbol{V}^{\mathrm{eq}}) + \sum_{k=1}^{N} \frac{\partial S_i}{\partial V_k} \left(V_k(t) - V_k^{\mathrm{eq}}\right) + \mathcal{O}(\delta^2). \tag{5.41}$$

By definition a monotonic relaxation ODE satisfies

$$S_i(\boldsymbol{V}) = 0 \quad \text{for} \quad V_i = V_i^{\mathrm{eq}}, \tag{5.42}$$

hence

$$\frac{\partial S_i}{\partial V_k} = 0 \quad \text{for} \quad k \neq i \tag{5.43}$$

at the point $\boldsymbol{V}^{\mathrm{eq}}$, and (5.41) reduces to

$$S_i(\boldsymbol{V}(t)) = \frac{\partial S_i}{\partial V_i} \left(V_i(t) - V_i^{\mathrm{eq}}\right) + \mathcal{O}(\delta^2). \tag{5.44}$$

As this holds for all t, we may write

$$S_i(\boldsymbol{V}(t)) = S_i(\boldsymbol{V}^n) \frac{V_i^{\mathrm{eq}} - V_i(t)}{V_i^{\mathrm{eq}} - V_i^n} + \mathcal{O}(\delta^2), \tag{5.45}$$

and using (5.8) we obtain

$$\varepsilon \frac{V_i^{\mathrm{eq}} - V_i^n}{S_i(\boldsymbol{V}^n)} \frac{\mathrm{d}V_i(t)}{V_i^{\mathrm{eq}} - V_i(t)} = \left(1 + \mathcal{O}(\delta)\right) \mathrm{d}t, \tag{5.46}$$

where we have used that

$$S_i(\boldsymbol{V}^n) \sim \mathcal{O}(\delta) \quad \forall i. \tag{5.47}$$

Integrating (5.46) we obtain

$$\frac{V_i^{\mathrm{eq}} - V_i(t+\Delta t)}{V_i^{\mathrm{eq}} - V_i^n} = \exp\left(-\frac{S_i(\boldsymbol{V}^n)\Delta t}{\varepsilon(V_i^{\mathrm{eq}} - V_i^n)}\right) + \mathcal{O}(\delta), \tag{5.48}$$

which can be rewritten as

$$V_i(t+\Delta t) = V_i^n + (V_i^{\mathrm{eq}} - V_i^n)\left(1 - \exp\left(-\frac{S_i(\boldsymbol{V}^n)\Delta t}{\varepsilon(V_i^{\mathrm{eq}} - V_i^n)}\right)\right) + \mathcal{O}(\delta^2). \tag{5.49}$$

We now recover (5.39) by using (5.29)–(5.30). □

5.3.4 Second-Order Accuracy

A general explicit two-stage Runge–Kutta scheme for the ODE (5.8) can be written in the form

$$\boldsymbol{V}^* = \boldsymbol{V}^n + a\frac{\Delta t}{\varepsilon}\boldsymbol{S}(\boldsymbol{V}^n) \tag{5.50}$$

$$\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + \frac{\Delta t}{\varepsilon}\left(b_1\boldsymbol{S}(\boldsymbol{V}^n) + b_2\boldsymbol{S}(\boldsymbol{V}^*)\right), \tag{5.51}$$

where for second-order accuracy the parameters a, b_1 and b_2 must satisfy (see for instance [26, Ch. 8]):

$$b_1 + b_2 = 1, \qquad ab_2 = \frac{1}{2}. \tag{5.52}$$

In this section, we make some preliminary investigations into higher-order versions of the ASY method by devising a similar two-stage application of (5.29).

Definition 9. *The numerical method given by*

$$V_i^* = V_i^n + (V_i^{\mathrm{eq}} - V_i^n)\left(1 - \exp\left(-a\frac{\Delta t}{\tau_i}\right)\right) \tag{5.53}$$

$$V_i^{n+1} = V_i^n + (V_i^{\mathrm{eq}} - V_i^n)\left(1 - b_1\exp\left(-\frac{\Delta t}{\tau_i}\right) - b_2\exp\left(-\frac{\Delta t}{\tau_i^*}\right)\right), \tag{5.54}$$

where

$$\tau_i = \varepsilon \frac{V_i^{\text{eq}} - V_i^n}{S_i(\boldsymbol{V}^n)}, \qquad \tau_i^* = \varepsilon \frac{V_i^{\text{eq}} - V_i^*}{S_i(\boldsymbol{V}^*)}, \tag{5.55}$$

and the parameters a, b_1 and b_2 satisfy

$$b_1 + b_2 = 1, \qquad ab_2 = \frac{1}{2}, \tag{5.56}$$

as well as

$$b_2 \in (0, 1], \tag{5.57}$$

will be denoted as the **ASY2** *method.*

Proposition 5.6. *The ASY2 method is second-order accurate in Δt when applied to a monotonic relaxation ODE.*

Proof. Expanding τ_i^* we obtain

$$\frac{1}{\tau_i^*} = \frac{1}{\tau_i}\left(1 + a\Delta t\left(\frac{1}{\tau_i} + \frac{1}{S_i(\boldsymbol{V}^n)}\sum_{k=1}^{N}\frac{\partial S_i}{\partial V_k}\frac{S_k(\boldsymbol{V}^n)}{\varepsilon}\right)\right) + \mathcal{O}(\Delta t^2), \tag{5.58}$$

where have used that

$$V_i^* = V_i^n + a\frac{\Delta t}{\varepsilon}S_i(\boldsymbol{V}^n) + \mathcal{O}(\Delta t^2), \tag{5.59}$$

$$S_i(\boldsymbol{V}^*) = S_i(\boldsymbol{V}^n) + a\frac{\Delta t}{\varepsilon}\sum_{k=1}^{N}\frac{\partial S_i}{\partial V_k}S_k(\boldsymbol{V}^n) + \mathcal{O}(\Delta t^2). \tag{5.60}$$

Substituting (5.58) into (5.54) and expanding the exponential function we obtain

$$\begin{aligned} V_i^{n+1} &= V_i^n + \frac{\Delta t}{\varepsilon}S_i(\boldsymbol{V}^n)\,(b_1 + b_2) \\ &+ \frac{1}{2}\frac{\Delta t^2}{\varepsilon^2}\left((2ab_2 - b_1 - b_2)\frac{S_i(\boldsymbol{V}^n)^2}{V_i^{\text{eq}} - V_i^n} + 2ab_2\sum_{k=1}^{N}\frac{\partial S_i}{\partial V_k}S_k(\boldsymbol{V}^n)\right) + \mathcal{O}(\Delta t^3), \end{aligned} \tag{5.61}$$

whereas the exact solution satisfies

$$V_i(t^n + \Delta t) = V_i^n + \frac{\Delta t}{\varepsilon}S_i(\boldsymbol{V}^n) + \frac{1}{2}\frac{\Delta t^2}{\varepsilon^2}\sum_{k=1}^{N}\frac{\partial S_i}{\partial V_k}S_k(\boldsymbol{V}^n) + \mathcal{O}(\Delta t^3). \tag{5.62}$$

Now using (5.56) we may write

$$V_i(t^n + \Delta t) - V_i^{n+1} = \mathcal{O}(\Delta t^3) \quad \forall V_i^n \neq V_i^{\mathrm{eq}}. \tag{5.63}$$

We finally observe that ASY2 method respects the limit

$$\lim_{V_i^n \to V_i^{\mathrm{eq}}} V_i^{n+1} = V_i^{\mathrm{eq}}. \tag{5.64}$$

□

Proposition 5.7. *The ASY2 method is monotonically asymptotically stable in the sense of Definition 7.*

Proof. The property MA1 follows immediately from Proposition 5.6. From (5.12), it follows that the exponential functions of (5.54) are bounded by the interval $(0, 1]$. The property MA2 then follows from (5.56)–(5.57). □

As might be expected, Proposition 5.5 also naturally extends to the ASY2 method.

