Diss. ETH No. 19875

Multi-Scale Approximation Models for the Boltzmann Equation

A dissertation submitted to
ETH Zürich

for the degree of
Doctor of Sciences

presented by

PETER KAUF
Dipl. Math. ETH
born December 27, 1981
from Felben-Wellhausen, Thurgau
Switzerland

Accepted on the recommendation of
Prof. Dr. Manuel Torrilhon, ETH Zürich, RWTH Aachen University, examiner
Prof. Dr. Rolf Jeltsch, ETH Zürich, co-examiner

August 2011

'Always look on the bright side of life.'

Eric Idle, Monty Python, 1979

Abstract

We are developing mathematical and computational approximation models to the Boltzmann Equation, exploiting its behaviour on multiple physical scales.

We will first give a concise overview of the appearing challenges.

In a second introductory part we will describe physics on various scales inside and around the Boltzmann equation. We will see how a molecular dynamics approach can be coarsened into a statistical description and how the Boltzmann distribution function relates to the macroscopic balance laws of continuum physics.

The third part consists of a mathematical analysis for kinetic models with linear collision operators. There, we will present the two main classical strategies of simplifying the Boltzmann equation, the asymptotic expansion in Knudsen number of Chapman Enskog and Grad's Hermite function approximation. Out of these two classical approaches we will construct a new, 'scale induced' method, based on the ideas in [49]. This method combines the physical accuracy in terms of Knudsen numbers as well as the convenient mathematical properties of Grad. The new strategy is tested numerically in the framework of a 16 discrete velocities model and, together with its mathematically proven convergence and stability properties, exhibits significant advantages to other methods. We will outline how this promising method can be applied also outside the framework of kinetic theory.

In the fourth part, we are developing a computationally motivated approximation to the Boltzmann equation. We will consider the BGK model for the collision term and derive a Galilei-invariant, temperature scaled weak formulation. The transformed Boltzmann distribution is non-linearly approximated by an equilibrium Maxwellian, disbalanced by a general series of perturbation functions. In order to ensure conservation of mass, momentum and energy, a major concern in schemes for the Boltzmann equation, we couple our formulation to the balance laws of continuum physics. Micro- / macro compatibility will be ensured either directly through the perturbation functions or through conditions on their series. The resulting equations evidence a considerable numerical challenge. In the Knudsen number regime of our interest, we will leverage physical diffusion to keep this challenge solvable. Our numerical scheme will be tested on a toy model (Grad's equations for 5 moments in one space and one velocity dimension), before we apply it to a full kinetic shocktube problem. The results for the full kinetic case look promising and will motivate future research for higher dimensional cases that are very interesting for applications.

Zusammenfassung

Wir entwickeln mathematische und numerische Modelle zur Näherung der Boltzmann Gleichung auf verschiedenen physikalischen Skalen. Zuerst besprechen wir in einem Übersichtsteil die dabei auftretenden Schwierigkeiten und Herausforderungen.

In einem zweiten Teil beschreiben wir die physikalischen Prozesse auf verschiedenen Skalen um die Boltzmanngleichung. Wir werden einen molekular dynamischen Ansatz grobkörniger machen und zu einer statistisch physikalischen Beschreibung transformieren. Aus dieser statistischen Beschreibung können wiederum die makroskopischen Bilanzgleichungen der Kontinuumsphysik hergeleitet werden.

Der dritte Teil der vorliegenden Arbeit ist eine mathematische Abhandlung über kinetische Modelle mit linearen Kollisionsoperatoren. Dabei werden wir die beiden klassischen Stragegien untersuchen, um die Boltzmanngleichung zu vereinfachen: Chapmann-Enskog Entwicklung in der Knudsenzahl und den Hermite Funktionen Ansatz von Grad. Mit Hilfe dieser klassischen Methoden werden wir eine neue 'skaleninduzierte' Strategie entwickeln, fussend auf den Ideen in [49]. Diese Strategie kombiniert physikalische Genauigkeit im Mass der Knudsenzahl sowie die günstigen mathematischen Eigenschaften des Ansatzes von Grad. Wir testen diese neue Strategie numerisch im Rahmen eines diskreten Models mit 16 Geschwindigkeiten. Zusätzlich zu den mathematisch beweisbaren Konvergenz- und Stabilitätseigenschaften zeigen sich dabei signifikant bessere Resultate als mit den klassischen Ansätzen von Chapman-Enskog und Grad. Wir werden skizzieren, wie diese vielversprechende Strategie auch ausserhalb der kinetischen Theorie angewendet werden kann.

Im vierten Teil entwickeln wir eine numerisch-physikalisch motivierte Näherung an die Boltzmann Gleichung. Wir werden das BGK Model für den Kollisionsterm verwenden und damit eine Galilei-invariante, temperaturskalierte schwache Formulierung der Boltzmann Gleichung herleiten. Die invariante Boltzmann Verteilung nähern wir nicht-linear mit Hilfe einer Gleichgewichts-Maxwell Verteilung, erweitert durch eine Störungsreihe. Um dabei Massen-, Impuls- und Energieerhaltung zu gewährleisten, was bei herkömmlichen numerischen Methoden für die Boltzmann Gleichung ein Problem darstellt, koppeln wir unsere schwache, invariante Formulierung an die Bilanzgleichungen der Kontinuumsphysik. Dies geschieht mit Hilfe des Wärmeflusses. Die Kompatibilität von mikroskopischen und makroskopischen Grössen werden wir entweder direkt mit der Wahl der entsprechenden Störungsfunktionen oder mit Bedingungen an die gesamte Störungsreihe sicherstellen. Die so entstehenden Gleichungen stellen eine numerische Herausforderung dar. In den Grössenordnungen der Knudsenzahl, die für uns interessant sind, wird physikalische Diffusion stark zur numerischen Lösbarkeit beitragen. Wir werden einen numerischen Lösungsalgorithmus an einem Spielzeugmodell testen (Gradgleichungen für 5 Momente in einer Raum- und Geschwindigkeitsdimension), bevor wir diesen auf das voll-kinetisches Schockwellenproblem anwenden. Die Resultate im voll-kinetischen Fall sehen vielversprechend aus und motivieren weitere Forschungsprojekte für höher dimensionale Fälle, die für die Praxis interessant sind.

Danksagung

Ich habe die vorliegende Dissertation zwischen Oktober 2006 und Juli 2011 am Seminar für angewandte Mathematik der ETH Zürich verfasst. An dieser Stelle möchte ich einigen Leuten danken, die Wesentliches zum Gelingen beigetragen haben.

An erster Stelle danke ich meinen Eltern, Ruth und Werner Kauf, die mich durch eine lange Ausbildungszeit hindurch begleitet, mit Grosszügigkeit unterstützt und mit gutem Willen angespornt haben.

Für den Ansporn wissenschaflticher Art möchte ich meinem Doktorvater Manuel Torrilhon meinen grosssen Dank aussprechen. Seine menschliche, motivierte Art mit Wissenschaft umzugehen, hat mich schon als Diplomstudent beeindruckt. So fiel die Wahl des Dissertationsgebiets im Juli 2006 nicht ersterhand aus fachlichen Überlegungen, sondern mehrheitlich aus Bauchgefühl und Sympathie. Manuel hat mir von Beginn weg grosses Vertrauen entgegengebracht und liess mich in Forschung und Lehre wertvolle Freiheit ausleben. Auf dem reichen Nährboden seines motivierenden Elans und seiner wissenschaftlichen Kompetenz ist die vorliegende Arbeit gewachsen.

Auch Manuel war einmal Doktorand am Seminar für angewandte Mathematik, nämlich bei Rolf Jeltsch. Dieser hat sich – trotz 'Ruhe'stand und vielen neuen Herausforderungen – die Zeit genommen, meine Arbeit zu koreferieren. Dafür möchte ich ihm hier ganz herzlich danken. Von Rolf Jeltsch durfte ich auch neben der Wissenschaft vieles Lernen, einerseits in interessanten Gesprächen über Vergangenheit und Zukunft, aber auch durch die Mithilfe bei ICIAM07, wo ich unter seiner Führung die IT-Infrastruktur organisierte.

Neben den beiden Professoren möchte ich den vielen grossartigen Kolleginnen und Kollegen am SAM danken. Einen wesentlichen Teil meines Promotions- und Assistenz-Alltags haben diese interessant und vielseitig gemacht, sei es bei gemeinsamen Sportaktivitäten, bei den erfrischenden Gesprächen über Mittag oder während der Klausurtagungen im Kloster Disentis. Die Zeit mit ihnen bleibt mir in bester Erinnerung.

Zürich, im August 2011

Peter Kauf

Contents

1 Overview

1.1 Approximating the Boltzmann Distribution - Introductive Overview

The Boltzmann distribution function f statistically describes states of interacting particles. It is a probability density function that depends on space ($\mathbb{R}^D$), time ($\mathbb{R}_+$) and particle velocities ($\mathbb{R}^D$), with $D = 1, 2, 3$. This description is most often used as basic modeling tool in fluid dynamics and extends to various other fields like e.g. debris flow research ([28]), traffic (jam) modeling ([61]) or even political opinion formation ([17]). The evolution of the distribution function f is given through the Boltzmann-equation, which incorporates free movement of particles together with an interaction model.

Even though the Boltzmann distribution models particles statistically, it still contains a large amount of information, incorporated in the 3-dimensional continuous velocity variable. This opens the challenge of finding simplified models that precisely capture the relevant information. These models usually depend on space and time, but only on very few discrete degrees of freedom in the velocity space.

A decisive parameter 'measuring' the amount of information necessary for accurate modeling is the 'Knudsen number' Kn ([51]). Kn is the ratio of mean free path λ (the mean path that a particle moves freely between two interactions) and system size L,

$$\mathrm{Kn} = \frac{\lambda}{L}.$$

As such it is a measure of rarefaction: the higher Kn the less interactions occur on the scale of L.

We will argue that the collisions drive particles into an 'equilibrium' state, thus the higher Kn, the further away the particles are from this equilibrium state.

In equilibrium, or very close to it, we have computationally efficient equations to approximate the solution of the Boltzmann equation in the quantities of interest, which are usually mass density, momentum density and temperature of a fluid.

However, in regimes with larger Knudsen numbers, extended modeling or a direct solution of the Boltzmann equation becomes necessary. Typically we find flows with large Knudsen numbers in atmospheric entry flights (large λ) or in micro devices (small L).

We distinguish 5 regimes of Knudsen numbers, see Fig. 1.1. This distinction is physically motivated through typical effects occuring in the different regimes. The quality of a model will be assessed partially by its computational efficiency and mainly by the ability to capture these specific, so called 'kinetic' effects. A very typical kinetic effect

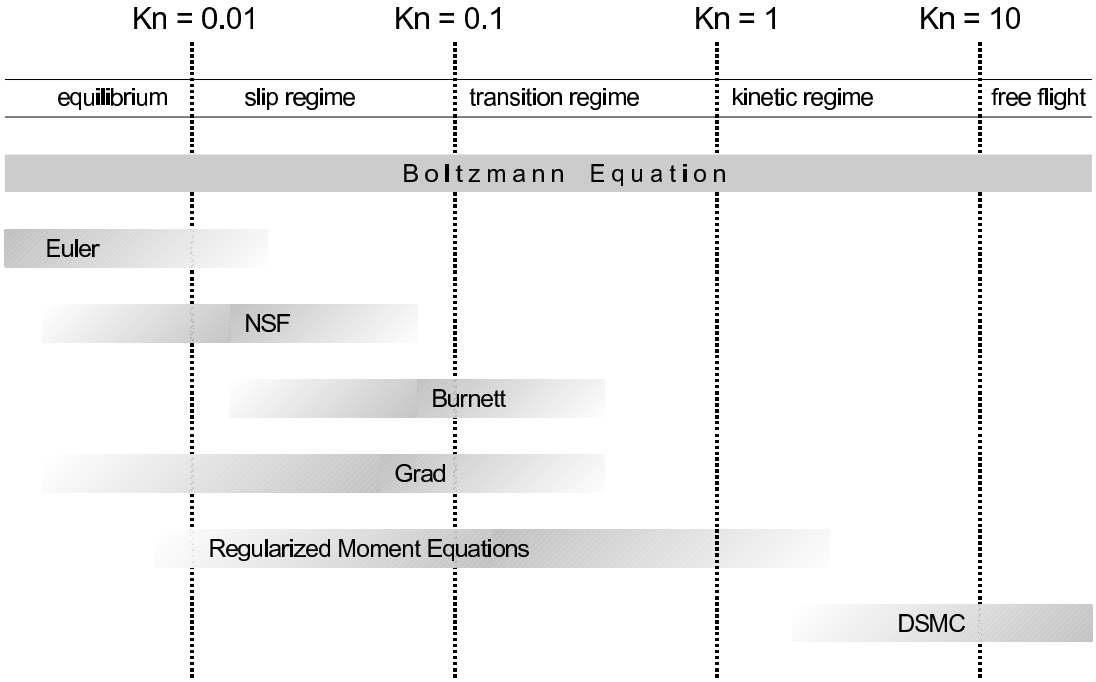


Figure 1.1: Models suitable for various ranges of Knudsen numbers. Around Kn = 1 there is a 'gap'.

is the Knudsen paradox ([57]): the mass flow through a tube decreases with the tube's diameter till the diameter reaches the order of the mean free path λ, then the mass flow increases again. A very interesting application that works with 'thermal transpiration' are Knudsen pumps ([1]). Knudsen pumps have no moving parts, but consist of a narrow channel with a cold and a hot end. In such a set up, particles will drift from the cold end to the hot end.

Knudsen paradox and Knudsen pumps are two physical examples, where the system sizes are in a range that kinetic effects become important.

It is very desirable that a model for the Boltzmann equation captures all the physical effects of its regime of validity. Unluckily, designing experiments whose output could be compared to corresponding simulations - for a verification of quantitative properties of a given model - is challenging.

A very detailed descritption of the models used in the first four regimes in Fig. 1.1 is given in Part 3, we will briefly outline them here.

If the Knudsen number is very low, we have equilibrium and can ommit any higher order effects. The distribution function is an isotropic Gaussian, determined through mass-, velocity- and energy-density, as we will discuss more precisely in Sect. 2.2.3. An appropriate model in this case are the Euler equations.

In the slip and transition regime, we have two classical strategies to derive appropriate models, Grad's moment methods and Chapman-Enskog asymptotic expansions of the distribution function. Grad ([22], [23]) approximates the distribution function f by a Hermite series which translates into a hierarchy of equations for moments of f. This hierarchy is truncated at a certain level. The convergence of this approach towards the true solution of the Boltzmann equation is rather slow, however the resulting equations are typically stable and locally hyperbolic, which makes them mathematically and numerically interesting.

Chapman and Enskog ([14]) expanded f into a series

$$f = f_{equilibrium} + \mathrm{Kn}\, f_1 + \mathrm{Kn}^2\, f_2 + ...$$

and derived equations by taking moments of the Boltzmann equation combined with matching terms of same orders in Kn. To zeroth order, this yields Euler's equations, the first order corresponds to the equations by Navier-Stokes and Fourier and the third order to Burnett equations (see Sect. 3.4.2). Here, the convergence order is more clear, however the equations at higher order become unstable. Chapman-Enskog expansion, as well as Grad's approach, yield a hierarchy of equations that needs to be truncated. This truncation is referred to as the 'closure problem'.

So called 'regularized' moment equations combine the advantages of Grad and Chapman-Enskog each, while avoiding their major drawbacks. They can be derived through a scale induced closure, which will be described in detail in Sect. 3.5. The most prominent set of such regularizued equations are the 'R13' equations, developed by Torrilhon and Struchtrup in [52]. While being linearly stable, they are able to reproduce typical kinetic effects that the Navier-Stockes-Fourier-equations fail to capture, among others the Knudsen paradox.

In principle, so called 'discrete velocity schemes' are available in all the regimes where we apply the Boltzmann equation. They are based upon a point discretization of the velocity-space ([3]). Their drawback is a huge computational cost for an adequate discretization. In real situations where quantitative data should be obtained, these schemes are not used, however with well chosen, relatively inexpensive discretizations ([3]), they can yield very valuable qualitative insight.

The direct simulation Monte Carlo (DSMC) method is only applicable at large Knudsen numbers. It uses states of particles that are evolved by statistically evaluating the collision operator of the Boltzmann equation ([6]). This method is very often applied in practice, a very prominent example is the simulation of the controlled descent of the MIR space station ([37]).

What the approaches for all the regimes have in common is that we approximate our distribution function f through some appropriate f_{model}. Whether f_{model} is (non-linearly) parametrized through statistical moments, or through a choice of discrete velocities, or, very crudely, through just an equilibrium Gaussian, influences our ability to capture physical effects.

By integrating (moments of) f or f_{model} over the velocity space, we obtain macroscopic fields (e.g. mass density, temperature), for details see Sect. 2.2.5. These fields are the final quantities of interest. The velocity integration is sensitive to quite some details of the distribution function, but not to all, and not always to the same ones, depending on the regimes we are interested in. This justifies the unse of models for the Boltzmann equation.

An effective such model should fulfill the following requirements:

1) It should be *accurate enough* to capture the relevant physics in the regime of interest.
2) It should be computationally feasible.
3) It should offer theoretical insight into the physics of the corresponding regime.
4) It should converge to the full Boltzmann solution in a mathematically predictable and numerically observable way.