Proposition 5.8. *When applied to a monotonic relaxation ODE, the ASY2 method is exact to first-order perturbations to the equilibrium state. More precisely, if we write*

$$\boldsymbol{V}^n = \boldsymbol{V}^{\mathrm{eq}} + \delta \tilde{\boldsymbol{V}}, \tag{5.65}$$

then for all $\Delta t \geq 0$ *the numerical solution* (5.53)–(5.54) *satisfies*

$$V_i(t^n + \Delta t) - V_i^{n+1} = \mathcal{O}(\delta^2) \quad \forall i, \tag{5.66}$$

where $\boldsymbol{V}(t)$ *is the exact solution.*

Proof. We have

$$S_i(\boldsymbol{V}^*) = \frac{\partial S_i}{\partial V_i} (V_i^* - V_i^{\mathrm{eq}}) + \mathcal{O}(\delta^2), \tag{5.67}$$

hence from (5.55) we obtain

$$\frac{1}{\tau_i^*} = \frac{1}{\tau_i} + \mathcal{O}(\delta^2). \tag{5.68}$$

Using (5.56), we may then write (5.54) as

$$\begin{aligned} V_i^{n+1} &= V_i^n + (V_i^{\mathrm{eq}} - V_i^n)\left(1 - \exp\left(-\frac{\Delta t}{\tau_i}\right)\left(b_1 + b_2(1 + \mathcal{O}(\delta^2)\right)\right) \\ &= V_i^n + (V_i^{\mathrm{eq}} - V_i^n)\left(1 - \exp\left(-\frac{\Delta t}{\tau_i}\right)\right) + \mathcal{O}(\delta^3). \end{aligned} \tag{5.69}$$

In other words, ASY1 and ASY2 coincide to second order in perturbations to the equilibrium state. The result then follows directly from Proposition 5.5. □

5.4 A Granular-Gas Flow Model

Granular gases have lately been the subject of considerable theoretical, numerical and experimental studies [15, 39, 38, 42, 40]. In this work we consider a continuum model for granular-gas flow, in which the dynamics are accounted for by a hyperbolic conservation law with relaxation. In addition to having been previously studied in the literature, this model is suitable for our current purposes for the following reasons.

- The relaxation part of the system is a monotonic nonlinear relaxation ODE.
- The equilibrium state corresponds to a granular temperature $T = 0$ and is hence easy to calculate.
- Numerically overshooting the equilibrium would be undesirable, as it would lead to the unphysical state $T < 0$.

5.4.1 Fluid-Mechanical Equations

The dynamics of a one-dimensional granular-gas flow under the influence of gravity, in the form considered by Serna and Marquina [42], can be described by the Euler-like equations

$$\partial_t \rho + \partial_x(\rho u) = 0, \tag{5.70a}$$

$$\partial_t(\rho u) + \partial_x(\rho u^2 + p) = \rho g, \tag{5.70b}$$

$$\partial_t E + \partial_x\, u(E + p) = \Theta + \rho g u. \tag{5.70c}$$

In the above, ρ is the density, u is the velocity, p is the pressure, g is the gravitational acceleration, E is the energy density and Θ is the rate of energy loss due to inelastic collisions. The energy density consists of both kinetic and internal energy and is given by $E = (1/2)\rho u^2 + (3/2)\rho T$, where T is the granular temperature.

Following Serna and Marquina [42], we use an energy-loss term based on Haff's cooling law [17], given by

$$\Theta(\rho, T) = -\frac{12}{\sqrt{\pi}} \frac{1-e^2}{\sigma} \rho T^{3/2} G(\nu), \tag{5.71}$$

where σ is the particle diameter and $e \in [0, 1]$ is the restitution coefficient. For $e = 1$ we recover a fully elastic model. The statistical correlation function $G(\nu)$ is given by

$$G(\nu) = \nu \left(1 - \left(\frac{\nu}{\nu_M}\right)^{\frac{3}{4}\nu_M}\right)^{-1}, \tag{5.72}$$

where $\nu = (\pi/6)\rho\sigma^3$ is the volume fraction and ν_M is the maximal volume fraction.

The pressure is determined by a granular equation of state (EOS), introduced by Goldshtein and Shapiro [15], given by

$$p(\rho, T) = T\rho(1 + 2(1 + e)G(\nu)). \tag{5.73}$$

5.4.2 The Relaxation ODE

Within the splitting (5.2a)–(5.2b), we obtain

$$\boldsymbol{U} = \begin{bmatrix} \rho \\ \rho u \\ E \end{bmatrix} \qquad \text{and} \qquad \frac{1}{\varepsilon}\boldsymbol{R}(\boldsymbol{U}) = \begin{bmatrix} 0 \\ 0 \\ \Theta(\rho, T) \end{bmatrix}. \tag{5.74}$$

For any initial condition

$$\boldsymbol{U}_0 = \begin{bmatrix} \rho_0 \\ \rho_0 u_0 \\ E_0 \end{bmatrix}, \tag{5.75}$$

this may be written in the reduced form (5.8) with

$$V(\boldsymbol{U}) = E, \tag{5.76}$$

$$\frac{1}{\varepsilon}S(V) = -\frac{4}{\sqrt{3\pi}}\frac{1-e^2}{\sigma}\rho_0\left(2\frac{V}{\rho_0} - u^2\right)^{3/2} G(\nu_0). \tag{5.77}$$

Furthermore, for any V we can reconstruct the full state vector $\boldsymbol{U}$ as

$$\boldsymbol{U}(V) = \begin{bmatrix} \rho_0 \\ \rho_0 u_0 \\ V \end{bmatrix}. \tag{5.78}$$

5.5 Numerical Tests

5.5.1 Verification of the Order of Convergence

The purpose of this section is to numerically verify the order of convergence of the monotonically asymptotically stable integrators presented in Section 5.3. Specifically, we wish to verify that the ASY1 scheme (5.29) is first-order accurate and that the ASY2 scheme (5.53)–(5.54) is second-order accurate. The two-stage ASY2 scheme is completely determined by the parameter a in the order conditions (5.56). For the calculations of this paper, we choose the parameter $a = 1$. By this choice, we only need two evaluations of the exponential function in (5.53)–(5.54).

We consider an initial value problem, based on the scalar relaxation ODE of the granular-gas model, given by

$$\partial_t E(t) = -\frac{4}{\sqrt{3\pi}}\frac{1-e^2}{\sigma}\rho_0\left(2\frac{E(t)}{\rho_0} - u_0^2\right)^{3/2} G(\rho_0), \quad E(0) = E_0. \tag{5.79}$$

For this numerical test we use $e = 0.97$, $\sigma = 10^{-3}\,\mathrm{m}$, $\rho_0 = 10.0\,\mathrm{kg/m^3}$ and $u_0 = 18.0\,\mathrm{m/s}$. The initial energy is given by

$$E_0 = 3966.5\,\mathrm{J/m^3}, \tag{5.80}$$

and the corresponding equilibrium energy is $E^{\mathrm{eq}} = (1/2)\rho_0 u_0^2 = 1620.0\,\mathrm{J/m^3}$.

A reference solution $E_{\mathrm{ref}}(1.0)$ was calculated using the second-order modified Euler scheme with a step size $\Delta t = 2^{-20}\,\mathrm{s}$. The modified Euler

Table 5.1: The error $\mathcal{E} = |E_{\text{ref}}(1.0) - \hat{E}(1.0)|$ in the numerical solution at $t = 1.0\,\text{s}$ using the ASY1 scheme, for different values of the step-size Δt. The number n indicates the estimated order of convergence.

Δt	$\mathcal{E}^i$	$\mathcal{E}^{i-1}/\mathcal{E}^i$	n
2^{-2}	4.02869674	-	-
2^{-3}	1.99680512	2.0176	1.0126
2^{-4}	0.99408608	2.0087	1.0063
2^{-5}	0.49597241	2.0043	1.0031
2^{-6}	0.24771960	2.0022	1.0016
2^{-7}	0.12379328	2.0011	1.0008
2^{-8}	0.06188003	2.0005	1.0004
2^{-9}	0.03093586	2.0003	1.0002
2^{-10}	0.01546689	2.0001	1.0001

scheme is given by the two-step explicit Runge–Kutta method (5.50)–(5.51) with $a = 0.5$. In order to estimate the order of convergence we calculate the error $\mathcal{E} = |E_{\text{ref}}(1.0) - \hat{E}(1.0)|$ for numerical solutions $\hat{E}$, using different step sizes. Let $\mathcal{E}^i$ be the error using step-size $\Delta t_i = 2^{-i}$, for $i \in \mathbb{N}$. For sufficiently small Δt, the order of convergence n is then given by

$$n = \log_2 \left(\frac{\mathcal{E}^{i-1}}{\mathcal{E}^i} \right). \tag{5.81}$$

Table 5.1 shows the error and the estimated order of convergence for the ASY1 scheme. The results are consistent with a first-order solver.

Table 5.2 shows the error and estimated order of convergence for the ASY2 scheme. These results agree with this being a second-order accurate scheme.

5.5.2 Numerical Method

In order to numerically test the ASY methods on the granular-gas model described in Section 5.4, we use a fractional-step method as described in Section 5.2.1. This means that we need a numerical solver for the

Table 5.2: The error $\mathcal{E} = |E_{\text{ref}}(1.0) - \hat{E}(1.0)|$ in the numerical solution at $t = 1.0\,\text{s}$ using the ASY2 scheme, for different values of the step-size Δt. The number n indicates the estimated order of convergence.

Δt	$\mathcal{E}^i$	$\mathcal{E}^{i-1}/\mathcal{E}^i$	n
2^{-2}	0.01629915	-	-
2^{-3}	0.00393750	4.1395	2.0494
2^{-4}	0.00096757	4.0695	2.0248
2^{-5}	0.00023981	4.0347	2.0125
2^{-6}	0.00005969	4.0173	2.0062
2^{-7}	0.00001489	4.0086	2.0031
2^{-8}	0.00000372	4.0041	2.0015
2^{-9}	0.00000093	4.0014	2.0005
2^{-10}	0.00000023	3.9982	1.9993

hyperbolic part (5.2a) to use in tandem with the exponential integrator.

A Multi-Stage Scheme

We consider a uniform grid in space and time, and denote $t^n = t_0 + n\,\Delta t$ and $x_j = x_0 + j\,\Delta x$. For a first-order accurate numerical scheme, we advance the solution $\boldsymbol{U}_j^n$ forward in time by using

$$\boldsymbol{U}_j^{n+1} = \boldsymbol{U}_j^n + \mathcal{F}_j^n \Delta t, \tag{5.82}$$

where

$$\mathcal{F}_j^n = \frac{1}{\Delta x}\left(\boldsymbol{F}_{j-1/2}^n - \boldsymbol{F}_{j+1/2}^n\right) + \boldsymbol{\mathcal{Q}}(\boldsymbol{U}_j^n). \tag{5.83}$$

In the above, $\boldsymbol{F}_{j+1/2}^n$ is the numerical approximation to the inter-cell flux and $\boldsymbol{\mathcal{Q}}(\boldsymbol{U}_j^n)$ are local source terms other than relaxation terms. For the granular-gas model, $\boldsymbol{\mathcal{Q}}(\boldsymbol{U})$ will be the gravity source terms.

In the Multi-Stage (MUSTA) approach, the inter-cell flux is calculated by solving the local Riemann problem at each cell interface on a local grid [44]. The solution on the local grid is then advanced in several stages giving an approximation to the inter-cell flux. In our application, we will use four local grid cells and two local iteration steps. The CFL number of the local grid is kept the same as on the global grid.

High Resolution

In a high resolution (second order) extension to the MUSTA scheme, we employ a second-order strong-stability-preserving (SSP) Runge–Kutta method to advance the solution forward in time. The two-stage scheme is given by

$$\begin{aligned} \boldsymbol{U}_j^* &= \boldsymbol{U}_j^n + \mathcal{F}_j^n \Delta t, \\ \boldsymbol{U}_j^{n+1} &= \frac{1}{2}\boldsymbol{U}_j^n + \frac{1}{2}\boldsymbol{U}_j^* + \frac{1}{2}\mathcal{F}_j^* \Delta t. \end{aligned} \tag{5.84}$$

In order to obtain second-order accuracy in space, a piecewise linear MUSCL interpolation [37, 45] was used. For the granular-gas model, the variables used in the interpolation were given by

$$\boldsymbol{W} = \begin{bmatrix} \rho & v & p \end{bmatrix}^T . \tag{5.85}$$

We reconstruct these variables to the right and to the left of the cell interface as

$$\boldsymbol{W}_{j+1/2}^{\mathrm{R}} = \boldsymbol{W}_{j+1} - \frac{\Delta x}{2}\boldsymbol{\sigma}_{j+1} \quad \text{and} \quad \boldsymbol{W}_{j+1/2}^{\mathrm{L}} = \boldsymbol{W}_j + \frac{\Delta x}{2}\boldsymbol{\sigma}_j, \tag{5.86}$$

respectively. The cell slopes $\boldsymbol{\sigma}_j$ are calculated using a *minmod* slope, given by

$$\sigma_{j,i} = \operatorname{minmod}\left(\frac{W_{j,i} - W_{j-1,i}}{\Delta x}, \frac{W_{j+1,i} - W_{j,i}}{\Delta x}\right), \tag{5.87}$$

where the minmod function is defined as

$$\operatorname{minmod}(a,b) = \begin{cases} 0 & \text{if } ab \leq 0 \\ a & \text{if } |a| < |b| \text{ and } ab > 0 \\ b & \text{if } |b| < |a| \text{ and } ab > 0 \end{cases}. \tag{5.88}$$

The reconstructed values at the interface are then used for the Riemann problem on the local MUSTA grid, in order to obtain second-order accuracy in space. We refer to the high-resolution scheme as MUSCL-MUSTA.

5.5.3 Case: Granular-Gas Tube

In this section we use the ASY integrators as a part of a fractional-step method in order to compare with previously reported results for the granular-gas model.

We consider the case of a granular gas in a vertical tube hitting a solid wall at the bottom end, as used by Serna and Marquina [42] and also Pareschi and Russo [38]. The 0.1 m tube is initially filled with a granular gas with volume fraction $\nu = 0.018$, velocity 0.18 m/s and pressure $p = 1589.26\,\mathrm{Pa}$. We use the gravitational acceleration $g = 9.8\,\mathrm{m/s}$, the restitution coefficient $e = 0.97$, the maximum volume fraction $\nu_M = 0.65$ and the particle diameter $\sigma = 10^{-3}\,\mathrm{m}$. The left boundary condition is given by an incoming flow consistent with the initial condition. At the right end of the domain we used a reflective boundary condition.