We will consider points 1), 3) and 4) in Part 3 in detail for several models in the setting of a general linear collision operator. The well known approaches of Chapman-Enskog and Grad will be presented in a mathematically concise way that will allow insight into the approximation principles of these models. Both models, Chapman-Enskog and Grad are computationally feasible, but show some severe instability (Chapman-Enskog) or convergence problems (Grad). In Part 3, we will combine the advantages of Chapman-Enskog and Grad into a new 'scale induced' strategy. We will be able to formulate mathematical theorems about convergence order and stability of this strategy and examplify its numerical approximation qualities in a discrete velocity model. Summarizing, Part 3 is yielding mathematical and theoretical insights in a simplified setting of linear collision operators.

In Part 4, we will illuminate the closure problem from a different angle: we will recast the Boltzmann equation into a very general invariant (weak) form. This will be mathematically much more complex than the original formulation and is motivated through computational efficiency. In this form, we will approximate the distribution function through an equilibrium Gaussian, that is perturbed with some arbitrary functions. This highly non-linear representation will allow us to capture properties of f that many approaches have difficulties to detect. The trade-off for the computational advantages is that we gain only little theoretical insight or mathematically provable approximation theorems with this approach, so points 3) and 4) are out of focus. The in-focus points

1) and 2) will be illustrated by several numerical convergence studies, including performance comparisons to Grad's moment approach.

Summarizing, the approaches in Part 3 and Part 4 have the same goal: constructing a computationally feasible model for the Botlzmann equation that captures relevant features and leaves out redundant information. Whereas Part 3 yields mathematical insights in trade of quantitatively realistic non-linear modeling, Part 4 focusses on numerical feasibility in trade of qualitative insight.

All the algorithms and theoretical considerations presented, both in Part 3 and Part 4 are conceptionally interesting and as such analysed on a conceptional level in one space and one velocity dimensions. Large scale real world engineering applications (3 space and 3 velocity dimensions) have not been implemented in this work - this is left for future projects.

2 Introduction

In this introductory part, we will explore the physics around (Sect. 2.1) and inside the Boltzmann equation (Sect. 2.2).

2.1 Physics around the Boltzmann Equation: Particles and Continuum

The Boltzmann equation incorporates the dynamics of a single particle system in a mathematically challenging continuum limit. Continuum field equations can be derived from it. These two descriptions, particles and continuum, conceptionally encircle the Boltzmann equation and deserve consideration in view of physical symmetries and conservation. We start by summarizing classical results for the movements of point particles in space and time, as essentially described by Sir Isaac Newton in [41] and discuss classical conservation properties in this *microscopic* view. Next we will have a look at continuum mechanics, describing processes in *macroscopic* terms of space and time dependent fields. There we will derive the balance laws of mass, momentum and energy in a macroscopic continuum setting.

2.1.1 How Particles Move

Let us consider N point particles of masses $m_i \in \mathbb{R}_+$, $i = 1, ..., N$ moving in time $t \in \mathbb{R}_+$ with space trajectories $\mathbf{x}_i(t) \in \mathbb{R}^3$, $i = 1, ..., N$. Then *Newton's second law of motion* states that the particle trajectories are determined by the N vectorial equations

$$m_i \ddot{\mathbf{x}}_i(t) = \mathbf{F}_i(\mathbf{x}_1, ..., \mathbf{x}_N, \dot{\mathbf{x}}_1, ..., \dot{\mathbf{x}}_N, t), \quad i = 1, ..., N, \tag{2.1}$$

given the initial positions and velocities

$$\mathbf{x}_1(0) = \mathbf{x}_1^{(0)}, ..., \mathbf{x}_N(0) = \mathbf{x}_N^{(0)}, \quad \dot{\mathbf{x}}_1(0) = \mathbf{v}_1^{(0)}, ..., \dot{\mathbf{x}}_N(0) = \mathbf{v}_N^{(0)}. \tag{2.2}$$

Here $\mathbf{F}_i \in \mathbb{R}^3$ is the *force* acting on particle i. Typically, this force is a superposition of pairwise internal forces $\mathbf{F}_{ik}^{(int)}$ independent of velocity and time and an external force $\mathbf{F}_i^{(ext)}$, independent of the particles $k \neq i$:

$$\mathbf{F}_i(\mathbf{x}_1, ..., \mathbf{x}_N, \dot{\mathbf{x}}_1, ..., \dot{\mathbf{x}}_N, t) = \sum_{k \neq i} \mathbf{F}_{ik}^{(int)}(\mathbf{x}_i, \mathbf{x}_k) + \mathbf{F}_i^{(ext)}(\mathbf{x}_i, \dot{\mathbf{x}}_i, t). \tag{2.3}$$

Newtons third law, *actio = reactio*, postulates that $\mathbf{F}_{ik}^{(int)} = -\mathbf{F}_{ki}^{(int)}$.

Newton's first law (see [41]) defines *inertial systems*:

Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus a viribus impressis cogitur statum illum mutare.

Every body persists in its state of being at rest or of moving uniformly in a straight line, except insofar as it is compelled by an impressed force to change this state.

All inertial systems are linked through the *Galilei transformations*

$$\begin{aligned} &\text{Time shift and time reversion:} \\ &\qquad t' = \lambda t + a, \quad \lambda = \pm 1, a \in \mathbb{R} \\ &\text{Uniform movement, rotated and shifted:} \\ &\qquad \mathbf{x}' = R\mathbf{x} + \mathbf{v}t + \mathbf{b}, \quad R \in O(3); \mathbf{v}, \mathbf{b} \in \mathbb{R}^3 \end{aligned} \tag{2.4}$$

Imposing that the law of motion (2.1) is invariant under Galilei transformations, it follows that the internal forces $\mathbf{F}_{ik}^{(int)}$ act along the line connecting particles i and k with a modulus depending only on their distance:

$$\mathbf{F}_{ik}^{(int)} = f_{ik}\left(|\mathbf{x}_i - \mathbf{x}_k|\right) \frac{\mathbf{x}_i - \mathbf{x}_k}{|\mathbf{x}_i - \mathbf{x}_k|}. \tag{2.5}$$

Integrating the scalar function $f_{ik}(r) = f_{ki}(r)$, we obtain a pair potential $V_{ik}(r)$, such that

$$\mathbf{F}_{ik} = -\nabla_{\mathbf{x}_i} V_{ik}\left(|\mathbf{x}_i - \mathbf{x}_k|\right) \tag{2.6}$$

The total potential V is the sum of all the pair potentials,

$$V(\mathbf{x}_1, ..., \mathbf{x}_N) = \sum_{i<k} V_{ik}\left(|\mathbf{x}_i - \mathbf{x}_k|\right). \tag{2.7}$$

The Galilei-invariance of (2.1) manifests itself in the symmetry of the potential,

$$V(R\mathbf{x}_1 + \mathbf{b}, ..., R\mathbf{x}_N + \mathbf{b}) = V(\mathbf{x}_1, ..., \mathbf{x}_N), \quad R \in O(3), \mathbf{b} \in \mathbb{R}^3. \tag{2.8}$$

Balance Laws

According to (2.8), the potential exhibits 7 symmetries (4 rotational[1], 3 displacement). These symmetries translate into *balance laws* for momentum, angular momentum and energy. Mass balance is trivial in this setting of particles with equal masses m.

Let us state

Definition 2.1.1 (momentum, angular momentum, energy)**.**

- *The momentum of a single particle is* $\mathbf{p}_i = m_i \dot{\mathbf{x}}_i$*. The total momentum of the* N *particles is*
$$\mathbf{P} = \sum_{i=1}^{N} \mathbf{p}_i \tag{2.9}$$
- *The total angluar momentum of the* N *particles is*
$$\mathbf{L} = \sum_{i=1}^{N} \mathbf{x}_i \wedge \mathbf{p}_i \tag{2.10}$$
- *The kinetic energy of a single particle is* $\frac{1}{2} m_i \dot{\mathbf{x}}_i^2$*. The total kinetic energy of the* N *particles is*
$$T = \frac{1}{2} \sum_{i=1}^{N} m_i \dot{\mathbf{x}}_i^2 \tag{2.11}$$

With these definitions at hand, we can derive (microscopic) balance laws

Lemma 2.1.2 (balance laws)**.**

momentum balance:
$$\frac{d}{dt}\mathbf{P} = \sum_{i=1}^{N} \mathbf{F}_i^{(ext)} \tag{2.12a}$$

angular momentum balance:
$$\frac{d}{dt}\mathbf{L} = \sum_{i=1}^{N} \mathbf{x}_i \wedge \mathbf{F}_i^{(ext)} \tag{2.12b}$$

energy balance:
$$\frac{d}{dt}(T+V) = \sum_{i=1}^{N} \mathbf{F}_i^{(ext)} \cdot \dot{\mathbf{x}}_i \tag{2.12c}$$

For the proof of Lemma 2.1.2 we need to exploit the symmetries of the potential, represented in the form (2.6).

Note that (2.12) is a system of ordinary differential equations for total (angular) momentum and energy.

[1] one rotational axis with 3 parameters and one angle describing the rotation around that axis

2.1.2 How Fields Evolve

We have seen in the last subsection that mircoscopically, balance laws are a result of fundamental symmetries, namely Galilei-invariance. A mathematical extension of Galilei-invariance and Newton's laws to fields of macroscopic quantities leads to conservation laws for fields of mass density, momentum density, energy density and angular momentum density.

A proper mathematical description of this extension process is given in [55]. Here, we will summarize the main results.

Let us consider a domain $\Omega_0 \subset \mathbb{R}^3$ at time $t = 0$.[2] The fundamental object of continuum mechanics is a bijective time evolution

$$\Phi^t : \Omega_0 \ni \mathbf{a} \mapsto \mathbf{x} \in \Omega_t = \Phi^t(\Omega_0). \tag{2.13}$$

We can describe the state of a system by two possible sets of variables that are linked through Φ:

$$(x, t) = (\Phi^t(a), t) \qquad \text{(Eulerian description)} \tag{2.14a}$$
$$(a, t) = (\Phi^{-t}(x), t) \qquad \text{(Lagrangian description)} \tag{2.14b}$$

We will mainly use the Eulerian description.

Mass Balance

The mass density is a field $\rho(\mathbf{x}, t)$ in the (Eulerian) space variable $\mathbf{x}$ and time t. Mass balance states that no mass is produced or destroyed as time evolves:

$$\frac{d}{dt} \int_{\Omega_t} \rho(\mathbf{x}, t) dx = 0. \tag{2.15}$$

The Reynolds transport theorem (see [55], p. 19) can be used to commute the (time dependent) integral with the time derivative. Defining the velocity field

$$v_i(\mathbf{x}, t) = \frac{\partial}{\partial t} \Phi_i(\mathbf{a}, t), \quad \mathbf{a} = \Phi^{-t}(\mathbf{x}), \tag{2.16}$$

we obtain the partial differential equation for the mass balance[3]

$$\partial_t \rho + \partial_j (\rho v_j) = 0. \tag{2.17}$$

[2] The choice of the initial time point does not change the physics. This corresponds to (2.4) in Sect. 2.1.1
[3] Notation see App. A.1

Momentum Balance

Generalizing Newton's second law to fields, we balance total momentum with the total forces. Forces can act on volumes (e.g. gravity, electric forces) or surfaces (e.g. friction forces).

Forces acting on surfaces are assumed to depend on time, on one space point $\mathbf{x}$ and on the normal $\mathbf{n}$ to the surface element they are acting on. Then (see [55], p. 47) the surface force $\mathbf{T}(\mathbf{x}, \mathbf{n})$ is a linear function of the direction $\mathbf{n}$,

$$T_i(\mathbf{x}, \mathbf{n}, t) = \sigma_{ij}(\mathbf{x}, t) n_j. \tag{2.18}$$

σ is called the *stress tensor*. The total force acting on our system is then the sum of

$$\mathbf{F}^{\text{volume}}(t) = \int_{\Omega_t} \mathbf{f}(\mathbf{x}, t) dx, \tag{2.19a}$$

$$\mathbf{F}^{\text{surface}}(t) = \int_{\partial\Omega_t} \sigma(\mathbf{x}, t) \cdot \mathbf{n}_{\Gamma_t} d\Gamma_t, \tag{2.19b}$$

$d\Gamma_t$ being the surface element of Ω_t. Balancing, we obtain

$$\frac{d}{dt} \int_{\Omega_t} \rho \mathbf{v}(\mathbf{x}, t) dx = \int_{\Omega_t} \mathbf{f} dx - \int_{\partial\Omega_t} \sigma(\mathbf{x}, t) \cdot \mathbf{n}_{\Gamma_t} d\Gamma_t, \tag{2.20}$$

leading to the partial differential equation of momentum balance

$$\partial_t(\rho v_i) + \partial_j \left(\rho v_i v_j + \sigma_{ij}\right) = f_i. \tag{2.21}$$

Balance of angular momentum is expressed through *symmetry* of the stress tensor σ. This can be seen by balancing the angular momentum analogously to what we have sketched above (see [55], p. 51).

Energy Balance

The energy of a system of monatomic matter consists of kinetic energy and of internal energy (non-motion energy) $e^{(int)}$[4]

$$E(\mathbf{x}, t) = \int_{\Omega_t} \rho(\mathbf{x}, t) \left(e^{(int)}(\mathbf{x}, t) + \frac{1}{2} \mathbf{v}(\mathbf{x}, t)^2 \right) dx \tag{2.22}$$

[4] For polyatomic gases, there is also rotational energy.

E changes due to exterior volume actions (radiation $r(\mathbf{x},t)$, energy of volume forces) and surface actions (heat flux $\mathbf{q}(\mathbf{x},t)$, power of surface forces[5]):

$$\frac{d}{dt}E = \int_{\Omega_t} r dx + \int_{\Omega_t} \mathbf{f}\cdot\mathbf{v} dx - \int_{\partial\Omega_t} \frac{1}{2}\mathbf{q}\cdot\mathbf{n}_t d\Gamma_t - \int_{\partial\Omega_t} \mathbf{v}\cdot\sigma\mathbf{n}_t d\Gamma_t. \tag{2.23}$$

This integral balance translates to the partial differential equation for (internal) energy conservation as

$$\partial_t\left(\rho e^{(int)} + \frac{1}{2}\rho v_k^2\right) + \partial_j\left(\left(\rho e^{(int)} + \frac{1}{2}\rho v_k^2\right)v_j + v_k\sigma_{kj} + \frac{1}{2}q_j\right) = f_k v_k + r \tag{2.24}$$

For ideal gases[6], we can reformulate (2.24) in terms of temperature (in energy units): through the ideal gas equation, we can relate pressure p, volume V and temperature T as

$$p = \rho\frac{k}{m}T, \tag{2.25}$$

where k is the Boltzmann constant and m the mass of the particles. Pressure on the other hand relates to internal energy and dimension d as

$$p = \frac{2}{d}\rho e^{(int)}. \tag{2.26}$$

Writing temperature in energy units as $\theta := \frac{k}{m}T$, we conclude that $e^{(int)} = \frac{d}{2}\theta$ (see [51]).

With this transform, (2.27) becomes

$$\partial_t\left(\rho\frac{d}{2}\theta + \frac{1}{2}\rho v_k^2\right) + \partial_j\left(\left(\rho\frac{d}{2}\theta + \frac{1}{2}\rho v_k^2\right)v_j + v_k\sigma_{kj} + \frac{1}{2}q_j\right) = f_k v_k + r \tag{2.27}$$

We will see in Sect. 2.2.5 that furthermore $\theta = \frac{1}{\rho}\operatorname{trace}\sigma$, with which we could reformulate (2.27) even further.

Closure Problem

Out of first principles, we have obtained the following system of partial differential equations, called *balance laws* of mass, momentum and energy,

[5] The prefactor $\frac{1}{2}$ to q in the balance equation is for mathematical convenience.
[6] A gas is ideal if the collision time of two particles is short compared to the free flight time, see [51].

$$\partial_t \rho + \partial_j (\rho v_j) = 0 \tag{2.28a}$$

$$\partial_t (\rho v_i) + \partial_j (\rho v_i v_j + \sigma_{ij}) = f_i \tag{2.28b}$$

$$\partial_t \left(\rho \frac{d}{2}\theta + \frac{1}{2}\rho v_k^2 \right) + \partial_j \left(\left(\rho \frac{d}{2}\theta + \frac{1}{2}\rho v_k^2 \right) v_j + v_k \sigma_{kj} + \frac{1}{2} q_j \right) = f_k v_k + r \tag{2.28c}$$

$$\sigma_{ij} - \sigma_{ji} = 0 \tag{2.28d}$$

Typically, we are interested in ρ, $\mathbf{v}$ and θ, while $\mathbf{f}$ and r are given, together with some boundary and initial conditions. But what about $\mathbf{q}$ and σ? Naively, there are more variables than equations - the system is not *closed* and can therefore not have a unique solution.

This is in intuitive accordance with physics, since we did not put any information about material properties into the balance laws. They are valid for fluids or solid bodies, for metals, water, honey, blood, gases or even plasmas. Obviously, the physical behaviour of these materials is very different.

The material properties enter through modeling assumptions. These modeling assumptions typically yield additional relations between the variables. One very prominent example of such relations are the Euler assumptions

$$\sigma_{ij} = \frac{1}{d} p \delta_{ij} = \frac{1}{3} \rho \theta \delta_{ij}, \quad q_i = 0. \tag{2.29}$$

These assumptions lead to the Euler equations of fluid dynamics (see [51], p. 59).