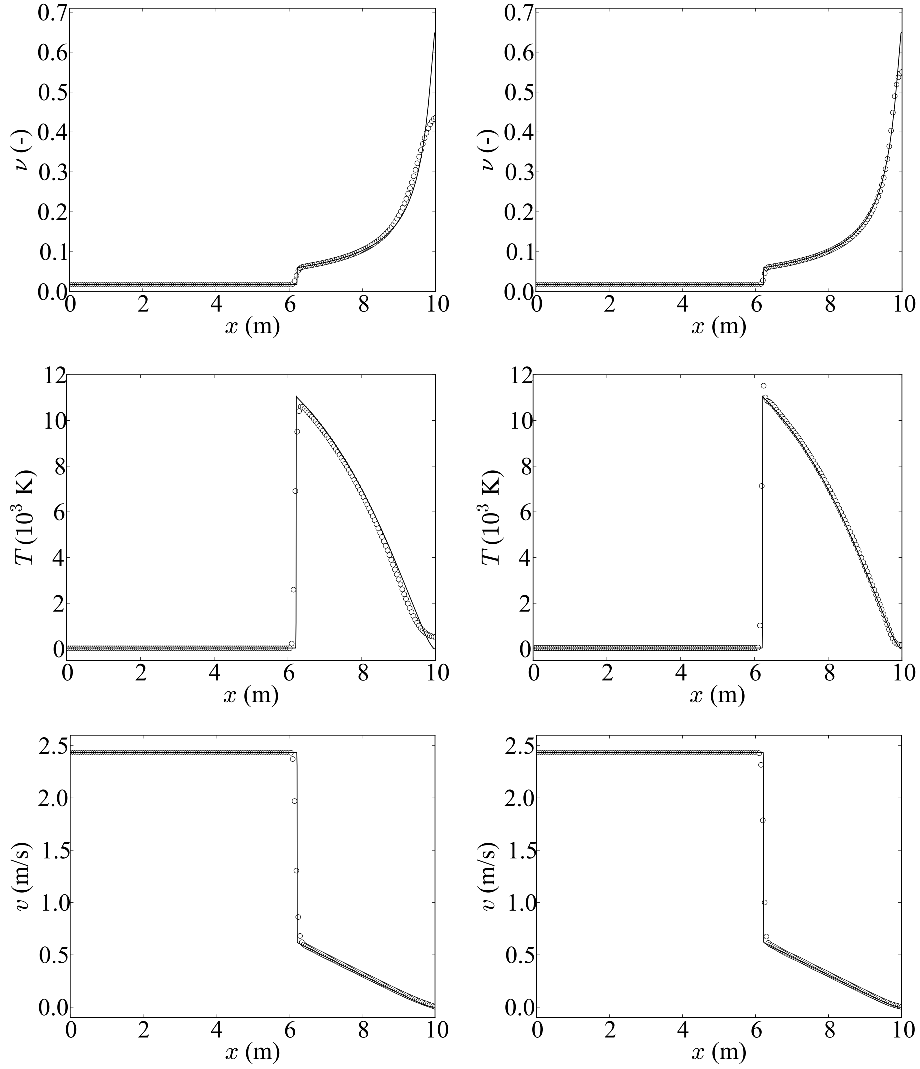
Simulations were carried out using 200 computational cells and a CFL number of 0.4. Figure 5.1 shows the results for the volume fraction, granular temperature and velocity at $t = 0.23\,\mathrm{s}$, using the MUSTA-ASY1 scheme with Godunov splitting and the MUSCL-MUSTA-ASY2 scheme with Strang splitting. The reference solution was computed using the MUSCL-MUSTA-ASY2 scheme with 10 000 cells.

The results show a shock being formed when the gas hits the solid wall. The shock propagates backwards and the gas continues to compress against the wall until the maximum volume fraction is reached at the right boundary. It is also at the right boundary the difference between the first and second-order schemes is most prominent, as illustrated in Figure 5.2. For the second-order MUSCL-MUSTA-ASY2 scheme some spurious oscillations can be observed near the shock, these are associated with the MUSCL interpolation in the hyperbolic step.

Our results do not compare unfavourably to those previously reported [38, 42] in terms of accuracy and numerical robustness.

5.6 Summary

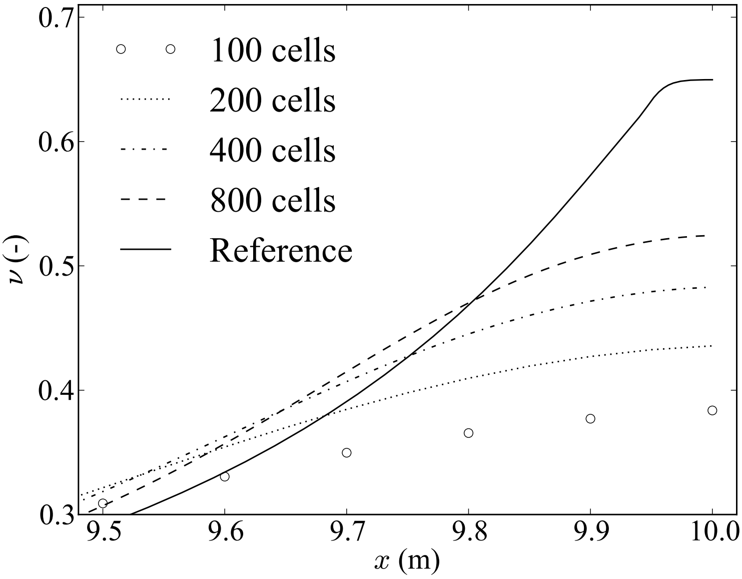
We have investigated a technique, based on exponential integration, for solving monotonic relaxation ODEs. First and second-order versions of the method have been presented. We have proved that the resulting methods possess desirable accuracy and stability properties. In particular, for first-



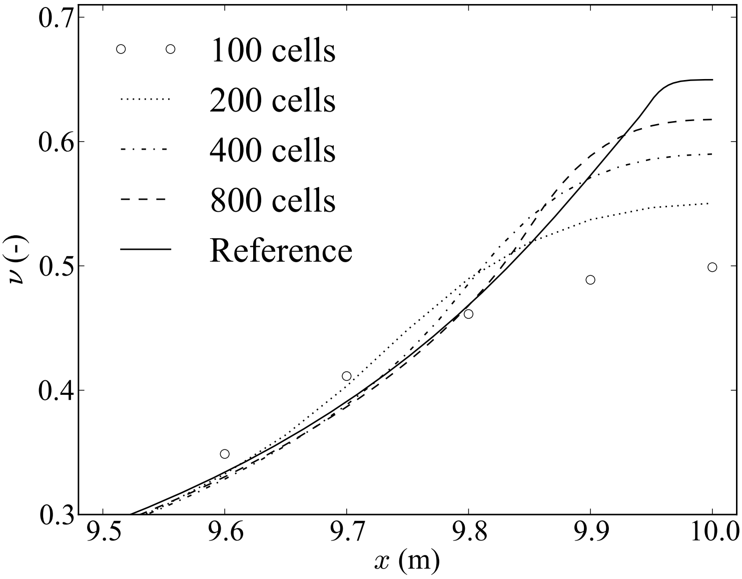
(a) MUSTA-ASY1

(b) MUSCL-MUSTA-ASY2

Figure 5.1: Granular-gas shock case at $t = 0.23\,\mathrm{s}$ for the MUSTA-ASY1 scheme and the MUSCL-MUSTA-ASY2 scheme. The solid line is the reference solution.



(a) MUSTA-ASY1



(b) MUSCL-MUSTA-ASY2

Figure 5.2: Convergence under grid refinement at the right boundary at $t = 0.23\,\mathrm{s}$.

order corrections to the equilibrium value, the methods yield the exact solution. Furthermore, the methods yield numerical solutions that are unconditionally bounded by the equilibrium state.

Through operator splitting, we have applied the methods to a system of hyperbolic conservation laws with relaxation, representing flow of granular gases. The simulations indicate that the currently selected approach, based on MUSCL interpolation in the hyperbolic step, is comparable to previously published results in terms of accuracy and appearance of numerical oscillations.

In summary, we have analytically demonstrated beneficial properties of the methods in the stiff and non-stiff limits of the time step. Our numerical experiments further verify the applicability of the methods for intermediate time steps. Hence the approach shows promise for solving hyperbolic relaxation processes where robustness in the relaxation step is essential, for instance to avoid vacuum or negative-temperature states.

Further work includes deriving higher-order conditions for general multi-stage versions of the method. In this context, it would also be of interest to derive unsplit versions of the approach, following for instance the ideas of Jin [22]. An extension to more general systems, through the matrix exponential, should also be investigated.

6 Numerical Simulations

This chapter is devoted to the numerical solution of a linear 2×2 relaxation model and its corresponding equilibrium model and Chapman–Enskog approximation.

6.1 The Example-Model

The same basic model which was used as an example-model in Chapter 4 will be used in the numerical simulations. The full relaxation system and the different derived models, are:

The Relaxation System

The hyperbolic relaxation system is given by

$$\partial_t u + \partial_x v = 0 \tag{6.1a}$$

$$\partial_t v + \lambda_R^2 \partial_x u = \frac{1}{\varepsilon}(\lambda_E u - v). \tag{6.1b}$$

where λ_R and λ_E are fixed parameters of the model.

The Homogeneous Relaxation System

The system

$$\partial_t u + \partial_x v = 0 \tag{6.2a}$$

$$\partial_t v + \lambda_R^2 \partial_x u = 0. \tag{6.2b}$$

will as usual be referred to as the *homogeneous* relaxation system. The model can be seen as the limit $\varepsilon \to \infty$ of the relaxation system, i.e. the limit where the effect of the relaxation term is negligible.

The Equilibrium Model

The equilibrium model is given by the local equilibrium condition

$$v = \lambda_E u \tag{6.3}$$

and the advection equation

$$\partial_t u + \lambda_E \partial_x u = 0 \tag{6.4}$$

in the reduced variable u. Note that $\pm\lambda_R$ is the characteristic speeds of the homogeneous relaxation system and λ_E is the characteristic speed of the equilibrium model. The sub-characteristic condition thus takes the form

$$\lambda_R^2 \geq \lambda_E^2. \tag{6.5}$$

The Chapman–Enskog-Approximation

In Section 3.4, the Chapman–Enskog approximation was derived for general linear 2 × 2 systems. Applied to the example-model (6.1a)–(6.1b), this approximation takes the form of the correction

$$v = \lambda_E u - \varepsilon(\lambda_R^2 - \lambda_E^2)\partial_x u \tag{6.6}$$

and the advection-diffusion equation

$$\partial_t u + \lambda_E \partial_x u = \varepsilon(\lambda_R^2 - \lambda_E^2)\partial_{xx} u. \tag{6.7}$$

6.2 Numerical Scheme

The purpose the present exercise is to numerically demonstrate some properties of the solutions to 2×2 hyperbolic relaxation systems. To this end, we seek a simple and robust numerical scheme; numerical accuracy and efficiency will not be considered in this section.

6.2.1 A Fractional-Step Method

In order to do numerical simulations on the models listed in Section 6.1, there is a need for numerical schemes for solving:

- Linear 2 × 2 Hyperbolic Relaxation Systems

$$\partial_t \boldsymbol{u}(x,t) + A\,\partial_x \boldsymbol{u}(x,t) = \frac{1}{\varepsilon} R \boldsymbol{u}(x,t) \tag{6.8}$$

- Advection-equations

$$\partial_t u(x,t) + a\,\partial_x u(x,t) = 0 \tag{6.9}$$

- Scalar advection-diffusion equations

$$\partial_t u(x,t) + a\,\partial_x u(x,t) = D\,\partial_{xx} u(x,t) \tag{6.10}$$

These problems can all be written in the general form

$$\partial_t \boldsymbol{u}(x,t) + A\,\partial_x \boldsymbol{u}(x,t) = \boldsymbol{S}[\boldsymbol{u}], \tag{6.11}$$

where the brackets indicate that $\boldsymbol{S}[\boldsymbol{u}]$ might have a non-local dependence on $\boldsymbol{u}$.

Equations in the form (6.11) can be solved by a fractional-step method [30, Ch. 17]. The idea is to split the problem into two sub-problems:

- Hyperbolic Conservation Law

$$\partial_t \boldsymbol{u}(x,t) + A\,\partial_x \boldsymbol{u}(x,t) = 0 \tag{6.12}$$

- Abstract ODE

$$\partial_t \boldsymbol{u}(x,t) = \boldsymbol{S}[\boldsymbol{u}] \tag{6.13}$$

Each of the sub-problems are then solved numerically in an alternating manner in order to be consistent with the original problem. A benefit of this approach is that we divide a composite problem into pieces that can more easily be solved by themselves, using standard methods.

Godunov Splitting

Let $\boldsymbol{u}^n$ denote the numerical solution a time $t = t^0 + n\,\Delta t$. A simple fractional-step method, often referred to as *Godunov Splitting*, consists of two steps: First, solve the hyperbolic conservation law in one time-step yielding an intermediate solution $\boldsymbol{u}^*$. Then use $\boldsymbol{u}^*$ as the initial condition for solving the abstract ODE in one time-step giving the solution in the next time-step $\boldsymbol{u}^{n+1}$. A schematic representation of Godunov-splitting for the problem (6.11) is given in Figure 6.1.

Godunov Splitting is first-order accurate in time if the numerical methods used in each sub-step are at least first-order accurate.

$$\begin{tikzcd}
\boldsymbol{u}^n \arrow[rr, dotted, "\partial_t \boldsymbol{u} + A\partial_x \boldsymbol{u} = \boldsymbol{S}[\boldsymbol{u}]", "\Delta t"'] \arrow[dr, "\Delta t", "\partial_t \boldsymbol{u} + A\partial_x \boldsymbol{u} = 0"'] & & \boldsymbol{u}^{n+1} \\
& \boldsymbol{u}^* \arrow[ur, "\Delta t", "\partial_t \boldsymbol{u} = \boldsymbol{S}[\boldsymbol{u}]"'] &
\end{tikzcd}$$

Figure 6.1: Illustration of Godunov splitting for problems in the form (6.11). The conservation law is solved in one time-step giving an intermediate solution $\boldsymbol{u}^*$. The intermediate solution is then used as the initial value for solving the ODE in one time-step.

6.2.2 The Lax–Friedrichs Scheme

A straightforward explicit finite-volume-method for solving the hyperbolic conservation law (6.12) of the fractional-step method is the Lax–Friedrichs Scheme.

Let space and time be discretized in the usual way as $x_i = x_0 + i\,\Delta x$ and $t^n = t^0 + n\,\Delta t$, respectively. If we denote the numerical solution as $\boldsymbol{u}(x_i, t^n) = \boldsymbol{u}_i^n$, the Lax–Friedrichs scheme for the linear hyperbolic equation (6.12) can be written as [30, Ch. 4]

$$\boldsymbol{u}_i^* = \frac{1}{2}\left(\boldsymbol{u}_{i+1}^n + \boldsymbol{u}_{i-1}^n\right) - \frac{1}{2}\frac{\Delta t}{\Delta x} A\left(\boldsymbol{u}_{i+1}^n - \boldsymbol{u}_{i-1}^n\right). \tag{6.14}$$

For stability of the scheme, we must require

$$\nu \equiv \max_j\{\lambda_j\}\frac{\Delta t}{\Delta x} \leq 1 \tag{6.15}$$

where $\{\lambda_j\}$ are the eigenvalues of A. This criterion is often called the CFL-criterion and ν the CFL-number, named after Courant, Friedrichs and Lewy.

6.2.3 Exponential Time-Differencing

For the full relaxation system (6.1a)–(6.1b), the abstract ODE (6.13) takes the form

$$\partial_t v = \frac{1}{\varepsilon}(\lambda_E u - v), \tag{6.16}$$

where the reduced variable u is treated as a constant in the ODE-step.

In order to solve the scalar relaxation ODE (6.16) numerically, there is a need for a simple scheme that behaves in a stable manner even in the stiff limit ($\varepsilon \to 0$). To meet this demand, the simple first-order exponential time-differencing scheme (5.29) introduced in Chapter 5 will be used.

For the relaxation ODE (6.16), the ASY1 scheme is given explicitly as

$$v^{n+1} = v^* + (\lambda_E u^* - v^*)\left[1 - \exp\left(-\frac{\Delta t}{\varepsilon}\right)\right], \tag{6.17}$$

where u^* and v^* are the intermediate values calculated in the conservation-law step (6.12) of the Godunov splitting.