We will soon see that the construction of such modeling assumptions or *closures* for the balance laws (2.28) is a delicate issue, especially if we consider *rarefied gases*.

2.2 Physics in the Boltzmann Equation

In the last section we have seen physical descriptions on a microscopic particle scale as well as macroscopic field equations that do not involve single particles anymore. The Boltzmann equation is something in between: it considers particles in the sense of a probability distribution of particle velocities in every space time point. Its time evolution is based on two ingredients: collisions of particles and free flight of particles.

We are deriving the Boltzmann equation in the next two subsections. We will start with describing particle collisions on the microscopic level in Sect. 2.2.1. Then we will sketch a mathematically complicated continuum limit in Sect. 2.2.2, explaining necessary physical assumptions on the system of colliding particles. In Sect. 2.2.3, we will derive quantities that are invariant under collisions and discuss some implications of the Boltzmann description like time irreversibility and entropy. In Sect 2.2.4 we are presenting some simplified collision models (Broadwell, BGK), and will finally connect the Boltzmann equation to the balance laws of continuum physics in Sect. 2.2.5.

2.2.1 Microscopic Particle Collisions - Atomistic Billiards

Before we proceed to a statistical description, we will consider again particles. We will derive what happens in a two-particle collision, using first principle conservation of momentum, angular momentum and energy. The full understanding of the collision process is not crucial for this thesis, since we will consider simplified collision models later on. However, the collision dynamics are the key ingredient to the Boltzmann equation and as such they allow for various applications of the Boltzmann equation to other fields (see Part 1). Therefore, we give the collision process some consideration.

We will start by exploiting all the symmetries that two-body interaction offers. Then we will compute cross sections of the scattering process in general and specifically for the case of hard spheres. We will be following [24] and [51].

Integration of two Body Interaction

We consider two particles at positions $\mathbf{x}_1$ and $\mathbf{x}_2$ with equal masses m_{12} in a potential $V(|\mathbf{x}_1 - \mathbf{x}_2|)$.

The equations of motion, (2.1), are

$$\ddot{\mathbf{x}}_1 = -\frac{1}{m_{12}} \nabla_{\mathbf{x}_1} V(|\,\mathbf{x}_1 - \mathbf{x}_2\,|) \tag{2.30a}$$

$$\ddot{\mathbf{x}}_2 = -\frac{1}{m_{12}} \nabla_{\mathbf{x}_2} V(|\,\mathbf{x}_1 - \mathbf{x}_2\,|) \tag{2.30b}$$

Introducing center of mass coordinates $\mathbf{X} = \mathbf{x}_1 + \mathbf{x}_2$, $\mathbf{x} = \mathbf{x}_1 - \mathbf{x}_2$ as well as $m = \frac{m_{12}}{2}$, (2.30) translates to[7]

$$\ddot{\mathbf{X}} = \mathbf{0} \tag{2.31a}$$

$$m\ddot{\mathbf{x}} = -\nabla_{\mathbf{x}} V(|\,\mathbf{x}\,|) \tag{2.31b}$$

The center of mass moves uniformly in a given direction, so we find a Galilei transform where it is at rest in $\mathbf{0}$. From now on we therefore assume without loss of generality that $\mathbf{X}(t) = \mathbf{0}$ for all t. For simplicity, we can also specify te coordinates such that the total angular momentum satisfies $L = (0, 0, 1)$.

With the two conserved quantities

$$\text{Energy:} \qquad E = \frac{1}{2} m\dot{\mathbf{x}}^2 + V(|\,\mathbf{x}\,|) \tag{2.32a}$$

$$\text{Angular momentum:} \qquad \mathbf{L} = m\,\mathbf{x} \wedge \dot{\mathbf{x}} := (0, 0, 1) \tag{2.32b}$$

[7] This also works for two different masses m_1, m_2. m is then chosen such that $\frac{1}{m} = \frac{1}{m_1} + \frac{1}{m_2}$.

we see that the trajectory $\mathbf{x}(t)$ is bound to the plane perpendicular to $\mathbf{L}$. In this plane, we use a polar coordinate basis with

$$\begin{aligned} \mathbf{e}_{\mathrm{r}} &= (\cos\phi, \sin\phi, 0) \qquad \text{(radial direction)} \\ \mathbf{e}_{\phi} &= (-\sin\phi, \cos\phi, 0) \quad \text{(tangential direction)} \end{aligned}$$

such that $\mathbf{x}(t) = r(t)\,\mathbf{e}_{\mathrm{r}}(t)$ and $\dot{\mathbf{x}} = \dot{r}\,\mathbf{e}_{\mathrm{r}} + r\dot{\phi}\,\mathbf{e}_{\phi}$.

In these coordinates, the constant modulus of the angular momentum is expressed as $l := |\,\mathbf{L}\,| = mr^2\dot{\phi}$. This allows to restate (2.31) in terms of E and l, using that $\dot{\phi} = \frac{l}{mr^2}$

$$\frac{1}{2}m\dot{r}^2 = E - \frac{l^2}{2mr^2} - V(r), \tag{2.33}$$

which can be integrated to

$$t(r) - t(r_0) = \pm \int_{r_0}^{r} \frac{1}{\sqrt{\frac{2}{m}\left(E - \frac{l^2}{2m\tilde{r}^2} - V(\tilde{r})\right)}} d\tilde{r} \tag{2.34a}$$

$$\phi(r) - \phi(r_0) = \pm \int_{r_0}^{r} \frac{l}{\tilde{r}^2\sqrt{2m\left(E - \frac{l^2}{2m\tilde{r}^2} - V(\tilde{r})\right)}} d\tilde{r}. \tag{2.34b}$$

To obtain the second equaqtion, we used $\frac{d\phi}{dr} = \dot{\phi}\frac{1}{\dot{r}} = \frac{l}{mr^2}\frac{1}{\dot{r}}$ together with (2.33).

Two Body Scattering Problem

With the help of (2.34), we will now discuss solutions of (2.30) with *unbounded trajectories*. We will see in Sect. 2.2.2 that stable orbits are not of interest, since we consider gases, where the time of interaction is 'short' compared to the time of free flight (ideal gas assumption).

Furthermore, we assume that the potential V is local, i.e.

$$V(r) \to 0 \text{ for } r \to \pm\infty. \tag{2.35}$$

Potentials that do not satisfy this assumption are unphysical.

In Fig. 2.1, we see the collision process: two particles are being infinitely apart from each other before the collision and at a relative speed $\dot{\mathbf{x}}(-\infty)$, determining the incoming (normalized) asymptote $\mathbf{e}_{in}$. The resting center of mass together with $\mathbf{e}_{in}$ determines the plane of motion. The coordinates in this plane are chosen such that $\mathbf{e}_r(\phi = 0) = \mathbf{e}_{in}$ and the origin is the center of mass. The conserved energy is $E = \frac{1}{2}m\dot{\mathbf{x}}^2$. The parameter $\mathbf{b}$ relates to the angular momentum. It describes the aberration, if $\mathbf{b} = 0$, the particles will experience a 'frontal collision'. $\mathbf{b}$ is perpendicular to $\mathbf{e}_{in}$ and lies in the plane

of motion. The conserved angular momentum, perpendicular to the plane of motion becomes $L = m\mathbf{x} \wedge \dot{\mathbf{x}} = m\mathbf{b} \wedge \dot{\mathbf{x}}$, and its modulus $l = |\mathbf{b}|\sqrt{2mE}$. $\mathbf{e}_{out}$ is the outgoing asymptote.

The quantity of interest is the collision angle $\theta = \theta(b, V, E)$. According to (2.34b), with

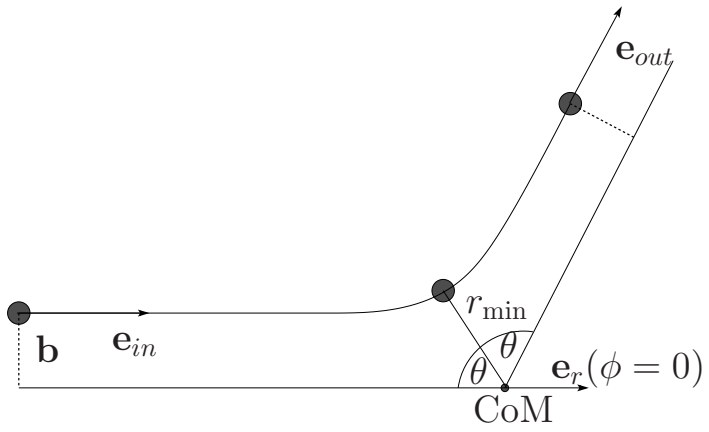


Figure 2.1: Binary collision in the plane perpendicular to $\mathbf{L}$.

the replacements of $l = |\mathbf{b}|\sqrt{2mE}$, we get

$$\theta = \pm \int_{r_{min}}^{\infty} \frac{|\mathbf{b}|}{\tilde{r}^2 \sqrt{1 - \frac{|\mathbf{b}|^2}{\tilde{r}^2} - \frac{V(\tilde{r})}{E}}} d\tilde{r}, \qquad r_{min}: \quad 1 - \frac{|\mathbf{b}|^2}{r_{min}^2} - \frac{V(r_{min})}{E} \stackrel{!}{=} 0. \tag{2.36}$$

We will evaluate this integral for the special cases of a hard sphere potential in the next subsection.

If E and $\mathbf{e}_{in}$ are given, the aberration $\mathbf{b}$ determines the outgoing asymptote $\mathbf{e}_{out}$. Mathematically, this relation allows for a mapping of the polar element $\Delta s := |\mathbf{b}|\Delta b \Delta\psi$ to the S^2 surface element $\Delta\Omega := \sin(\chi)\Delta\chi\Delta\psi$, with $\chi = \pi - 2\theta$, as shown in Fig. 2.2. The angle ψ corresponds to deviations from $\mathbf{b}$, see the left side of Fig. 2.2. In the limit, the ratio $\frac{\Delta s}{\Delta\Omega}$ defines the differential cross section

$$\frac{ds}{d\Omega} = \left| \frac{|\mathbf{b}|}{\sin(\chi)} \left(\frac{d\chi}{d|\mathbf{b}|} \right)^{-1} \right|. \tag{2.37}$$

Integrating (2.37) over the whole sphere, yields the total area s, that contributes to non-trivial collisions. This area s is called 'cross section',

$$s = \int_{\psi=0}^{2\pi} \int_{\chi=0}^{\pi} \frac{ds}{d\Omega} d\Omega. \tag{2.38}$$

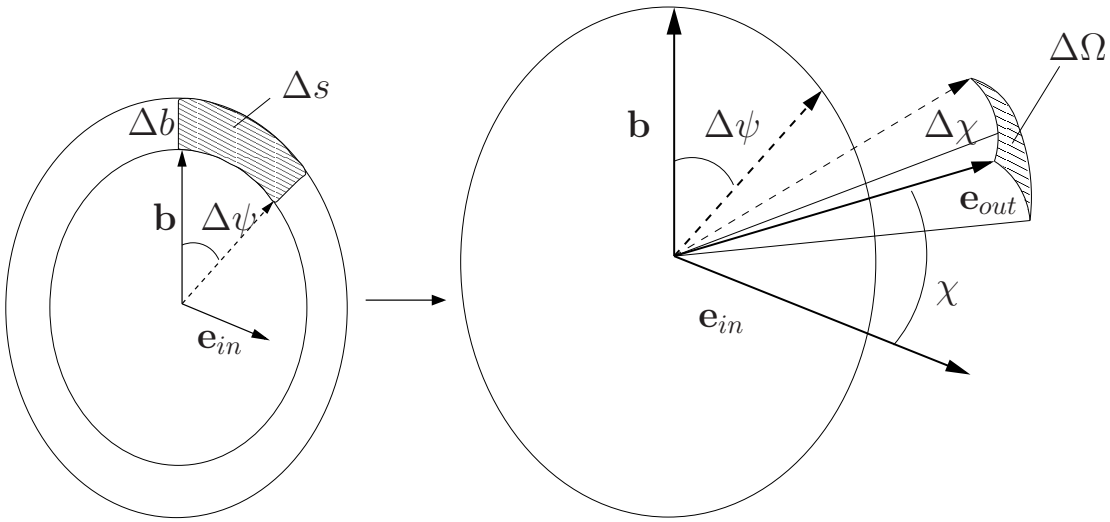


Figure 2.2: Mapping of polar element Δs to S^2-surface element $\Delta\Omega$. The left figure corresponds to the (3-dimensional) pre-collision situation, the right one to the post-collision situation.

Two body scattering for hard spheres

In the case of hard spheres with diameter d, we have an interaction potential (see [51])

$$\phi_{hs}(r) = \begin{cases} 0 & r > d \\ \infty & r \leq d \end{cases} \tag{2.39}$$

With this potential, the collision angle (2.36) can be determined analytically as

$$\theta(\mathbf{b}, E) = \arcsin\left(\frac{|\mathbf{b}|}{d}\right). \tag{2.40}$$

Note, that collisions only occur if $b < d$.

The differential cross section becomes

$$\frac{ds}{d\Omega} = \frac{d^2}{4}, \tag{2.41}$$

leading to a total cross section of $s = \pi d^2$. This would have been clear directly from the model, since hard spheres collide with each other only if their distance is smaller than their radius.

Geometry of Collisions

We can summarize the description of collisions in a more geometric and intuitive way. We call the given pre-collision velocities $\mathbf{c}_1$ and $\mathbf{c}_2$, and the unknown post collision velocities $\tilde{\mathbf{c}}_1$ and $\tilde{\mathbf{c}}_2$. From molecular dynamics, we assume the conservations of momentum and energy through a collision (with all equal particle masses m), namely

$$\mathbf{c}_1 + \mathbf{c}_2 = \tilde{\mathbf{c}}_1 + \tilde{\mathbf{c}}_2 \tag{2.42a}$$

$$\mathbf{c}_1^2 + \mathbf{c}_2^2 = \tilde{\mathbf{c}}_1^2 + \tilde{\mathbf{c}}_2^2. \tag{2.42b}$$

Since the precollision velocities $\mathbf{c}_1$ and $\mathbf{c}_2$ are given, this yields a total of 4 equations for $4 \times 3 - 6 = 6$ parameters (in 3 dimensions), so we are left with a two-dimensional solution manifold (1 dimensional in 2 dimensions, no freedom in 1 dimension).

There is a very nice geometrical representation of these solution manifolds, the Thalescircle (2 dimensions) or the Thalessphere (3 dimensions). We construct it as follows:

1. Subtract $\mathbf{c}_1$ from all the velocities, and define new velocities $\mathbf{c} := \mathbf{c}_2 - \mathbf{c}_1$, and $\hat{\mathbf{c}}_i := \tilde{\mathbf{c}}_i - \mathbf{c}_1$, $i = 1, 2$.
2. Draw a circle / sphere Θ of radius $|\mathbf{c}|/2$ around the point $\mathbf{c}/2$ (Thalescircle / Thalessphere).
3. All possible combinations of shifted post collision velocities $\hat{\mathbf{c}}_1$ and $\hat{\mathbf{c}}_2$ that have their endpoints on the circle / sphere solve the equations (2.42).[8].
4. Shift the velocities back by adding $\mathbf{c}_1$.

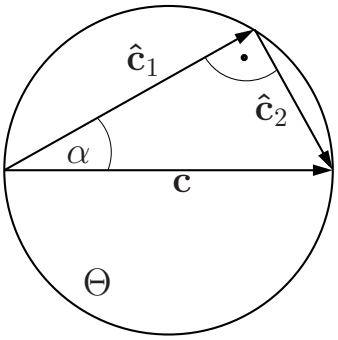


Figure 2.3: Thalescircle.

[8] This is due to the construction of the Thalescircle, vectors starting at 0 and $\mathbf{c}$ meeting in the same endpoint on the circle are orthogonal. This can be checked easily in $\mathbb{R}^2$: take a circle centered around 0 with radius 1, denote the points on the circle as $\{\begin{pmatrix} x \\ y \end{pmatrix} : x^2 + y^2 = 1\}$ and observe that the scalar product between the two vectors $\begin{pmatrix} x+1 \\ y \end{pmatrix}$ and $\begin{pmatrix} x-1 \\ y \end{pmatrix}$ is zero.

The above recipe describes a 2, resp. 1 parameter mapping

$$(\mathbf{c}_1, \mathbf{c}_2) \mapsto (\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2) = T^{2\mathrm{d}}_{\alpha}(\mathbf{c}_1, \mathbf{c}_2) \tag{2.43a}$$

$$(\mathbf{c}_1, \mathbf{c}_2) \mapsto (\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2) = T^{3\mathrm{d}}_{\alpha,\beta}(\mathbf{c}_1, \mathbf{c}_2). \tag{2.43b}$$

Geometrically it can be verified that T is an involution, i.e. $T^{2\mathrm{d}}_{\alpha} \circ T^{2\mathrm{d}}_{\alpha} = id_2$ and $T^{3\mathrm{d}}_{\alpha,\beta} \circ T^{3\mathrm{d}}_{\alpha,\beta} = id_3$.

The representation of post-collision velocities in dependence on precollision velocities with the Thalessphere is important computationally for the DSMC method (see [6] and the end of Sect. 2.2.2).

Note here that we can also model an inelastic collision process (e.g. for granular media). In that case, we would not have energy conservation through the collision, but some (to be specified) energy dissipation, see e.g. [4].