6.2.4 The Crank–Nicolson Method

For the Chapman–Enskog approximation, the second step (6.13) of the fractional-step method is the diffusion problem

$$\partial_t u(x,t) = D\,\partial_{xx} u(x,t), \quad u(x,0) = u^*(x), \tag{6.18}$$

for $t \in [0, \Delta t]$. The Crank-Nicolson method applied to the problem (6.18) yields the semi-implicit discretization [33]

$$\frac{u_i^{n+1} - u_i^*}{\Delta t} = \frac{D}{2\,\Delta x^2}\left[\left(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}\right) + \left(u_{i+1}^* - 2u_i^* + u_{i-1}^*\right)\right]. \tag{6.19}$$

Remark 6.1. *The Crank-Nicolson method is second-order accurate in space, and has the benefit of being unconditionally stable for the diffusion equation. The latter is the main reason for using the scheme in our application; we wish to focus on robustness, not accuracy.*

Rearranging (6.19) by moving terms dependent on the $n+1$ time-level to the left-hand side gives

$$-ru_{i+1}^{n+1} + (1+2r)u_i^{n+1} - ru_{i-1}^{n+1} = ru_{i+1}^* + (1-2r)u_i^* + ru_{i-1}^*, \tag{6.20}$$

where we have introduced the shorthand

$$r \equiv \frac{D\Delta t}{2\,\Delta x^2}. \tag{6.21}$$

With the simple extrapolation boundary condition

$$u_0 = u_1 \quad \text{and} \quad u_{N+1} = u_N, \tag{6.22}$$

the equations (6.20) can be written in matrix form as

$$\begin{bmatrix} (1+r) & -r & & & 0 \\ -r & (1+2r) & -r & & \\ & \ddots & \ddots & \ddots & \\ & & -r & (1+2r) & -r \\ 0 & & & -r & (1+r) \end{bmatrix} \begin{bmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_{N-1}^{n+1} \\ u_N^{n+1} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{N-1} \\ b_N \end{bmatrix}, \tag{6.23}$$

where

$$b_i = r u_{i+1}^* + (1-2r) u_i^* + r u_{i-1}^*. \tag{6.24}$$

The matrix equation (6.23) must then be solved for the unknown vector $\boldsymbol{u}^{n+1}$ in each time step. Note that the matrix is in a tri-diagonal form, the system can therefore be solved efficiently using the standard Thomas Algorithm [10].

6.3 Case: Single Discontinuity

For the numerical simulations, we use initial conditions corresponding to the Riemann-problem

$$u(x,0) = \begin{cases} 1.0 & \text{if} \quad x \leq 0.5 \\ 1.2 & \text{if} \quad x > 0.5 \end{cases}. \tag{6.25}$$

We let $v(x,0) = \lambda_E\, u(x,0)$ to ensure that the initial state is an equilibrium state. For simplicity, extrapolation boundary conditions are used, but the simulations are stopped before any waves can propagate from the initial discontinuity to the boundary.

The spatial domain $x \in [0,1]$ is divided into 2000 equally spaced computational cells, and a CFL-number of 0.9 is used.

6.3.1 Wave-Dynamics of the Relaxation Model

The relaxation model (6.1a)–(6.1b) with the initial condition (6.25) was solved numerically using $\lambda_R = 1.0$ and $\lambda_E = 0.2$. Figure 6.2 shows the solution for different values of the relaxation time ε, compared to the solution of the equilibrium model and the homogeneous relaxation system.

The results clearly show that in the stiff limit ($\varepsilon \to 0$) of the relaxation model, the solutions approach the equilibrium solution. Also, in the non-stiff limit, the results seem to converge to that of the homogeneous relaxation model.

Moreover, the equilibrium solution consists of one right-going wave with wave-speed λ_E, while the homogeneous relaxation model has two distinct waves with wave-speeds $\pm\lambda_R$. The relaxation-model on the other hand, is dispersive and has no well-defined wave-speed for a finite ε. However, as ε becomes smaller, more and more Fourier-components in the solution become *equilibrium-like* and the wave-dynamics changes from two waves to a single equilibrium wave.

6.3.2 Validity of the Chapman–Enskog Approximation

The Chapman–Enskog approximation (6.7) was solved numerically for different relaxation times ε, see Figure 6.3.

In Section 4.4 we showed that—to first order in ε—the solution of the relaxation model is equivalent to the solution of the Chapman–Enskog approximation.

The results showed in Figure 6.3 indicate, as expected, that the Chapman–Enskog approximation is valid for small relaxation times. However, for larger ε the solution of the relaxation model breaks off into two distinct waves and the advection-diffusion equation is no longer a good approximation.

6.3.3 Breaking the Sub-Characteristic Condition

In Section 3.3.1 the sub-characteristic condition was shown to be equivalent to the linear stability of the solution of the relaxation model.

To test this statement, the relaxation model was solved numerically for $\lambda_E = 0.2$, $\lambda_E = 0.8$ and $\lambda_E = 1.4$; the parameter $\lambda_R = 1.0$ is kept fixed.

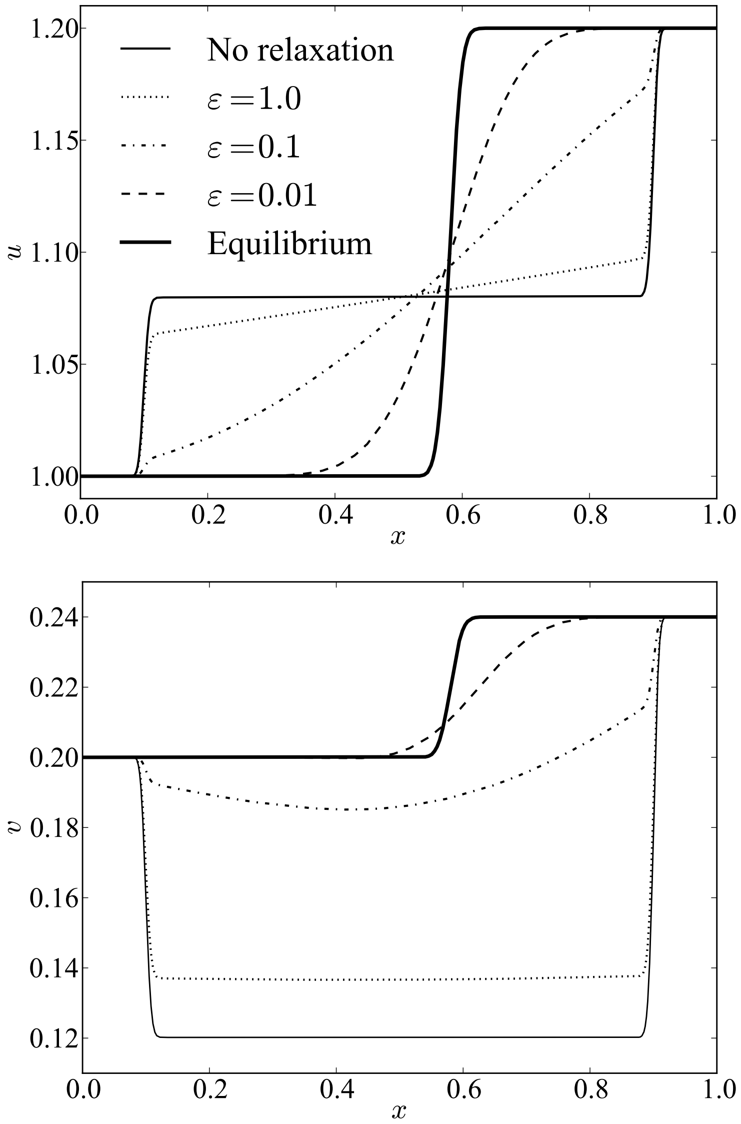
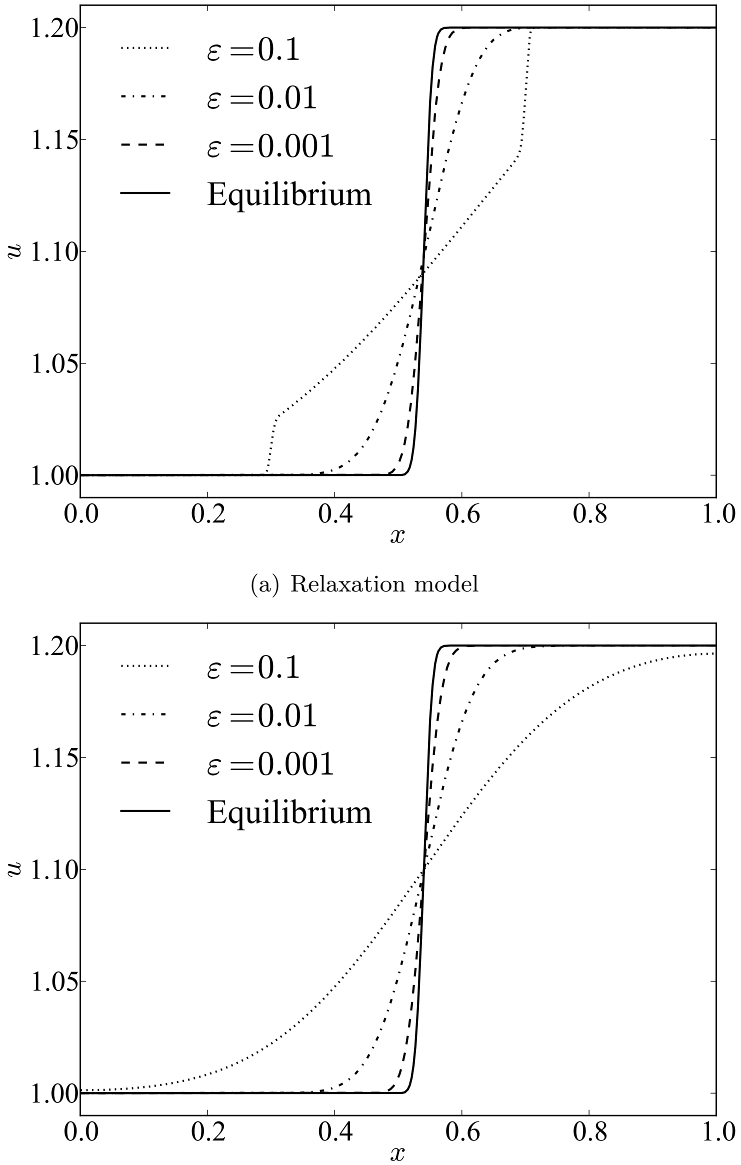