2.2.2 In Between Particles and Continuum: the Boltzmann Distribution

Ludwig Boltzmann (1844-1906, see [12], [13] for mathematical and historical facts about Ludwig Boltzmann) considers a model in between the continuum picture of Sect. 2.1.2 and the particle description of Sect. 2.1.1.

The Liouville Equation

To every space-time point $(\tau, \xi) \in \mathbb{R}_+ \times \mathbb{R}^3$, we assign a probability density function[9]

$$\mathbb{R}^3 \ni \mathbf{c} \mapsto \mathcal{F}(\tau, \xi, \mathbf{c}) \in \mathbb{R}_+. \tag{2.44}$$

This description requires a limit of the particle system, where the number of particles $N \to \infty$ and the 'diameter' or interaction radius of the particles $R \to 0$.

Physically, (2.44) means that $\mathcal{F}(\tau, \xi, \mathbf{c})\Delta\xi\Delta\mathbf{c}$ gives the number of particles at time τ in a small cuboid centered around ξ of sidelengths $\Delta\xi$, that have speeds in the cuboid

$$[c_1 - \frac{1}{2}\Delta c_1, c_1 + \frac{1}{2}\Delta c_1] \times [c_2 - \frac{1}{2}\Delta c_2, c_2 + \frac{1}{2}\Delta c_2] \times [c_3 - \frac{1}{2}\Delta c_3, c_3 + \frac{1}{2}\Delta c_3]. \tag{2.45}$$

In order to determine the evolution of $\mathcal{F}$, we consider particle trajectories $(\tau, \xi(\tau), \mathbf{c}(\tau))$, with $\mathbf{c}(\tau) = \dot{\xi}(\tau)$ and a (sufficiently smooth) evolution operator[10]

$$\Phi^{\tau} : \mathbb{R}^3 \times \mathbb{R}^3 \ni (\xi(0), \mathbf{c}(0)) \mapsto \Phi^{\tau}(\xi(0), \mathbf{c}(0)) = (\xi(\tau), \mathbf{c}(\tau)) \tag{2.46}$$

[9] For a more mathematically profound derivation see [2]: while we are assuming that the probability measure is absolutely continous with respect to the Lebesgue measure (with density $\mathcal{F}$), Babovsky in [2] does the analysis for more general probability measures.

[10] The mathematical existence of such an operator as a limit from a multiparticle system is not at all trivial. Details can be found in [33].

We *postulate* conservation of probability for any domain $\Omega_0 \subset \mathbb{R}^3 \times \mathbb{R}^3$ as

$$\frac{d}{d\tau} \int_{\Phi^\tau(\Omega_0)} \mathcal{F}(\tau, \xi, \mathbf{c}) d\xi dc = 0, \tag{2.47}$$

which, using Reynolds transport theorem ([55], p. 19), translates to the *Liouville equation*

$$\partial_\tau \mathcal{F}(\tau, \xi, \mathbf{c}) + \partial_{\xi_i} \mathcal{F}(\tau, \xi, \mathbf{c}) c_i + \partial_{c_i} \mathcal{F}(\tau, \xi, \mathbf{c}) \frac{F_i}{m} = 0. \tag{2.48}$$

Comparable to the splitting in (2.3), the forces $\mathbf{F} = m\dot{\mathbf{c}}$ can be split into an interior interparticle force $\mathbf{F}^{int}(\tau, \xi, \mathbf{c})$ and an exterior contribution $\mathbf{F}^{ext}(\tau, \xi)$. This is leaves us with a reformulation of the Liouville equation as

$$\partial_\tau \mathcal{F}(\tau, \xi, \mathbf{c}) + \partial_{\xi_i} \mathcal{F}(\tau, \xi, \mathbf{c}) c_i + \partial_{c_i} \mathcal{F}(\tau, \xi, \mathbf{c}) \frac{F_i^{ext}}{m} = -\partial_{c_i} \mathcal{F}(\tau, \xi, \mathbf{c}) \frac{F_i^{int}}{m}. \tag{2.49}$$

From Liouville to Boltzmann - Coarse Graining

The Liouville equation (2.48) offers a fine scale description of particle interactions, the scales (ξ, τ) are resolving collision distances and times.

Since this scale is too fine, we do 'coarse graining' and average $\mathcal{F}$ to

$$f(\mathbf{x}, t, \mathbf{c}) = \frac{1}{4\Delta\xi\Delta\tau} \int_{\mathbf{x}-\Delta\xi}^{\mathbf{x}+\Delta\xi} \int_{t-\Delta\tau}^{t+\Delta\tau} \mathcal{F}(\xi, \tau, \mathbf{c}) d\tau d\xi, \tag{2.50}$$

This new scale $(\mathbf{x}, t)$ does not resolve the collisions anymore. The description $f(\mathbf{x}, t, \mathbf{c})$ connects to the Boltzmann-Grad limit, where we let the particle number $N \to \infty$ and the potential range $d \to 0$.[11]

Expressions for the interior particle forces out of the collision process for $\mathcal{F}$ could be derived. We will proceed a bit differently and take a limit directly from a particle description of the collisions to the coarse scale distribution $f(\mathbf{x}, t, \mathbf{c})$. Like this, collissions will produce a gain (G) and a loss term (L) to the evolution of f at speed $\mathbf{c}$,

$$\partial_t f(\mathbf{x}, t, \mathbf{c}) + \partial_i c_i f(\mathbf{x}, t, \mathbf{c}) = G(\text{collisions to } \mathbf{c}) - L(\text{collisions away from } \mathbf{c}) \tag{2.51}$$

For the computation of the gain and loss terms for (2.51), we consider only two particle collisions. Ternary and higher order collisions are assumed to be negligible (dilute gas assumption, see [33]).

[11] While $N \to \infty$ and $d \to 0$, $N \to \infty$ the Boltzmann Grad limit requires $Nd^2 = const.$ For details see [33].

Since the coarse graining idea of letting $d \to 0$ includes the ideal gas assumption that requires d to be small with respect to the mean free flight path[12], it is sufficient to consider only interactions of particles with unbounded trajectories. This has been done in Sect. 2.2.1. With bounded trajectories, the particles would remain inside their potential radius for comparably long times and would therefore violate the ideal gas assumption.

Given $\mathbf{c}$ and $\mathbf{c}_2$ the speeds of the two colliding particles, we can construct the situation on the left of Fig. 2.2. Since in the Boltzmann Grad limit we will let the potential radius go to zero, we will not distinguish the space points of the particles. We furthermore assume that the potential range is very short (see above), so we can also assume that the particle distribution function changes only at the very moment of the collision.

Stosszahlansatz

The number of center of masses of particles passing through the shaded area in the left picture of Fig. 2.2 within a time intervall Δt is

$$\underbrace{f^{(2)}(\mathbf{x}, t, \mathbf{c}, \mathbf{c}_1)\Delta c \Delta c_1}_{\#\text{particles per space volume}} \underbrace{|\mathbf{c} - \mathbf{c}_1|\Delta t \Delta s}_{\text{space volume}}. \tag{2.52}$$

Here, $f^{(2)}$ is the two particle velocity distribution. The assumption of molecular chaos ('Stosszahlansatz') postulates that the particles are independent, meaning that $f^{(2)}$ factorizes into

$$f^{(2)}(\mathbf{x}, t, \mathbf{c}, \mathbf{c}_1) = f(\mathbf{x}, t, \mathbf{c}) f(\mathbf{x}, t, \mathbf{c}_1) \tag{2.53}$$

The validity of this assumption has been mathematically proven by Lanford in [33] for short time-intervalls and given that the initial velocity distribution factorizes. The 'Stosszahlansatz' is a very strong assumption, it decorrelates the particle velocities and we will argue later that this is what destroys time reversibility of the Boltzmann equation.

With the molecular chaos assumption, (2.52) becomes

$$f(\mathbf{x}, t, \mathbf{c}) f(\mathbf{x}, t, \mathbf{c}_1) \Delta c \Delta c_1 |\mathbf{c} - \mathbf{c}_1| \Delta t \Delta s. \tag{2.54}$$

Derivation of Gain and Loss Operators

With the differential crosssection $\frac{\Delta s}{\Delta \Omega}(\chi, \psi)$ (see (2.37)), we can rewrite (2.54) in a post-collision form as

$$f(\mathbf{x}, t, \mathbf{c}_1) \Delta c_1 f(\mathbf{x}, t, \mathbf{c}) \Delta c |\mathbf{c} - \mathbf{c}_1| \Delta t \frac{\Delta s}{\Delta \Omega} \overbrace{\sin(\chi) \Delta \chi \Delta \psi}^{\Delta \Omega}. \tag{2.55}$$

[12] The ideal gas assumption can also be interpreted as particles having a potential energy that is much smaller than the mean kinetic energy, or that collision times are very short compared to free flight times (see [51]).

To obtain the loss term L in (2.51), we need a number of collisions per unit time, so we divide (2.55) by Δt.

Since collisions to any velocities lead to a loss in $\mathbf{c}$, we integrate over all possible collision results and partner velocities $\mathbf{c}_1$. Normalized by dc, this integration yields

$$L(\mathbf{c},\mathbf{x},t) = \int_{\mathbf{c}_1\in\mathbb{R}^3}\int_{\psi\in[0,2\pi]}\int_{\chi\in[-\pi/2,\pi/2]} f(\mathbf{x},t,\mathbf{c}_1)f(\mathbf{x},t,\mathbf{c}) \cdot |\mathbf{c}_1-\mathbf{c}|\frac{ds}{d\Omega}(\chi,\psi)\sin(\chi)d\chi d\psi dc_1. \tag{2.56}$$

For the gain term $G(\mathbf{c})$, we exploit the involution property of the collision mapping $T_{\alpha,\beta}$ as defined in (2.43). For velocities $(\mathbf{c},\mathbf{c_1})$ and $(\tilde{\mathbf{c}}_1,\tilde{\mathbf{c}}_2)$ that are linked through a collision, we can show that

$$\mathcal{S}(\mathbf{c},\mathbf{c_1},\tilde{\mathbf{c}}_1,\tilde{\mathbf{c}}_2) = \mathcal{S}(\tilde{\mathbf{c}}_1,\tilde{\mathbf{c}}_2,\mathbf{c},\mathbf{c_1}), \tag{2.57}$$

where

$$|\mathbf{c}_1-\mathbf{c}|\frac{ds}{d\Omega}(\chi,\psi)\sin(\chi) := \mathcal{S}(\underbrace{\mathbf{c},\mathbf{c}_1}_{\text{precollision}}, \underbrace{\tilde{\mathbf{c}}_1(\mathbf{c}_1,\mathbf{c},\chi,\psi),\tilde{\mathbf{c}}_2(\mathbf{c}_1,\mathbf{c},\chi,\psi)}_{\text{postcollision}}). \tag{2.58}$$

With this, the gain term becomes

$$G(\mathbf{c},\mathbf{x},t) = \int_{\mathbf{c}_1\in\mathbb{R}^3}\int_{\psi\in[0,2\pi]}\int_{\chi\in[-\pi/2,\pi/2]} f(\mathbf{x},t,\tilde{\mathbf{c}}_1(\mathbf{c},\mathbf{c}_1,\chi,\psi))f(\mathbf{x},t,\tilde{\mathbf{c}}_2(\mathbf{c},\mathbf{c}_1,\chi,\psi)) \cdot |\mathbf{c}_1-\mathbf{c}|\frac{ds}{d\Omega}(\chi,\psi)\sin(\chi)d\chi d\psi dc_1. \tag{2.59}$$

Balancing gain and loss, the full Boltzmann equation with a general particle interaction potential is

$$\partial_t f(\mathbf{x},t,\mathbf{c}) + \partial_i c_i f(\mathbf{x},t,\mathbf{c}) = \int_{\mathbf{c}_1\in\mathbb{R}^3}\int_{\psi\in[0,2\pi]}\int_{\chi\in[-\pi/2,\pi/2]} \{f(\mathbf{x},t,\tilde{\mathbf{c}}_1(\mathbf{c},\mathbf{c}_1,\chi,\psi))f(\mathbf{x},t,\tilde{\mathbf{c}}_2(\mathbf{c},\mathbf{c}_1,\chi,\psi)) - f(\mathbf{x},t,\mathbf{c}_1)f(\mathbf{x},t,\mathbf{c})\} \cdot |\mathbf{c}_1-\mathbf{c}|\frac{ds}{d\Omega}(\chi,\psi)\sin(\chi)d\chi d\psi dc_1. \tag{2.60}$$

Observe here that a numerical evaluation of the collision integral requires extensive computational ressources: for every $\mathbf{c}\in\mathbb{R}^3$, we have to compute a 5-dimensional integral. In practice, Monte Carlo methods are used to stochastically approximate this integral with the help of the geometric representation that we have seen in Sect. 2.2.1.[13]

[13] This is not yet the 'Direct Simulation Monte Carlo' (DSMC) method, see Sect. 2.2.4.

To finish this derivation sketch, let us summarize again the assumptions:

- Ideal gas: The collision time is negligible compared to free flight time, in other words: the potential has a short range compared to the mean free path.
- Dilute gas: we neglect all ternary or higher order collisions.
- Molecular Chaos / Stosszahlansatz: Two-particle distribution functions factorize.

In the frame of these assumptions, the coarse graining Boltzmann-Grad limit can be taken and leads to the Boltzmann equation (2.60).

2.2.3 Collision Invariants, Equilibrium and Entropy

From now on, we will abbreviate the collision integral in (2.60) as

$$\mathcal{C}[\tilde{f}\tilde{f}_1 - f f_1](\mathbf{x}, t, \mathbf{c}) := \int_{\psi\in[0,2\pi]} \int_{\chi\in[-\pi/2,\pi/2]} \quad ... \quad d\chi d\psi. \tag{2.61}$$

Independent of the interaction potential, conservation of mass, momentum and energy for the collisions on the molecular level translate into collision invariants $\phi(\mathbf{c})$ of the Boltzmann collision kernel $\mathcal{C}$,

$$\int \phi(\mathbf{c})\mathcal{C}[\tilde{f}\tilde{f}_1 - f f_1](\mathbf{x}, t, \mathbf{c})dc = 0 \tag{2.62}$$

The property (2.62) can be proven for $\phi(\mathbf{c}) = 1$, $\phi(\mathbf{c}) = \mathbf{c}$ and $\phi(\mathbf{c}) = \mathbf{c}^2$ by using the microscopic relations (2.42).

One can show that, a collision invariant ϕ satisfies

$$\phi(\mathbf{c}_1) + \phi(\mathbf{c}_2) - \phi(\tilde{\mathbf{c}}_1) - \phi(\tilde{\mathbf{c}}_2) = 0, \qquad (\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2) = T_{\alpha,\beta}(\mathbf{c}_1, \mathbf{c}_2) \tag{2.63}$$

It can be proven (see [25]) that $(1, \mathbf{c}, \mathbf{c}^2)$ and their linear combinations are the only (continuous) collision invariants[14].

Equilibrium Distribution

The equilibrium distribution $f^{(eq)}$ is defined as one (not necessarily unique) function that does not change anymore by further particle collisions, thus it satisfies

$$\int \begin{pmatrix} 1 \\ \mathbf{c} \\ \mathbf{c}^2 \end{pmatrix} \mathcal{C}[\tilde{f}^{(eq)}\tilde{f}_1^{(eq)} - f^{(eq)} f_1^{(eq)}](\mathbf{x}, t, \mathbf{c})dc = 0 \tag{2.64}$$

[14] This can be seen by considering collisions with input velocities $(\mathbf{c}, -\mathbf{c})$ as well as $(\mathbf{0}, \mathbf{c})$ and deriving from those that any continuous collision invariant $\phi(\mathbf{c})$ is either constant, linear or quadratic in $\mathbf{c}$.

for the collision invariants 1, $\mathbf{c}$, $\mathbf{c}^2$. This means that $\tilde{f}^{(eq)}\tilde{f}_1^{(eq)} - f^{(eq)}f_1^{(eq)} = 0$, thus $\ln f^{(eq)}$ satisfies (2.63) and is therefore a collisional invariant. Hence, it must be a linear combination of $(1, \mathbf{c}, \mathbf{c}^2)$,

$$\ln f^{(eq)} = a + \mathbf{B} \cdot \mathbf{c} + D\mathbf{c}^2 \tag{2.65}$$

By requiring that $f^{(eq)}$ produces the same first 5 moments as f,

$$\begin{aligned} \rho &:= \int f dc \overset{!}{=} \int f^{(eq)} dc, \\ \rho\mathbf{v} &:= \int \mathbf{c} f dc \overset{!}{=} \int \mathbf{c} f^{(eq)} dc, \\ \rho\theta &:= \int (\mathbf{c} - \mathbf{v})^2 f dc \overset{!}{=} \int (\mathbf{c} - \mathbf{v}) f^{(eq)} dc, \end{aligned} \tag{2.66}$$

we obtain the unique local Maxwellian equilibrium distribution in dimension d as

$$f^{(eq)}(\mathbf{x}, t, \mathbf{c}) = \frac{\rho(\mathbf{x}, t)}{\sqrt{2\pi\theta(\mathbf{x}, t)}^d} \exp\left(-\frac{(\mathbf{v}(\mathbf{x}, t) - \mathbf{c})^2}{2\theta(\mathbf{x}, t)}\right) \tag{2.67}$$

Entropy

Entropy is an important conceptional and mathematical quantity. It provides a link between a macroscopic observable and the microscopic states of a gas (see (2.70)). This linking idea motivated Boltzmann's choice (see [51], p. 38)

$$\eta = -k \int f \ln \frac{f}{y} dc, \quad \Phi_i := -k \int c_i f \ln \frac{f}{y} dc, \quad \Sigma := \int \mathcal{C}[-k \ln \frac{f}{y}](\mathbf{x}, t, \mathbf{c}) dc, \tag{2.68}$$

with some constants y and k, leading to the entropy transport equation

$$\partial_t \eta + \partial_i \Phi_i = \Sigma \tag{2.69}$$

It can be shown that $\Sigma \geq 0$, and thus (2.69) is in accordance with the second law of thermodynamics, stating that entropy can never decrease in a closed system (see e.g. [43], starting p. 70).

Boltzmann could show that, with a proper choice of the constant y, the total entropy $H := \int \eta dx$ is proportional to the number of possibilities W to distribute N particles into a given number of cells in the phase space, leading to the famous statement that Max Planck made carve onto Boltzmann's grave stone,

$$H = k \ln W. \tag{2.70}$$

Note that the concept of non-decreasing entropy is in contradiction to time reversibility in molecular dynamics (see Sect. 2.1.1). This is due to model assumptions in the derivation

of the Boltzmann equation, and can be mathematically located in the Boltzmann-Grad limit of the so called BBGKY hierarchy as $R \to 0$ (see [33] for more details).

In more physical terms, the time irreversibility can be explained through the decorrelation of particles by the stosszahlansatz. This decorrelation is plausible if the particles did not interact with each other for some (long) time. Just after a collision, however, they are strongly correlated, their respective speeds strongly depend on the other particle. If we reverse time just after the collision, the assumption of molecular chaos will not be valid, so the Boltzmann equation cannot be time reversible.

2.2.4 'Approximative' Collision Models

We have already argued that a full evaluation of the Boltzmann collision kernel (2.60) is extremely expensive in computation time. In the next subsections, we are therefore presenting some simplified collision models.

Ideally, simplified collision models should describe important physics as accurately as possible. Major issues are the collision invariants, the entropy theorem and the equilibrium Maxwellian distribution. There are also some physically motivated constants that are desirable to be reproduced, but these also depend on the regimes that we are interested in.

Most of the direct approximation qualities of simplified collision models are not strictly mathematically proven, but heuristically motivated. This complies with the challenging framework of modeling in the transition between molecular dynamics and continuum physics that we have encountered in the derivation of the Boltzmann collision operator.

BGK-Approximation

The BGK (Bhatnagar-Gross-Krook) model (see [5]) assumes a non-linear approximation

$$\mathcal{C}[\tilde{f}^{(eq)}\tilde{f}_1^{(eq)} - f^{(eq)}f_1^{(eq)}](\mathbf{x}, t, \mathbf{c}) \approx \frac{1}{\tau}\left(f_M[f](\mathbf{x}, t, \mathbf{c}) - f(\mathbf{x}, t\mathbf{c})\right), \tag{2.71}$$

where τ is the mean free flight time or Knudsen number. This model satisfies an entropy equation, its equilibrium distribution is very clearly a Maxwellian and it conserves mass, momentum and energy. Its derivation can be heuristically motivated, see [51] p. 46.

The simplification (2.71) ofers a significant computational advantage (in trade of physical accuracy) over the full collision kernel: no binary terms and collision angles need to be integrated. Due to this advantage, we will use the BGK-approximation in Part 4.

Discrete Velocity Models

Discrete velocity type models for the collision kernel are motivated through a pointwise numerical discretization of the velocity space. They represent the velocity space $\mathbb{R}^d$ ($d = 1, 2, 3$) by a (small) set of velocity points $\{\mathbf{c}_1, ..., \mathbf{c}_N\}$ and describe the collision operator by setting rules for the collisions of two particles at speeds $\mathbf{c}_i$ and $\mathbf{c}_j$, $i, j = 1, ..., N$. In meaningful models, these rules can be constructed such that physically reasonable collision invariants (mass, momentum and energy) are respected, and some notion of (non-Maxwellian) equilibrium is possible. Typically, discrete velocity models lead to collision kernels that are bilinear in f.

The first such models have been developed by Broadwell (see [10]), with e.g. $N = 4$. Some more recent approaches by Babovsky (see [3]) use more velocity points and imitate the geometric features of the collision process (compare Sect. 2.2.1).

In Sect. 3.9.2, we will consider a two dimensional 16-discrete velocity model. The collision rules are illustrated in Fig. 2.4. If e.g. two particles at speeds $\mathbf{c}_7$ and $\mathbf{c}_2$ collide, their post-collision velocities will be $\mathbf{c}_6$ and $\mathbf{c}_3$. Or, if we have a collision between particles at speeds $\mathbf{c}_7$ and $\mathbf{c}_{15}$, we obtain post-collision velocities of $\mathbf{c}_{10}$ and $\mathbf{c}_{12}$. In (A.22) and (A.23), we find the full bilinear collision operator for this model.

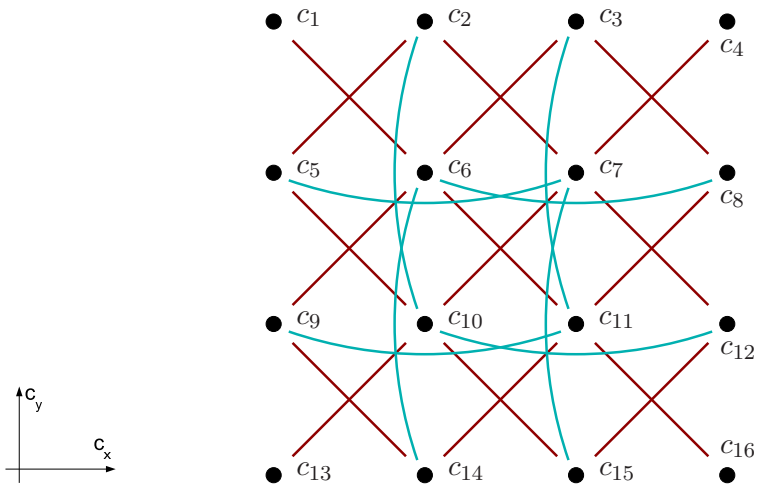


Figure 2.4: 2-D velocity space with interactions in the 16 discrete velocities model.

It is clear that with discrete velocity models at small N, we cannot perform simulations with precise quantitative agreement to physical reality. Such models are very interesting for the study of qualitative effects or can serve as useful mathematical models like in Sect. 3.9.2.

Direct Simulation Monte Carlo

The 'Direct Simulation Monte Carlo' (DSMC) method is a statistical approach to approximate the physics of the Boltzmann equation (see [6]). It uses statistical ensembles of particles that are evolved in time. Thus, this method does not only simplify the expensive, high dimensional evaluation of the collision integral (2.60) by a stochastic approximation. DSMC indeed consists of a complete, particle oriented approach to the whole time evolution of the Boltzmann equation.

The DSMC method has proven very successfull for innumerous cases of extremely rarefied gases (see Sect. 1.1).

The nature of the DSMC method is clearly compuational, it thus allows only little insight into the physical processes involved. In this sense it is a very challenging playground for implementations (parallelization, optimization) and also stochastic numerical mathematics.

2.2.5 Relation to Macroscopic Balance Laws

After exploring the connections of the Boltzmann equation to molecular dynamics, we would like to conclude this overview with linking the Boltzmann equation to the other side of the scale, to macroscopic field equations as shown in Sect. 2.1.2.

The Boltzmann equation reproduces (2.28) out of averaging monomials over the whole velocity space,

$$\partial_t \int \begin{pmatrix} 1 \\ c_i \\ \frac{1}{2}c_i^2 \end{pmatrix} f(\mathbf{x},t,\mathbf{c})dc + \partial_k \int \begin{pmatrix} 1 \\ c_i \\ \frac{1}{2}c_i^2 \end{pmatrix} c_k f(\mathbf{x},t,\mathbf{c})dc = 0. \tag{2.72}$$

With the physically intuitive definitions of mass density, momentum density, stress tensor and temperature (in energy units) as

$$\rho(x_j,t) = \int_{\mathbb{R}^3} f(x_j,t,c_j)dc \tag{2.73a}$$

$$\rho v_i(x_j,t) = \int_{\mathbb{R}^3} c_i f(x_j,t,c_j)dc \tag{2.73b}$$

$$\sigma_{ij}(x_j,t) = \int_{\mathbb{R}^3} (c_i - v_i)(c_j - v_j) f(x_j,t,c_j)dc \tag{2.73c}$$

$$\rho\theta(x_j,t) = \int_{\mathbb{R}^3} (c_i - v_i)^2 f(x_j,t,c_j)dc = \operatorname{trace}\sigma. \tag{2.73d}$$

Calculations reveal that (2.72) is indeed equivalent to (2.28). Note that the closure problem emerges out of (2.72) through the multiplication of $\partial_k f$ with c_k. No matter how many monomials we will be using, there will always be more unknowns than equations.

The closure problem is the source of challenges in modeling macroscopic equations that imitate the physics of the Boltzmann equation as accurately as possible. In the present work, we will explore several approaches to deal with these modeling issues, some of them will give qualitative insights (Part 3), others will be more computationally oriented (Part 4).

3 Analysis of Approximations to the Linear Boltzmann Equation

Most of this part has been published in [30], in collaboration with Manuel Torrilhon and Michael Junk. The main elaboration of the publication has been done by the author of this thesis, otherwise, contributions cannot be clearly allocated to the single authors. The published work is a result of various discussions and compromises among all the authors.

Note: The Knudsen number is called 'ε' in this part. The variable name 'τ', as used before, does not fit into the typical framework of mathematics and asymptotic expansions.

3.1 Introduction

Kinetic theory describes the flow of gases by means of a stochastic description based on the distribution function of the particle velocities, as we have seen in Part 1. The distribution function obeys the Boltzmann equation - an integro-differential equation that considers free streaming and collisions of the particles. This description of gases is a detailed, complex, microscopic approach reflected in the fact that the state of the gas at a spatial point is given by a function, i.e., an infinite dimensional object. In contrast, gases in classical fluid dynamics are described by a low dimensional vector of variables, typically density, velocity and temperature in each space point.

The aim of approximation methods in kinetic theory is to reduce the high dimensional particle description rigorously to a low-dimensional continuum model. Classical approaches are given by asymptotic analysis and function approximation theory. The Chapman-Enskog expansion conducts an asymptotic analysis where the smallness parameter is the Knudsen number, see for example the textbook [14]. This expansion successfully derives the fluid dynamic laws of Navier-Stokes and Fourier, but fails to produce useful higher order results beyond the first order. Instead, the Burnett- and super-Burnett-equations have been shown to be unstable in [8]. Grad's moment approach uses approximation theory and represents the distribution function as series of Hermite functions, see [22, 23]. In the limit, this series is supposed to reproduce any

distribution function. Truncations of the series give rise to moment equations that approximate Boltzmann's equation. However, the approximation converges slowly and also unphysical artifacts, like subshocks, are produced, see e.g. [59].

Various attempts exist to remedy the drawbacks of the Chapman-Enskog expansion. The work [29] introduced a hyperbolic form of the Burnett equations which is stable, while in [7] it was shown that a variable transformation may be able to remove unstable terms from the second order Chapman-Enskog result. Moment equations have been popular for their mathematical structure, see [36], and also for some success in describing physical processes, see the textbook [40]. A combination of Grad's moment method and an asymptotic approach has been introduced in [52].

Recently, in [49], a new derivation of macroscopic equations was presented that was claimed to be different from both Chapman-Enskog and Grad. This so-called *order-of-magnitude method* is based on general moment equations and follows the scale of the variables for a closure, see also the textbook [51]. The resulting equations exhibit an inherent asymptotic accuracy in the sense of Chapman-Enskog and they are stable. The method succeeded to derive generalized 13-moment-equations in [50] and also showed that the R13-equations of [52] are a correct, stable, third order accurate approximation of Boltzmann's equation. This may explain the success of the R13-equations as demonstrated in [54, 57, 56, 59]. The R13-equations even allow to construct reasonable boundary conditions, see [26, 53, 60].

In this part, we extend the order-of-magnitude method to the level of kinetic equations. So far, this method was only applied to the full non-linear moment hierarchy with little chance to gain insight into the general mathematical idea and structure of the closure. Our aim is to develop a formal theory of the new closure and to apply it to general kinetic equations. Here, we restrict ourselves to a linear kinetic model equation and demonstrate the relation of the new method to the classical approaches of Chapman-Enskog and Grad. We prove the asymptotic accuracy of the resulting closure and show the existence of an entropy law and L^2-stability, once specific variables are chosen. Our findings clearly show how the method exploits the scaling of the distribution function and the structures that this scaling creates in the phase space. Hence, it is reasonable to call this method a *scale-induced closure*.

This part is organized as follows: The next section briefly resumes the order-of-magnitude method as applied to the moment hierarchy in [51] and discusses the results. Sect. 3.3 introduces the linear kinetic model and Sect. 3.4 discusses the classical closure theories in their application to the model. The new scale-induced closure is derived in Sect. 3.5, the asymptotic accuracy and stability are proven Sect. 3.6 and Sect. 3.7. A generalization to higher orders is sketched in Sect.3.8. As examples of the new method, Sect. 3.9 discusses the generalized 13-moment-system of [50], the application of the new closure to a 16 discrete velocities scheme and an application to a more general, high dimensional "kinetic type" equation. In that setting, the classical closures are compared to the scale induced closure, and the advantage of the latter is clearly shown. We finish this part with

a conclusion, Sect. 3.10. Some technical details are shown in an appendix to this part, see Sect. A.3 and A.4. Some mathematical background about hyperbolicity, stability and entropy is given in the appendix, Sect.A.2.

3.2 Struchtrup's Order-of-Magnitude Approach

In the papers [49] and [50], Struchtrup proposes an order-of-magnitude approach to derive macroscopic transport equations in kinetic gas theory based on Boltzmann's equation. We briefly summarize the results of his method which will be generalized in the later sections. For details we refer to the original papers and the textbook [51].

The Boltzmann equation

$$\frac{\partial f}{\partial t} + c_i \frac{\partial f}{\partial x_i} = \frac{1}{\varepsilon} J(f, f) \tag{3.1}$$

describes the evolution of the distribution function f of the particle velocities in a monatomic gas. The value of $f(\mathbf{x}, t, \mathbf{c})\, d\mathbf{c}$ gives the number density of particles in $\mathbf{x}$ at time t with velocities in $[\mathbf{c}, \mathbf{c} + d\mathbf{c}]$ with $\mathbf{c}$ defined with respect to an absolute reference. The collision operator J is an integral functional which depends quadratically on f (see, for example, [12]). We assume, that the equation is normalized in such a way that the Knudsen number, i.e., the ratio between the mean free path and a macroscopic length, appears as scaling parameter ε.

Relevant for macroscopic equations are the equilibrium moments density, momentum density and energy density

$$\varrho = m \int_{\mathbb{R}^3} f \, d\mathbf{c}, \qquad \varrho \boldsymbol{v} = m \int_{\mathbb{R}^3} \mathbf{c}\, f \, d\mathbf{c}, \qquad \frac{3}{2}\rho\theta + \frac{1}{2}\varrho v^2 = \frac{1}{2} m \int_{\mathbb{R}^3} c^2 \, f \, d\mathbf{c} \tag{3.2}$$

from which average velocity $\boldsymbol{v}$ and temperature θ (in energy units) are derived. Additional higher order non-equilibrium moments are defined as

$$u^s_{i_1 \cdots i_n} = m \int_{\mathbb{R}^3} C^{2s} C_{\langle i_1} \cdots C_{i_n \rangle} \, (f - f_M) \, d\mathbf{C}. \tag{3.3}$$

Here, f_M is the Maxwell distribution and $\mathbf{C} = \mathbf{c} - \boldsymbol{v}$ is the peculiar velocity. Indices in angular brackets denote the symmetric and trace-free part of the corresponding tensor. Evolution equations for the moments follow from integration of (3.1). They form an infinite hierarchy with a closure problem.

The order-of-magnitude approach closes the system of equations in three steps. As first step, a Chapman-Enskog expansion is conducted (e.g., $u^s_i = \varepsilon\, u^s_{i|1} + \varepsilon^2\, u^s_{i|2} + ...$) on the infinite hierarchy in order to assign an order of magnitude in terms of the Knudsen number to all moments. In the first expansion, only vectorial and second degree tensors with arbitrary number of traces are non-zero and we obtain

$$u^s_{i|1} = -\kappa_s\, \rho\theta^s \frac{\partial \theta}{\partial x_i}, \quad u^s_{ij|1} = -\mu_s\, \rho\theta^{s+1} \frac{\partial v_{\langle i}}{\partial x_{j \rangle}}, \quad u^s_{i_1 \cdots i_n} = 0 \ (n > 2). \tag{3.4}$$

The subscript 1 denotes the first expansion, κ_s and μ_s are pure numbers. All other moments vanish to first order in ε. Obviously, heat flux $q_{i|1} = \frac{1}{2}u^1_{i|1}$ and stress tensor $\sigma_{ij|1} = u^0_{ij|1}$ are among the first order moments. The coefficients κ_s and μ_s stem from the production terms of the moment equations. For a more specific representation in terms of quantities involving the collision operator, we refer to [52].