Figure 6.2: Numerical solution at $t = 0.4$ for the single-discontinuity case, for different values of the relaxation time ε. The solutions of the relaxation model are compared to the solutions of the homogeneous relaxation system and the equilibrium model.



(a) Relaxation model

(b) Chapman–Enskog approximation

Figure 6.3: The numerical solution at $t = 0.2$ for the relaxation model and the Chapman–Enskog approximation for the single-discontinuity case, for different values of the relaxation time ε.

For the first two cases the sub-characteristic condition is fulfilled, while for the third case it is not. The relaxation time was $\varepsilon = 0.1$ for all cases.

The results are shown in Figure 6.4, and demonstrate that when the sub-characteristic condition is not fulfilled, an unstable growing peak appears near the right-going wave.

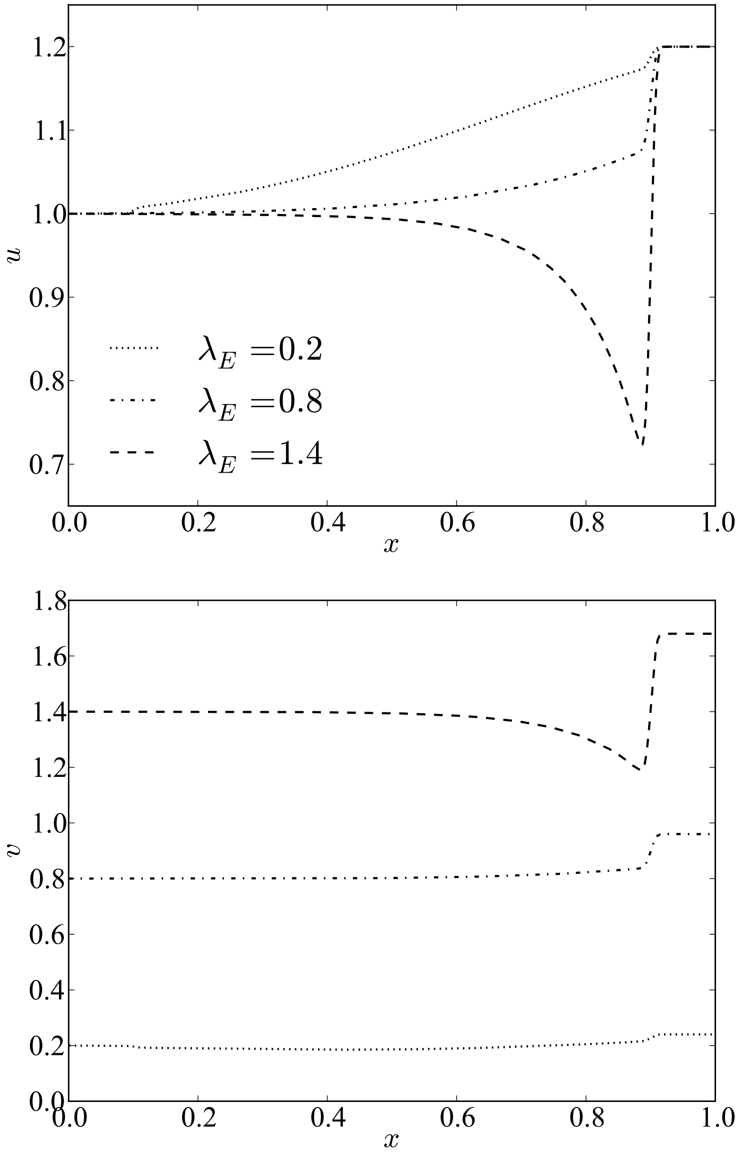


Figure 6.4: Numerical solution at $t = 0.4$ for the relaxation model for the single-discontinuity case, using different values for the equilibrium speed λ_E. $\lambda_R = 1.0$ for all cases.

7 Conclusions and Recommendations for Further Work

7.1 Conclusions

This thesis has treated hyperbolic relaxation systems, with topics both theoretical and numerical in nature. The main conclusions from the work are given below.

7.1.1 The Sub-Characteristic Condition and Dissipativity

Chapter 3 was devoted to discussing some of the properties of 2×2 relaxation systems. Particular attention was given to the relationship between the sub-characteristic condition and the stability of the general solution.

The Chapman–Enskog approximation was derived for linear 2×2 systems in general spatial dimensions. In Proposition 3.3, the dissipativity of the diffusion term of the Chapman–Enskog approximation was showed to be equivalent to the sub-characteristic condition in the 1-D case—a result known from literature.

This relationship was then investigated for higher spatial dimensions. In Proposition 3.5 the relationship between dissipativity and the sub-characteristic condition was shown to hold also in the 2-D case.

7.1.2 Wave-Dynamics of 2×2 Systems

In Chapter 4, the wave-dynamics of 2×2 systems was studied in detail through linear analysis. The main results from this analysis are those related to the wave-speeds of the Fourier-components for the solution.

In Proposition 4.1 and Proposition 4.2, the limit behavior of the wave-dynamics was established. Also, the damping-mechanism responsible for changing the 2-wave dynamics of the homogeneous relaxation system into the 1-wave dynamics of the equilibrium system was identified.

Moreover, in Proposition 4.3, the transitional wave-speeds were shown to be monotonic functions of ξ, which together with the limit behavior proves a *transitional* sub-characteristic condition for 2×2 systems. This result has—to the authors knowledge—not been previously shown.

7.1.3 Exponential Time-Differencing for Relaxation Systems

Chapter 5 was devoted to a new way of numerically solving monotonic relaxation systems using exponential time-differencing with a fractional-step method.

First and second-order schemes were derived, and their order of convergence was numerically verified. The method was shown to be unconditionally stable in the ODE-step of the fractional-step method, with regard to a strong stability-requirement. Moreover, the method turns out to be the exact solution in the ODE-step to first order in the relaxation time ε.

In order to demonstrate the practical use of the method, it was used to numerically solve a granular-gas case previously used in literature. The results did not compare unfavorably to those previously reported.

7.2 Topics for Further Work

It is the opinion of the author that there is no compelling reason why the relationship between the sub-characteristic condition and the dissipativity of the Chapman–Enskog approximation should be limited to 1-D and 2-D. Therefore, a natural topic for further work could be to investigate this for higher dimensions. While the specific techniques used in this thesis to prove this relationship in the 2-D case might prove cumbersome in higher dimensions, it might be possible to construct a more general proof using the properties of the special orthogonal group $\mathrm{SO}(n)$ in n dimensions.

In this thesis it was showed that—for 2×2 systems—the wave-speeds of the individual Fourier-waves will satisfy a transitional sub-characteristic condition. To the author's knowledge, this has not yet been shown or commented in literature. A generalization of this property to $N \times N$ systems might be a topic for further work.