The fact that all vectorial and 2-tensorial moments are of the same order of magnitude is used as constitutive relation in the second step of the method. Indeed, up to an error of second order in the Knudsen number, two of all these moments suffice to calculate the value of the others. As natural candidates for a basis we choose heat flux $u^1_{i|1}$ and stress tensor $u^0_{ij|1}$ and eliminate the gradient expressions in (3.4). The result are local constitutive equations for all higher moments accurate up to an error of second order in ε. They read

$$u^s_{i|1} = \frac{\kappa_s}{\kappa_1}\,\theta^{s-1}u^1_{i|1} \qquad (s>1), \tag{3.5}$$

$$u^s_{ij|1} = \frac{\mu_s}{\mu_0}\theta^s u^0_{ij|1} \qquad (s>0), \tag{3.6}$$

$$u^s_{i_1\cdots i_n\,|1} = 0 \qquad (s>0,\, n>2). \tag{3.7}$$

In the last step of the method, these relations are inserted into the moment hierarchy and all expressions that have been shown to be of higher order in ε than two, are simply set to zero. The final equations form a closed system based on quantities and expressions with consistent order of magnitude. It is important to note, that the closure (3.5)/(3.6) depends on the collision integral through the parameters κ_s and μ_s.

The order-of-magnitude method is, in principle, capable to produce equations at any order of Knudsen number, see [51]. However, only equations up to third order have been derived, so far.

3.3 Linear Kinetic Model

In the following we will recast the order-of-magnitude approach into a general kinetic framework and demonstrate attractive properties of the resulting equations.

The theory is developed for a generic linear kinetic model which includes discrete velocity models with finite velocity sets $\mathcal{C} \subset \mathbb{R}^d$ as well as the continuous case $\mathcal{C} = \mathbb{R}^d$.

Definition 3.3.1 (Kinetic Model). *Starting from an open spatial domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$ and a velocity set $\mathcal{C} \subset \mathbb{R}^d$ we identify distribution functions $f : \mathbb{R}^+ \times \Omega \times \mathcal{C} \to \mathbb{R}$ with elements $f_{t,\mathbf{x}} : \mathcal{C} \to \mathbb{R}^+$ of a suitable Hilbert space V of real valued functions on $\mathcal{C}$. A solution of the linear kinetic model is a distribution function which satisfies*

$$\partial_t f(t,\mathbf{x},\mathbf{c}) + \mathbf{c}\cdot\nabla f(t,\mathbf{x},\mathbf{c}) + \frac{1}{\varepsilon}Kf(t,\mathbf{x},\mathbf{c}) = 0, \qquad (t,\mathbf{x},\mathbf{c}) \in \mathbb{R}^+\times\Omega\times\mathcal{C} \tag{3.8}$$

with Knudsen number ε and a linear collision operator $K : V \to V$, independent of $(t, \mathbf{x})$, with the following properties:

1. *K has a p-dimensional kernel ($p \in \mathbb{N}$) injectively parametrized by an equilibrium distribution*
$$M : \mathbb{R}^p \to V, \quad \rho \longmapsto M\,\rho \tag{3.9}$$
satisfying $K\,M = 0$. The operator M does not depend on $(t, \mathbf{x})$. The function $(M\rho)(t, \mathbf{x}, \mathbf{c})$ plays the role of the Maxwellian distribution function. The components of ρ are called equilibrium parameters.
2. *There exists a surjective equilibrium operator generalizing the mapping to the equilibrium moments, which is independent of $(t, \mathbf{x})$*
$$E_0 : V \to \mathbb{R}^p, \quad f \longmapsto \rho = E_0 f \tag{3.10}$$
and satisfies the conservation property $E_0\,K = 0$ as well as $E_0\,M = id_{\mathbb{R}^p}$. Note that the combination $Q = ME_0$ is a projection onto the kernel of K, the so called equilibrium projection. Accordingly, $P = id - Q$ is called non-equilibrium projection. With this projections we have the decomposition $V = V_0 \oplus V_{NE}$ with the equilibrium space $V_0 = QV$ and the non-equilibrium space $V_{NE} = PV$.
3. *There exists a linear mapping $K^\dagger : V \to V$ with the properties*
$$K^\dagger Q = 0, \qquad K^\dagger K = K K^\dagger = P \tag{3.11}$$

The condition $E_0 M = id_{\mathbb{R}^p}$ clearly implies that E_0 inverts the action of M, or in other words, that E_0 is a pseudo-inverse of M.

Definition 3.3.2. *Let X, Y be vector spaces and $A : X \to Y$ be linear. A linear mapping $B : Y \to X$ is called a pseudo inverse of A (abbreviated as $B = A^\dagger$), provided*
$$ABA = A, \qquad BAB = B. \tag{3.12}$$
If X, Y are Hilbert spaces, B is called Moore-Penrose-inverse of A if in addition to (3.12) the operators AB and BA are self-adjoint, i.e.
$$(AB)^* = AB, \qquad (BA)^* = BA. \tag{3.13}$$

One can show that the Moore-Penrose-inverse is unique (see, for example, [19] and App. A.3) and, in the case of injective A and finite dimensional X, it is given by $B = (A^*A)^{-1}A^*$.

Applied to our situation with $M = A$ and $E_0 = B$, we first see that $E_0 M = id_{\mathbb{R}^p}$ implies (3.12) so that E_0 is indeed a pseudo-inverse of M. Moreover, the identity $E_0 M$ is self-adjoint with respect to any scalar product on $\mathbb{R}^p$ and the self-adjointness of the converse product $Q = ME_0$ is equivalent to the orthogonality of the projection Q. In particular,

the Moore-Penrose-inverse $M^\dagger = (M^*M)^{-1}M^*$ can serve as equilibrium operator E_0 provided $M^\dagger K = 0$. Since

$$M^\dagger K = (M^*M)^{-1}M^*K = (M^*M)^{-1}(K^*M)^*,$$

we see that this condition is satisfied when K is self-adjoint, i.e. $K^* = K$, because $KM = 0$. This case will be of importance in section 3.7.

Using the properties of K and $K^\dagger$ one can show (3.12)

$$\begin{aligned} KK^\dagger K &= KP = K - KQ = K \\ K^\dagger K K^\dagger &= K^\dagger P = K^\dagger - K^\dagger Q = K^\dagger \end{aligned}$$

so that $K^\dagger$ is really a pseudo-inverse of K. If K is self adjoint and E_0 is chosen as Moore-Penrose-inverse of M, then Q and P are also self-adjoint. In this case, $K^\dagger$ is the Moore-Penrose-inverse of K because (3.13) is also satisfied.

Further properties of K and $K^\dagger$ which will be frequently used later can also directly be deduced from the basic assumptions:

$$\begin{aligned} KQ = QK = 0, \qquad KP = PK = K, \\ K^\dagger P = PK^\dagger = K^\dagger, \qquad K^\dagger Q = QK^\dagger = 0. \end{aligned} \tag{3.14}$$

In the case of the Boltzmann equation, the space V would be some weighted $\mathbf{L}^2$ space. The equilibrium parameters are $\rho = E_0 f = \int \psi f$ with $\psi = (1, \mathbf{c}, c^2)^T$ and $p = 2 + d$. Furthermore, $M\,\rho$ would be given by the Maxwell distribution $f_M(\varrho, \boldsymbol{v}, \theta; \mathbf{c})$.

The vector of equilibrium parameters ρ is a mapping

$$\rho : \Omega \times \mathbb{R}^+ \to \mathbb{R}^p. \tag{3.15}$$

The modelling task in kinetic theory is to find reasonable evolution equations for ρ by using a projected space with much lower dimension than V. The following theory will achieve this goal.

3.4 Classical Approximations

Classical asymptotic limits and approximations of kinetic equations include the Euler equations, Chapman-Enskog expansion and Grad's method. We review these results here for our model since the new approach is built upon them and shows various connections to them.

In equation (3.8), the limit $\varepsilon \to 0$ formally leads to $Kf = 0$ so that the distribution function is asymptotically given by an equilibrium $M\,\rho = Qf$. Any extension beyond equilibrium will be written

$$f = Qf + Pf = M\,\rho + f^{(\mathrm{NE})} \tag{3.16}$$

with a non-equilibrium disturbance $f^{(\mathrm{NE})}$.

3.4.1 Equilibrium Closure

The Euler equations arise if we apply the equilibrium operator E_0 to (3.8) and obtain

$$\partial_t \rho + E_0 \mathbf{c} \cdot \nabla M\rho + E_0 \mathbf{c} \cdot \nabla f^{(\mathrm{NE})} = 0 \tag{3.17}$$

The closure assumption $f^{(\mathrm{NE})} = 0$ produces the Euler equations.

3.4.2 Chapman-Enskog Closure

The Chapman-Enskog expansion asks for the structure of the disturbance $f^{(\mathrm{NE})} = Pf$. It is easy to find an evolution equation for this quantity by applying the non-equilibrium projection P to (3.8) and observing (3.14)

$$\partial_t f^{(\mathrm{NE})} + P\mathbf{c} \cdot \nabla f^{(\mathrm{NE})} + P\mathbf{c} \cdot \nabla M\rho + \frac{1}{\varepsilon} K\, f^{(\mathrm{NE})} = 0. \tag{3.18}$$

Inserting the expansion $f^{(\mathrm{NE})} = \varepsilon\, f_1^{(\mathrm{NE})} + \varepsilon^2 f_2^{(\mathrm{NE})} + \ldots$, applying $K^\dagger$ and using (3.11) and (3.14), we obtain under the condition that all coefficients $f_k^{(\mathrm{NE})}$ and their derivatives are bounded with respect to ε

$$P f_1^{(\mathrm{NE})} + K^\dagger \mathbf{c} \cdot \nabla\, M\rho = O(\varepsilon). \tag{3.19}$$

In the Chapman-Enskog approach, this necessary condition on $f_1^{(\mathrm{NE})}$ is replaced by the sufficient but more strict requirement

$$f_1^{(\mathrm{NE})} = -K^\dagger \mathbf{c} \cdot \nabla\, M\rho. \tag{3.20}$$

Using $\varepsilon f_1^{(\mathrm{NE})}$ as approximation for $f^{(\mathrm{NE})}$ in (3.17), we can close the equation in a more accurate way, leading to the general Navier-Stokes-Fourier equations

$$\partial_t \rho + E_0 \mathbf{c} \cdot \nabla M\rho = \varepsilon\, E_0 (\mathbf{c} \cdot \nabla)\, K^\dagger (\mathbf{c} \cdot \nabla) M\rho. \tag{3.21}$$

Going one order further and using relation (3.20) for $f_1^{(\mathrm{NE})}$, we get from (3.18)

$$K f_2^{(\mathrm{NE})} - K^\dagger \mathbf{c} \cdot \nabla M \partial_t \rho - P\mathbf{c} \cdot \nabla K^\dagger \mathbf{c} \cdot \nabla M\rho = O(\varepsilon) \tag{3.22}$$

which can be solved in the form

$$P f_2^{(\mathrm{NE})} = K^\dagger K^\dagger \mathbf{c} \nabla M \partial_t \rho + K^\dagger \mathbf{c} \cdot \nabla K^\dagger \mathbf{c} \cdot \nabla M\rho + O(\varepsilon) \tag{3.23}$$

Again, dropping the non-equilibrium projection and the possible $O(\varepsilon)$ contribution, the necessary condition is replaced by a more strict requirement in the classical Chapman-Enskog approach. In these so called Burnett relations for $f_2^{(\mathrm{NE})}$, the time derivatives $\partial_t \rho$

can be replaced by $-E_0\mathbf{c}\cdot\nabla M\rho$ (Euler equations) with no loss of order. The equations then read

$$\begin{aligned}\partial_t\rho + E_0\mathbf{c}\cdot\nabla M\rho =& \varepsilon\, E_0(\mathbf{c}\cdot\nabla)\, K^\dagger(\mathbf{c}\cdot\nabla)M\rho \\ &- \varepsilon^2 E_0(\mathbf{c}\cdot\nabla)K^\dagger(\mathbf{c}\cdot\nabla)K^\dagger(\mathbf{c}\cdot\nabla)M\rho \\ &+ \varepsilon^2 E_0(\mathbf{c}\cdot\nabla)K^\dagger K^\dagger(\mathbf{c}\nabla M)E_0(\mathbf{c}\cdot\nabla)M\rho.\end{aligned} \tag{3.24}$$

However, (3.24) can be proven to be unstable in the realistic cases of the full Boltzmann collision operator, see [8]. Higher order expansions like super-Burnett equations, turn out to be unstable as well, [51]. This failure of the expansion indicates that the assumptions on the coefficients are too strict in the higher order cases. In fact, for a model problem (see [11]) one can show that less rigid assumptions help to avoid the stability breakdown.

There exist various attempts to stabilize the Burnett equations, for example [7] and [29] which can be seen as particular choices of the right hand side in (3.23).

3.4.3 Grad Closure

Grad in [22] and [23] assumes a specific form of the distribution function which we summarize as

$$f = M\,\rho + G\,\mu + \tilde{f}. \tag{3.25}$$

Here, the non-equilibrium part is composed of the Grad distribution $G\mu$ and a remainder $\tilde{f}$, where $G : \mathbb{R}^q \to V$ maps certain non-equilibrium parameters $\mu \in \mathbb{R}^q$ onto a distribution function. The dependencies of G on the equilibrium variables ρ are neglected in accordance with a linear theory. The range of the mapping G can be viewed as vectors of the distribution space V opening a subspace additional to the equilibrium space given by M. In the original Grad theory, this subspace is spanned by Hermite polynomials. The parameters μ are typically defined in terms of higher order moments, for example, as non-equilibrium parts of the fluxes of the equilibrium variables. More generally, we assume that $\mu = E_1 f$ with a linear mapping $E_1 : V \to \mathbb{R}^q$ which satisfies

$$E_1 G = id_{\mathbb{R}^q}, \quad E_1 M = 0, \quad E_0 G = 0. \tag{3.26}$$

As a consequence, $S = GE_1$ is a projection which decomposes P into two parts S and $R = P - S$, the latter one being the projection onto the remainder term.

Application of E_0 and E_1 to the equation (3.8) yields evolution equations for ρ and μ. We find

$$\partial_t\rho + E_0\mathbf{c}\cdot\nabla M\rho + E_0\mathbf{c}\cdot\nabla G\mu + E_0\mathbf{c}\cdot\nabla\tilde{f} = 0 \tag{3.27}$$

and

$$\partial_t\mu + E_1\mathbf{c}\cdot\nabla M\rho + E_1\mathbf{c}\cdot\nabla G\mu + E_1\mathbf{c}\cdot\nabla\tilde{f} + \frac{1}{\varepsilon}E_1 K\, G\,\mu + \frac{1}{\varepsilon}E_1 K\,\tilde{f} = 0, \tag{3.28}$$

where in Grad's approach $\tilde{f} = 0$ leads to a closure of the system. Grad's equations can typically be shown to be stable.

From a geometric point of view, Grad's approach amounts to a splitting of the non-equilibrium space V_{NE} into a resolved and an unresolved subspace where the resolved subspace $V_1 = \mathrm{Im}(G)$ is parametrized through an - up to conditions (3.26) - arbitrary choice of higher order moments E_1 (see Figure 3.1). Hence, the asymptotic accuracy in terms of Knudsen number remains unclear for Grad's equations.

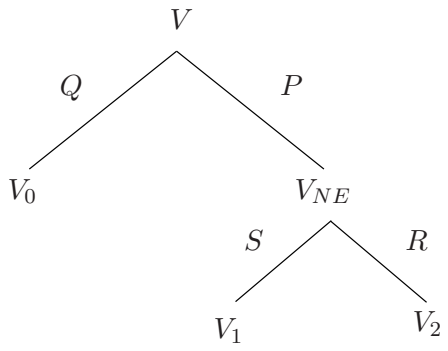


Figure 3.1: Splitting of the phase space into an equilibrium subspace V_0 with projection $Q = ME_0$ and the non-equilibrium remainder $V_{NE} = PV$ which is again split into the primary non-equilibrium subspace $V_1 = SV$ with Grad-projection $S = GE_1$ and the secondary non-equilibrium subspace $V_2 = RV$.

3.5 Scale-Induced Closure

The order-of-magnitude approach wants to derive stable moment equations which are asymptotically accurate in the sense of a Chapman-Enskog expansion. Burnett equations satisfy the accuracy condition, but are unstable. On the other hand, Grad's equations are stable but the closure is based on a distribution function which is arbitrarily reconstructed through higher moments and has no a-priori asymptotic properties.

3.5.1 Derivation

The Chapman-Enskog expansion implies a distribution function in the form

$$f = M\,\rho + \varepsilon\, f_1^{(\mathrm{NE})} + \varepsilon^2 f_2^{(\mathrm{NE})} + \mathcal{O}(\varepsilon^3), \tag{3.29}$$

while in Grad's approach, the distribution function is structured according to

$$f = M\,\rho + G\,\mu + f_R \tag{3.30}$$

with equilibrium part $M\rho \in V_0$, the primary non-equilibrium contribution $G\mu \in V_1$ and the secondary contribution $f_R \in V_2$ (see Figure 3.1).

A compatibility between the two representations (3.29) and (3.30) may be achieved if V_1 is constructed in such a way that it contains $\varepsilon f_1^{(\mathrm{NE})}$. Thus, the task is to appropriately define G and moments μ with their operator E_1 such that, apart from the basic requirements

$$E_1 G = id_{\mathbb{R}^q}, \qquad E_1 M = 0 \qquad E_0 G = 0, \tag{3.31}$$

also $\varepsilon f_1^{(\mathrm{NE})} = G\mu \in \mathrm{Im}(G) = V_1$ is possible. In contrast to Grad's moment approach where the distribution function is specified, for example, as Hermite series independent of the kinetic equation, the condition $\varepsilon f_1^{(\mathrm{NE})} \in V_1$ combines the phase space splitting with the structure of the kinetic equation.

Using the equilibrium projection Q, and the projections $S = GE_1$ and $R = P - S$ related to the primary and secondary non-equilibrium, we can derive equations for ρ, μ and f_R. Applying R to (3.8) and using (3.30), we obtain

$$\varepsilon^2 \partial_t \hat{f}_R + R\mathbf{c} \cdot \nabla M\rho + \varepsilon R\mathbf{c} \cdot \nabla G\,\hat{\mu} + \varepsilon^2 R\mathbf{c} \cdot \nabla \hat{f}_R + RK\,G\,\hat{\mu} + \varepsilon RK\,\hat{f}_R = 0, \tag{3.32}$$

where we scaled the moments $\mu = \varepsilon\hat{\mu}$ and $f_R = \varepsilon^2 \hat{f}_R$. If we choose the primary non-equilibrium G such that for some suitable $\hat{\mu}$

$$G\,\mu = \varepsilon G\,\hat{\mu} = -\varepsilon K^\dagger \mathbf{c} \cdot \nabla M\rho \tag{3.33}$$

we automatically satisfy the following equivalent requirements

1. the first expansion coefficient $\varepsilon f_1^{(\mathrm{NE})}$ in (3.20) can be written in the form $G\,\mu$.
2. the evolution of the remainder f_R in (3.32) is governed only by quantities at least first order in ε.
3. the distribution $G\,\mu$ is given by the leading order term of the expansion of the distribution function f in powers of ε conducted on (3.18).

To see Item 2, we do a short calculation: combining the zeroth order terms in (3.32) we have

$$R\mathbf{c} \cdot \nabla M\rho + RKG\hat{\mu} \overset{(3.33)}{=} R\left(\mathbf{c} \cdot \nabla M\rho - P\mathbf{c} \cdot \nabla M\rho\right) \overset{RP=R}{=} 0. \tag{3.34}$$

In order to derive an expression for G from (3.33) we will write it in the form

$$G\,\hat{\mu} = -K^\dagger \mathbf{c} \cdot M\,\nabla\rho, \tag{3.35}$$

where now the operator $-K^\dagger \mathbf{c} \cdot M$ acts on $p \times d$ gradients $\nabla\rho =: A \in \mathbb{R}^{p\times d}$ according to

$$-K^\dagger \mathbf{c} \cdot MA = -K^\dagger \sum_{i=1}^{p} \sum_{\alpha=1}^{d} c_\alpha M\,\mathbf{e}_i A_{i\alpha} \tag{3.36}$$

where $\mathbf{e}_i$ are the $\mathbb{R}^p$ unit vectors. The same convention is applied to operators with the same structure.

In the next two sections we consider two alternative paths to the specification of the operators G and E_1. In order to keep notation simple, we use μ for the scaled higher moments $\hat{\mu}$ in the following.

3.5.2 Constructing the Distribution

In this section we choose a moment operator $E_1 : V \to \mathbb{R}^q$ with some restrictions and determine the distribution function G from it. This point of view corresponds to constructing a closure $G\mu$ for the infinite moment hierachy by saying that $f = M\rho + G\mu$ for given moments $\mu = E_1 f$. This directly corresponds to Struchtrup's order-of-magnitude approach. Note, that the moment production terms can be computed without further assumptions on f, as long as the operator $K^\dagger \mathbf{c} \cdot M$ is not pathologic (see below). This simplifies the process originally developed by Struchtrup in [50].

The projector E_1 cannot be chosen entirely arbitrarily. Since we require $E_1 G = id_{\mathbb{R}^q}$ and want to replace gradients by moments in (3.35) we have to choose E_1 such that the linear equation

$$\mu = E_1 G\mu = -\varepsilon E_1 K^\dagger \mathbf{c} \cdot M \nabla \rho \tag{3.37}$$

is essentially solvable for $\nabla\rho$. We expect that this leaves quite some freedom for the choice of E_1.

Let us make this restriction a bit more precise: In general, the operator $K^\dagger \mathbf{c} \cdot M$ has a non-trivial nullspace $\ker\left(K^\dagger \mathbf{c} \cdot M\right) \in \mathbb{R}^{p \times d}$. We define $V_1 := \mathrm{Im}(K^\dagger \mathbf{c} \cdot M)$ and choose E_1 injective on V_1, i.e. $\ker(E_1) \cap V_1 = \{0\}$, meaning that E_1 should not enlarge the kernel of $K^\dagger \mathbf{c} \cdot M$. Furthermore we require the basic relation that $E_1 M = 0$ and define q such that $E_1 : V_1 \to \mathbb{R}^q$ is surjective.

Defining the projections T_0 onto $\ker\left(K^\dagger \mathbf{c} \cdot M\right)$ and T_1 onto any subspace complementary to $\ker(K^\dagger \mathbf{c} \cdot M)$ in $\mathbb{R}^{p \times d}$, we can write $\nabla\rho = T_0 \nabla\rho + T_1 \nabla\rho$. We then solve

$$T_1 \nabla\rho = -\frac{1}{\varepsilon} \left(E_1 K^\dagger \mathbf{c} \cdot M\right)^\dagger \mu, \tag{3.38}$$

which now determines the relevant part of $\nabla\rho$ in terms of μ. The symbol $\dagger$ denotes any pseudoinverse, see Sect. 3.3. This procedure should be compared to the elimination of gradient expressions in Section 3.2 for the order-of-magnitude method conducted on moments.

For G we then compute

$$\begin{aligned} G\mu &= -\varepsilon K^\dagger \mathbf{c} \cdot M \left(T_0 \nabla\rho + T_1 \nabla\rho\right) && (3.39)\\ &= -\varepsilon K^\dagger \mathbf{c} \cdot M \left(T_0 \nabla\rho - \frac{1}{\varepsilon} \left(E_1 K^\dagger \mathbf{c} \cdot M\right)^\dagger \mu\right) && (3.40) \end{aligned}$$

$$= K^\dagger \mathbf{c} \cdot M \left(E_1 K^\dagger \mathbf{c} \cdot M \right)^\dagger \mu \tag{3.41}$$

and thus

$$G = K^\dagger \mathbf{c} \cdot M \left(E_1 K^\dagger \mathbf{c} \cdot M \right)^\dagger . \tag{3.42}$$

Lemma 3.5.1 (Scale-Induced Distribution)**.**
Under the assumptions $\ker(E_1) \cap V_1 = \{0\}$ *with* $V_1 = \operatorname{Im}\left(K^\dagger \mathbf{c} \cdot M\right)$ *and* $E_1 M = 0$ *for the moment projector* $E_1 : V \to \mathbb{R}^q$ *with* $q = \dim V_1$*, we have for the distribution function (3.42):*

1. *The construction is in accordance with the requirements* (3.31)*. The non-equilibrium space can be split into* $V_{NE} = V_1 \oplus V_2$*, with* $V_1 := G\mathbb{R}^q = GE_1V_{NE} = SV_{NE}$ *and* $V_2 = (P - S)V_{NE} = RV_E$*.*
2. *The construction satisfies* $\operatorname{Im}(G) = \operatorname{Im}\left(K^\dagger \mathbf{c} \cdot M\right) = V_1$ *in agreement with the condition (3.35) as well as* $\ker(G) = \{0\}$*.*
3. V_1 *constains contributions to the distribution function up to order* $\mathcal{O}(\varepsilon)$ *and* V_2 *contains all orders higher than* ε^2 *in a Chapman-Enskog expansion.*

Proof.

1. We clearly have $E_1 G = E_1 K^\dagger \mathbf{c} \cdot M \left(E_1 K^\dagger \mathbf{c} \cdot M\right)^\dagger = id_{\mathbb{R}^q}$ due to the requirement $\ker(E_1) \cap V_1 = \{0\}$. The condition $E_1 M = 0$ was required for E_1 a priori. Furthermore $PG = G$ follows from $PK^\dagger = K^\dagger$ and implies that $E_0 G = E_0 PG = 0$. The splitting follows from these three requirements. For the decomposition, we observe that $G\mathbb{R}^q = V_1 \subset V_{NE}$ since $QG = ME_0G = 0$. With $E_1 M = 0$ we have that $E_1 V_0 = E_1 M E_0 V = \{0\}$ and with $E_1 R = E_1 P - E_1 S = E_1 - E_1 = 0$, we have $E_1 V_2 = E_1 RV = \{0\}$ and thus $\operatorname{Im} E_1 = E_1 V_1$ follows, and with this $V = V_0 \oplus V_1 \oplus V_2$.
2. Since $\ker\left(E_1 K^\dagger \mathbf{c} \cdot M\right) = \ker\left(K^\dagger \mathbf{c} \cdot M\right)$, it follows that

   $$\operatorname{Im}\left(E_1 K^\dagger \mathbf{c} \cdot M\right)^\dagger \cap \ker\left(K^\dagger \mathbf{c} \cdot M\right) = \{0\},$$

   and with that $\operatorname{Im}(G) = \operatorname{Im}\left(K^\dagger \mathbf{c} \cdot M\right)$. It follows from the definition of q that $G\mu = 0$ implies $\mu = 0$.
3. The order of magnitude of the subspaces follows directly from the definition of $G\mu$ as in (3.42).

□

Herewith, the structure of the distribution function has been deduced from the kinetic equation. It strongly depends on the collision operator K and arises from the requirement of a scale separation in the non-equilibrium subspace according to an asymptotic expansion in ε.

The typical separation in the phase space of the distribution function is that into equilibrium and non-equilibrium as in (3.16). The order-of-magnitude method given above now shows that there exists an additional natural separation of the non-equilibrium phase space that follows from the kinetic equation itself. The first order contribution opens a subspace V_1 in non-equilibrium that can be described by a low-dimensional set of moments μ that all scale by ε. The remainder space V_2 contains all high order contributions to the distribution function when Chapman-Enskog expanded.

The result is a scale-induced closure whose distribution structure strongly depends on the collision operator K. It is characterized not by slow and fast relaxation times but instead through the scale of the contributions of the asymptotic expansions to the distribution function.

Note that this construction is extendable to higher orders, leading to a more detailed decomposition of the non-equilibrium phase space V_{NE}.

In the derivation of G the higher moments $\mu = E_1 f$ are specified only by the solvability of the system (3.37). This is possible, but E_1 will not be unique. This situation is equivalent to the result (3.4) for the order-of-magnitude method applied to the moments directly. In (3.4) some moments had to be chosen as basis in which the others are represented. In the calculation of this representation, gradient expressions of equilibrium variables had to be eliminated.

3.5.3 Constructing the Moment Operator

To some extend the approach in Sect. 3.5.2 above mimics the procedure used in [49] and [50] conducted on the moment equations. An alternative path to exploiting the condition (3.35) is to first specify the distribution G and then derive the moment operator E_1.

The easiest way to construct G in accordance with (3.35) is to just choose $G = K^\dagger \mathbf{c} \cdot M$. However, having condition (3.31) in mind, $E_1 G = id_{\mathbb{R}^q}$ cannot be fullfilled if $K^\dagger \mathbf{c} \cdot M$ is not injective. To improve our definition of G, we introduce q linearly independent vectors which generate a complementary subspace to $\ker(K^\dagger \mathbf{c} \cdot M)$. Then we define the surjective map

$$D : \mathbb{R}^{p \times d} \to \mathbb{R}^q, \ \nabla \rho \mapsto \hat{\mu}, \tag{3.43}$$

such that $D(\ker(K^\dagger \mathbf{c} \cdot M)) = \{0\}$. Note that this leaves quite some freedom for the choice of D.

With D we can adjust our definition of G to

$$GD = -K^\dagger \mathbf{c} M \tag{3.44}$$

as an operator acting on $\nabla \rho$ according to (3.36), giving

$$GD \nabla \rho = -K^\dagger \mathbf{c} M \nabla \rho = -K^\dagger \mathbf{c} \cdot \nabla M \, \rho \tag{3.45}$$

in agreement with condition (3.33). Using a pseudoinverse of D the distribution G is explicitly given by

$$G = -K^\dagger \mathbf{c} M D^\dagger \tag{3.46}$$

as a mapping from $\mathbb{R}^q$ to V. In Sect. 3.5.4, we will compare (3.46) to (3.42), which was resulting from the choice of a specific operator E_1.

By construction, G is injective on the moment space $\mathbb{R}^q$ with

$$\operatorname{Im} G = \operatorname{Im}\left(K^\dagger \mathbf{c} M\right) = V_1 \subset V.$$

Hence $G : \mathbb{R}^q \to V_1$ is bijective and we can use its inverse to construct E_1 on V_1.

We remark that this definition automatically entails the condition $E_0 G = 0$: In fact, according to (3.14), we have $K^\dagger = P K^\dagger$ so that $G = PG$ and hence

$$E_0 G = E_0 P G = 0.$$

It remains to specify the moment mapping E_1 in such a way that the remaining conditions $E_1 M = 0$ and $E_1 G = id_{\mathbb{R}^q}$ in (3.31) are satisfied. While $E_1 M = 0$ fixes the behavior of E_1 on the equilibrium subspace V_0, the condition $E_1 G = id_{\mathbb{R}^q}$ shows that E_1 has to invert G on $V_1 = \operatorname{Im}(G)$. This can be summarized by saying that E_1 has to be a pseudoinverse of G whose kernel includes V_0.

The only information about the behavior of E_1 on subspaces complementary to $V_0 \oplus V_1$ is that V_2 should be the nullspace of the projection $S = GE_1$. Since G is injective, the nullspace of S is identical to the nullspace of E_1. Hence, the complete construction follows by choosing a space V_2 with the property $V_1 \oplus V_2 = V_{NE}$ and setting $E_1 = 0$ on $V_0 \oplus V_2$ and $E_1 = G^{-1}$ on V_1. Then all conditions on G and E_1 are satisfied. Summarizing, we obtain a decomposition of V into generally non-orthogonal subspaces

$$V = V_0 \oplus V_1 \oplus V_2, \qquad V_0 = QV, \quad V_{NE} = PV = V_1 \oplus V_2. \tag{3.47}$$

and

$$E_1 f = E_1(f_0 \oplus \varepsilon f_1^{(\mathrm{NE})} \oplus f_R) = G^{-1} \varepsilon f_1^{(\mathrm{NE})}, \quad \text{with } f_0 \in V_0,\ \varepsilon f_1^{(\mathrm{NE})} \in V_1,\ f_R \in V_2. \tag{3.48}$$

If we get orthogonal sums in (3.47), then $S = GE_1$ is a symmetric projector. With that, additional to (3.31) and (3.48), we obtain the unique Moore-Penrose-inverse $E_1 = G^\dagger$, see Sect. 3.3. For a proof see App. A.3 and [62].

In Sect. 3.7 we will show that the specific construction leading to $E_1 = G^\dagger$ as above produces desirable properties of the evolution equation for ρ and μ. However, if not stated otherwise, we will use a general non-orthogonal decomposition $V = V_0 \oplus V_1 \oplus V_2$.

From the construction of E_1 we can again clearly see that the order of magnitude method is based on a natural separation of the non-equilibrium phase space V_{NE} that follows

from the kinetic equation itself. It should be noted that, also here, the construction of E_1 is not unique due to some arbitrariness in the choice of D and, following from this, the moment space $\mathbb{R}^q$. This situation corresponds again to the result (3.4) for the order-of-magnitude method applied to the moments directly. In (3.4) some moments had to be chosen as basis in which the others are represented.

3.5.4 Comparison

In Sect. 3.5.2 we started with the construction of a projection $E_1 : V \to \mathbb{R}^q$ with certain restrictions and computed G from it, whereas in Sect. 3.5.3 above, we started with the specification of $G : \mathbb{R}^q \to V$ by choosing the operator D and then determined E_1 as inverse of G. In both cases the restriction of $G\mu$ as obtained in (3.35) was used.

The following Lemma shows how the constructions in Sect. 3.5.2 and Sect. 3.5.3 are related.

Lemma 3.5.2 (Relation of Different Constructions)**.**
Let $G\mu$ be determined through (3.35)*.*

1. *Consider the derivation in Sect. 3.5.3. If an appropriate operator D as in (3.43) gives rise to a distribution G as in (3.46) and a moment operator E_1 as in (3.48), then*
$$D = -E_1 K^\dagger \mathbf{c} \cdot M \quad \text{and} \quad G = K^\dagger \mathbf{c} \cdot M \left(E_1 K^\dagger \mathbf{c} \cdot M \right)^\dagger \tag{3.49}$$
in agreement with the definition (3.42) of G in the derivation of Sect. 3.5.2.
2. *Consider the derivation in Sect. 3.5.2. If a moment operator E_1 satisfying the condition described in Sect. 3.5.2 gives rise to the distribution G as in (3.42) and additionally the projector GE_1 is symmetric, then there exists an operator $E_1^{5.3} = G^{-1}|_{V_1}$ as in* (3.48)*. In particular, $E_1^{5.3} = G^\dagger = E_1$ and the two approaches agree.*

Proof.