In the context of the work on the exponential time-differencing scheme, a possible topic for further work would be to try to construct higher-order methods. To achieve this, it might be possible to use the same Runge–Kutta-analog that was used to derive the second-order scheme.

Bibliography

[1] A. Aw and M. Rascle. Resurrection of second order models for traffic flow. *J. Appl. Math.*, 60:916–938, 2000.

[2] B. Barker, M. Johnson, L. Rodrigues, and K. Zumbrun. Metastability of solitary roll wave solutions of the St. Venant equations with viscosity. *Arxiv preprint arXiv:1007.5262*, 2010.

[3] H. Berland, B. Owren, and B. Skaflestad. B-series and order conditions for exponential integrators. *SIAM J. Numer. Anal.*, 43:1715–1727, 2005.

[4] S. Boscarino. Error analysis of IMEX Runge–Kutta methods derived from differential-algebraic systems. *SIAM J. Numer. Anal.*, 45:1600–1621, 2007.

[5] S. Boscarino and G. Russo. On a class of uniformly accurate IMEX Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *SIAM J. Sci. Comput.*, 31:1926–1945, 2009.

[6] C. Cerecignani. *The Boltzmann Equation and its Applications.* Springer Verlag, New York, NY, USA, 1988.

[7] S. Chapman and T. Cowling. *The mathematical theory of non-uniform gases*, volume 1. Cambridge University Press, Cambridge, UK, 1991.

[8] G.-Q. Chen, C. D. Levermore, and T.-P. Liu. Hyperbolic conservation laws with stiff relaxation terms and entropy. *Commun. Pure Appl. Math.*, 47:787–830, 1994.

[9] I. Chern. Long-time effect of relaxation for hyperbolic conservation laws. *Commun. Math. Phys.*, 172(1):39–55, 1995.

[10] S. Conte and C. Boor. *Elementary numerical analysis: an algorithmic approach.* McGraw-Hill, New York, NY, USA, 1980.

[11] S. M. Cox and P. P. Matthews. Exponential time differencing for stiff systems. *J. Comput. Phys.*, 176:430–455, 2002.

[12] B. H. Ehle and J. D. Lawson. Generalized Runge–Kutta processes for stiff initial-value problems. *J. Inst. Math. Applics.*, 16:11–21, 1975.

[13] T. Flåtten and H. Lund. Relaxation two-phase flow models and the subcharacteristic condition. *Math. Mod. Meth. Appl. S., accepted for publication.*

[14] T. Flåtten, A. Morin, and S. T. Munkejord. Wave propagation in multicomponent flow models. *SIAM J. Appl. Math.*, 70:2861–2882, 2010.

[15] A. Goldshtein and M. Shapiro. Mechanics of collisional motion of granular materials: Part 1. general hydrodynamic equations. *J. Fluid Mech.*, 282:75–114, 1995.

[16] Z. Gu, N. Nefedov, and R. O'Malley. On singular singularly perturbed initial value problems. *SIAM J. Appl. Math.*, 49(1):1–25, 1989.

[17] P. K. Haff. Grain flow as a fluid mechanical phenomenon. *J. Fluid Mech.*, 134:401–430, 1983.

[18] M. Hochbruck and A. Ostermann. Explicit exponential Runge–Kutta methods for semilinear parabolic problems. *SIAM J. Numer. Anal.*, 43:1069–1090, 2005.

[19] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19: 1552–1574, 1998.

[20] M. Hochbruck, A. Ostermann, and J. Schweitzer. Exponential Rosenbrock-type methods. *SIAM J. Numer. Anal.*, 47:786–803, 2009.

[21] H. Holden, K. H. Karlsen, K.-A. Lie, and N. H. Risebro. Splitting methods for partial differential equations with rough solutions. *EMS Series of Lectures in Mathematics, EMS Publishing House, Zürich*, 2010.

[22] S. Jin. Runge–Kutta methods for hyperbolic systems with stiff relaxation terms. *J. Comput. Phys.*, 122:51–67, 1995.

[23] S. Jin and Z. Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Commun. Pure Appl. Math.*, 48 (3):235–276, 1995.

[24] W. Jintang and Z. Yongshu. Some results on hyperbolic systems with relaxation. *Acta Mathematica Scientia*, 26(4):767–780, 2006.

[25] J. Keizer. *Statistical Thermodynamics of Nonequilibrium Processes.* Springer Verlag, New York, NY, USA, 1987.

[26] D. Kincaid and W. Cheney. *Numerical analysis: Mathematics of scientific computing.* American Mathematical Society, Providence, RI, USA, 2009.

[27] S. Krogstad. Generalized integrating factor methods for stiff PDEs. *J. Comput. Phys.*, 203:72–88, 2005.

[28] C. Lattanzio and P. Marcati. The zero relaxation limit for 2×2 hyperbolic systems. *Nonlinear Anal.*, 38:375–389, 1999.

[29] J. D. Lawson. Generalized Runge–Kutta processes for stable systems with large Lipschitz constants. *J. Numer. Anal.*, 4:372–380, 1967.

[30] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge University Press, New York, NY, USA, 2002.

[31] T.-P. Liu. Hyperbolic conservation laws with relaxation. *Commun. Math. Phys.*, 43:153–175, 1987.

[32] M. Mei and B. Rubino. Convergence to traveling waves with decay rates for solutions of the initial boundary problem to a relaxation model. *J. Differ. Equations*, 159(1):138–185, 1999.

[33] K. Morton and D. Mayers. *Numerical solution of partial differential equations: An introduction.* Cambridge University Press, New York, NY, USA, 2005.

[34] S. T. Munkejord. A numerical study of two-fluid models with pressure and velocity relaxation. *Adv. Appl. Math. Mech.*, 2:131–159, 2010.

[35] R. Natalini. Convergence to equilibrium for the relaxation approximations of conservation laws. *Commun. Pure Appl. Math.*, 49(8): 795–823, 1996.

[36] R. Natalini. Recent results on hyperbolic relaxation problems. Analysis of systems of conservation laws. *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*, 99:128–198, 1997.

[37] S. Osher. Convergence of generalized MUSCL schemes. *SIAM J. Numer. Anal.*, 22(5):947–961, 1985.

[38] L. Pareschi and G. Russo. Implicit-explicit Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25:129–155, 2005.

[39] M. Pelanti, F. Bouchut, and A. Mangeney. A Roe-type scheme for two-phase shallow granular flows over variable topography. *ESAIM: M2AN*, 42:851–885, 2008.

[40] E. C. Rericha, C. Bizon, M. D. Shattuck, and H. L. Swinney. Modelling phase transition in metastable liquids: application to cavitating and flashing flows. *J. Fluid Mech.*, 607:313–350, 2008.

[41] A. Saurel, F. Petitpas, and R. Abgrall. Modelling phase transition in metastable liquids: application to cavitating and flashing flows. *J. Fluid Mech.*, 607:313–350, 2008.

[42] S. Serna and A. Marquina. Capturing shock waves in inelastic granular gases. *J. Comput. Phys.*, 209:787–795, 2005.

[43] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[44] E. F. Toro. MUSTA: A multi-stage numerical flux. *Appl. Numer. Math.*, 56:1464–1479, 2006.

[45] B. van Leer. Towards the ultimate conservative difference scheme V. A second-order sequel to Godunov's method. *J. Comput. Phys.*, 32: 101–136, 1979.

[46] G. Whitham. *Linear and nonlinear waves*, volume 226. Wiley, New York, NY, USA, 1974.

[47] W. Yong. Singular perturbations of first-order hyperbolic systems with stiff source terms. *J. Differ. Equations*, 155(1):89–132, 1999.

[48] W. Yong. Basic structures of hyperbolic relaxation systems. *P. Roy. Soc. Edinb. A*, 132:1259–1274, 2002.

[49] A. Zein, M. Hantke, and G. Warnecke. Modeling phase transition for compressible two-phase flows applied to metastable liquids. *J. Comput. Phys.*, 229:2964–2998, 2010.