1. From (3.44) it follows $D = -G^\dagger K^\dagger \mathbf{c} \cdot M$ and since $K^\dagger \mathbf{c} \cdot M$ maps to V_1, we have $D = -E_1 K^\dagger \mathbf{c} \cdot M$. The second equality follows with (3.46).
2. $E_1^{5.3} = G^\dagger = E_1$ follows from the symmetry of S together with condition (3.31), stating $E_1 G = id_{\mathbb{R}^q}$. In the case where V is finite dimensional, this is a standard argument using singular value decomposition. For details see App. A.3. If V is a generally infinite dimensional Hilbert space, we refer to Theorem 9.1.3 in [62].

□

Remark: Note that starting with E_1 as in Sect. 3.5.2, then, according to (3.42), constructing $G = K^\dagger \mathbf{c} \cdot M \left(E_1 K^\dagger \mathbf{c} \cdot M\right)^\dagger$, and finally defining $\widetilde{E}_1$ as in (3.48) does not necessarily yield $\widetilde{E}_1 = E_1$, unless the appearing pseudo-inverses are consistently chosen, which automatically happens in the orthogonal case.

Usually, the approach in Sect. 3.5.3 is less practical since typically the distribution function is to be constructed after the choice of specific moments to describe the process.

Note once more that this situation is equivalent to the result (3.4) for the order-of-magnitude method applied to the moments directly. In (3.4) some moments had to be chosen as basis in which the others are represented. In the calculation of this representation, gradient expressions of equilibrium variables had to be eliminated.

Finally, we want to stress that the basic idea of the construction presented here is extendable to higher orders, leading to a more detailed decomposition of the non-equilibrium phase space V_{NE}, see Sect. 3.8.

3.6 Asymptotic Order

Assuming, as in Grad's closure, $f = M\,\rho + G\,\mu$ with G and E_1 satisfying (3.48), we find the evolution equations

$$\partial_t \rho + E_0 \mathbf{c} \cdot \nabla M\,\rho + E_0 \mathbf{c} \cdot \nabla G\,\mu = 0 \tag{3.50}$$

$$\partial_t \mu + E_1 \mathbf{c} \cdot \nabla M\,\rho + E_1 \mathbf{c} \cdot \nabla G\,\mu + \frac{1}{\varepsilon} E_1 K G\,\mu = 0. \tag{3.51}$$

Note that in accordance with (3.29/3.30) and (3.33), $\mu = \varepsilon\mu_1 + \mathcal{O}(\varepsilon^2)$. We have chosen this scaling to compare (3.50/3.51) to Grad's equations (3.27/ 3.28).
We are interested in the asymptotic accuracy in terms of powers of ε of the evolution of ρ with respect to the full kinetic equation. The question is, whether the evolution for μ in (3.51) when expanded in ε and inserted into (3.50) reproduces the equations for ρ resulting from the Chapman-Enskog expansion of the full kinetic model.

3.6.1 Order Analysis

The following theorem completely characterizes the asymptotic behavior of the system (3.50)/(3.51).

Theorem 3.6.1 (Asymptotic Accuracy)**.**
The system (3.50)/(3.51) with primary non-equilibrium distribution G and moment operator E_1 satisfying (3.48) describes an evolution of ρ that is of the following Chapman-Enskog orders:

1) *first order in the Knudsen number ε, if the operator $E_1 K G$ is invertible on $\mathbb{R}^q$.*

2) *second order in ε, if E_1KG is invertible on $\mathbb{R}^q$, and if*

$$\widetilde{E}K^\dagger R = 0 \qquad \textit{and} \qquad \widetilde{E}K^\dagger G = \widetilde{E}G\,(E_1KG)^{-1}\,, \tag{3.52}$$

where $\widetilde{E} = E_0\mathbf{c}$.

Proof.

1) The kinetic model produces an evolution for ρ that is given by (3.24). We have to show that, up to first order, the asymptotic expansion of μ leads to the same equation. We introduce $\mu = \varepsilon\mu_1 + \varepsilon^2\mu_2$ into (3.51) and obtain for the first order contribution

$$E_1\mathbf{c}\cdot\nabla M\,\rho + E_1KG\,\mu_1 = 0. \tag{3.53}$$

We note that multiplying $E_1M = 0$ with E_0 from the right yields $E_1Q = 0$ so that $E_1P = E_1$. Using further $KK^\dagger = P$, we obtain $E_1 = E_1KK^\dagger$ and hence (3.53) reads

$$E_1K\left(K^\dagger\mathbf{c}\cdot\nabla M\,\rho + G\,\mu_1\right) = 0 \tag{3.54}$$

The expression in the bracket is contained in $V_1 = \mathrm{Im}(G) = \mathrm{Im}(K^\dagger\mathbf{c}\cdot\nabla M)$ and since $SV_1 = V_1$, we have

$$E_1KS\left(K^\dagger\mathbf{c}\cdot\nabla M\,\rho + G\,\mu_1\right) = 0. \tag{3.55}$$

Using the definition $S = GE_1$, the invertibility of E_1KG and the relation $E_1G = id_{\mathbb{R}^q}$, we conclude

$$\mu_1 = -E_1K^\dagger\mathbf{c}\cdot\nabla M\,\rho \tag{3.56}$$

and with it $G\mu_1 = -K^\dagger\mathbf{c}\cdot\nabla M\,\rho$.

Using this result, we can compute the leading order of the μ dependent expression in (3.50)

$$E_0\mathbf{c}\cdot\nabla G\mu = \varepsilon\,E_0(\mathbf{c}\cdot\nabla)G\mu_1 + O(\varepsilon^2) \tag{3.57}$$

$$= -\varepsilon\,E_0(\mathbf{c}\cdot\nabla)K^\dagger(\mathbf{c}\cdot\nabla)M\,\rho + O(\varepsilon^2) \tag{3.58}$$

which is precisely the first order contribution given in (3.21).

2) Balancing the next order of (3.51) yields

$$\partial_t\mu_1 + E_1(\mathbf{c}\cdot\nabla)G\mu_1 + E_1KG\mu_2 = 0 \tag{3.59}$$

where μ_1 has to be inserted from above. The relevant term that enters the evolution of ρ reads

$$\widetilde{E}\cdot\nabla G\mu_2 = -\widetilde{E}\cdot\nabla G\,(E_1KG)^{-1}\,(\partial_t\mu_1 + E_1(\mathbf{c}\cdot\nabla)G\mu_1) \tag{3.60}$$

The relation corresponding to (3.59) within the Chapman-Enskog expansion of the full kinetic equation can be written

$$\partial_t \left(-K^\dagger \mathbf{c} \cdot \nabla M \, \rho \right) + \mathbf{c} \cdot \nabla \left(-K^\dagger \mathbf{c} \cdot \nabla M \, \rho \right) + K f_2^{(\mathrm{NE})} = 0$$

where the expression in brackets can be replaced by $G\mu_1$ in the current context. As in Sect. 3.4.2, we can multiply by $K^\dagger$ and drop the non-equilibrium projection in front of $f_2^{(NE)}$, which yields

$$f_2^{(\mathrm{NE})} = -K^\dagger G \left(\partial_t \mu_1 + E_1 (\mathbf{c} \cdot \nabla) G \mu_1 \right) - K^\dagger R \left(\mathbf{c} \cdot \nabla \right) G \mu_1$$

which influences the evolution of ρ in the form

$$\widetilde{E} \cdot \nabla f_2^{(\mathrm{NE})} = -\widetilde{E} \cdot \nabla K^\dagger G \left(\partial_t \mu_1 + E_1 (\mathbf{c} \cdot \nabla) G \mu_1 \right) - \widetilde{E} \cdot K^\dagger R \left(\mathbf{c} \cdot \nabla \right) G \mu_1. \quad (3.61)$$

Equality with the expression obtained for μ_2 is given if

$$\widetilde{E} \cdot \nabla G \left(E_1 K G \right)^{-1} = \widetilde{E} \cdot \nabla K^\dagger G - \widetilde{E} \cdot K^\dagger R \left(\mathbf{c} \cdot \nabla \right) G. \quad (3.62)$$

This is guaranteed by the assumptions.

□

The theorem shows that the system resulting from the scale induced closure can be of second order, that is, of the same accuracy as the Burnett equations, and as such go beyond the first order Navier-Stokes equations. The scale induced closure combines Grad's method with the asymptotic properties of a Chapman-Enskog expansion.

Second order is given in non-trivial cases where the special conditions given in the theorem are satisfied. In general, additional expressions have to be added on the right hand side of the system stemming from higher moment equations. This can be seen in [50] where generalized 13-moment-equations are derived. Interestingly, for Maxwell-molecules the condition for direct second order is satisfied, hence the original Grad equations are of Burnett order, see [51]. For a given K, sufficient and also necessary conditions for any order can be obtained through direct comparison of the asymptotic expansion with the corresponding Chapman-Enskog expansion of (3.8). This is exemplified in App. A.4.3.

Note that the operator $\widetilde{E} = E_0 \mathbf{c}$ can be interpreted as equilibrium projection of higher moments. By asking that $\widetilde{E} K^\dagger R = 0$, we demand in a weak sense that $K^\dagger$ does not map any elements of the secondary non-equilibrium V_2 to a lower order (non-)equilibrium. This is also related to condition 3) in Sect. 3.6.2 below.

3.6.2 Various Conditions for First and Higher Order

The conditions given in Theorem 3.6.1 are sufficient. In this section we give an overview of some more sufficient conditions for first and even higher order.

We begin with analyzing the mapping KG. To check injectivity, we note that $KG\mu = 0$ implies $G\mu \in \ker(K)$, i.e. $G\mu = M\rho$ for some $\rho \in \mathbb{R}^p$, so that condition (3.48) yields $\mu = E_1 M\rho = 0$. Hence, $KG\mathbb{R}^q = KV_1 = PKV_1 \subset V_{NE}$ is a q-dimensional subspace of V_{NE}. Now there are two possibilities depending on whether the intersection of KV_1 and V_2 is empty or not. Interestingly, this alternative decides about the first order accuracy condition.

Lemma 3.6.2. *The following conditions are equivalent*

$$1)\, E_1KG \text{ invertible} \qquad 2)\, V_1 = SKV_1 \qquad 3)\, KV_1 \cap V_2 = \{0\}.$$

Proof. Assuming (1), we see that $G(E_1KG)$ is injective. Since $S = GE_1$ is a projection onto V_1, we conclude that $SKV_1 = SKG\mathbb{R}^q$ is a q-dimensional subspace of V_1 and thus identical to V_1. Next, we assume (2) and $KV_1 \cap V_2 \neq \{0\}$ for a proof by contradiction. Then there exists some $f = G\mu \in V_1$ such that $0 \neq Kf \in V_2$ and hence $SKf = 0$ which shows that $SKG : \mathbb{R}^q \to V_1$ is not injective. Consequently, it cannot be surjective which contradicts the assumption (2). Finally assuming (3), we check the injectivity of E_1KG. If $\mu \in \mathbb{R}^q$ satisfies $E_1KG\mu = 0$, then also $SKG\mu = GE_1KG\mu = 0$ so that $KG\mu \in \ker(S) = V_2$. Since also $KG\mu \in KV_1$ we find $KG\mu = 0$. The injectivity of KG then yields $\mu = 0$ which finally shows (1). □

In view of this reformulation, we should recall the construction of the operator E_1. In the construction, we had the freedom to choose V_2 and now we see that it may be beneficial to select V_2 in such a way that it intersects KV_1 only at the origin.

In order to satisfy the second order condition (3.52) in Thm. 3.6.1, we can choose V_1 as an invariant subspace of K, i.e.

$$KV_1 = V_1,$$

This works because

$$G = PG = K^\dagger KG = K^\dagger SKG = K^\dagger GE_1KG,$$

and by applying the inverse of E_1KG, we find

$$G(E_1KG)^{-1} = K^\dagger G,$$

which is sufficient for the second condition in (3.52). If we have in addition $\widetilde{E}K^\dagger R = 0$, we get second order accuracy if V_1 is an invariant subspace of K. Note that asking $KV_1 = V_1$ is stronger a condition than just requiring $SKV_1 = V_1$.

Another sufficient condition for first, and in fact also higher order is given by

$$G(E_1KG)^{-1}E_1 = K^\dagger$$

which needs the invertibility of E_1KG but is a much stronger condition. In fact, one can show with asymptotic expansion that it implies $V_1 = V_{NE}$, or equivalently $V_2 = \{0\}$. Hence, we easily see that it leads to arbitrary accuracy, since higher order contributions are trivial. For our purpose, however, this case is of little interest, because the complexity of the kinetic equation is not reduced by applying the scale induced closure.

3.6.3 Regularization

The above construction uses the zeroth order of the evolution of the distribution $\hat{f}_R$ in (3.32). The first order gives additional accuracy and leads to regularized equations similar to the R13-system in [49] and [52].

Under the first order accuracy condition, the evolution of $\hat{f}_R$ in (3.32) reduces to

$$\varepsilon^2\partial_t\hat{f}_R + \varepsilon^2 R\mathbf{c}\cdot\nabla\hat{f}_R + \varepsilon R\mathbf{c}\cdot\nabla G\,\hat{\mu} + \varepsilon RK\,\hat{f}_R = 0 \tag{3.63}$$

where zero-order terms have vanished. The first order terms in this equation are balanced by choosing

$$RK\hat{f}_R = -R\mathbf{c}\cdot\nabla G\hat{\mu} \tag{3.64}$$

Assuming that $(RK)^\dagger RK\hat{f}_R = \hat{f}_R$, we can write

$$\hat{f}_R = -(RK)^\dagger R\mathbf{c}\cdot G\nabla\hat{\mu}. \tag{3.65}$$

This gives a first contribution to the secondary non-equilibrium in (3.30) based on gradients of the non-equilibrium variables μ. The elaboration of this procedure is left for future work. In fact, an additional first order term has been suppressed above. This regularization procedure has been successfully conducted on the moment hierarchy for Maxwell-molecules in [49].

3.7 Stability

In this section we assume that

1) V is a Hilbert space with scalar product $\langle\cdot,\cdot\rangle_V$.
2) The collision operator K and the equilibrium projection Q are self-adjoint. Furthermore K is positive semi-definite.
3) The multiplication operator $c_\alpha : V \to V$ defined by $(c_\alpha f)(\mathbf{v}) = v_\alpha f(\mathbf{v})$ is self-adjoint.
4) The projector $S = GE_1$ as constructed in Sect. 3.5.1 - 3.5.4 is self-adjoint.

(3.66)

The distribution M is defined in Sect. 3.3.1. After introducing symmetric positive definite operators based on the adjoints of the distributions, we will show that the equations (3.50)/(3.51) possess an entropy law and are therefore stable.

Definition 3.7.1 (Adjoint)**.**
We denote $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ the standard scalar product in $\mathbb{R}^n$

1) *We define the adjoint M^* of M through the Riesz representation theorem as the unique linear operator $M^* : V \to \mathbb{R}^p$ such that*

$$\langle x, M^* f\rangle_{\mathbb{R}^p} = \langle Mx, f\rangle_V, \qquad \forall f \in V,\ x \in \mathbb{R}^p \tag{3.67}$$

2) *Similarly we define $G^* : V \to \mathbb{R}^q$, such that*

$$\langle y, G^* f\rangle_{\mathbb{R}^q} = \langle Gy, f\rangle_V, \qquad \forall f \in V,\ y \in \mathbb{R}^q \tag{3.68}$$

3) *Based on the adjoints we define*

$$B := M^*M : \mathbb{R}^p \to \mathbb{R}^p \text{ and } L := G^*G : \mathbb{R}^q \to \mathbb{R}^q. \tag{3.69}$$

The matrices B and L are symmetric, positive definite by construction. They will give rise to symmetric forms which constitute the entropy of the moment system. Again we keep notation simple, and use μ for the scaled higher moments $\hat{\mu}$ in the following. Generally the stability result does not depend on the scaling of μ.

In this orthogonal setting, the pseudoinverse $G^\dagger$ is the unique Moore-Penrose inverse. Due to injectivity of G, it can be computed as $G^\dagger = (G^*G)^{-1}G^*$. Furthermore we have through (3.66) 2) and 4) that indeed E_0 and E_1 are the unique Moore-Penrose inverses of M and G respectively (see Sect. 3.3 and 3.5.4).

Theorem 3.7.2 (Stability)**.**
Let the moment system (3.50)/(3.51) be given, based on the operators M, G and E_0, E_1 defined in Sect. 3.3 and 3.5.3. Then the system features the convex entropy

$$\eta = \frac{1}{2}\langle\rho, B\rho\rangle_{\mathbb{R}^p} + \frac{1}{2}\langle\mu, L\mu\rangle_{\mathbb{R}^q}, \qquad \rho \in \mathbb{R}^p, \quad \mu \in \mathbb{R}^q \tag{3.70}$$

with associated negative definite entropy production. In particular, the system is symmetric hyperbolic[1].

Proof.

Convexity of η: $\langle\cdot, B\cdot\rangle_{\mathbb{R}^p}$ and $\langle\cdot, L\cdot\rangle_{\mathbb{R}^q}$ define scalar products based on the symmetric, positive definite matrices from (3.69). The function η is therefore convex. Note that η shows similarities with the entropies in [7] and [53].

[1] For an overview of hyperbolicity, entropy and stability, see Sect. A.2.

Entropy law: Multiplying (3.50) with $\langle \rho, B\cdot\rangle$ and (3.51) with $\langle \mu, L\cdot\rangle$ yields

$$\langle \rho, B\partial_t \rho\rangle_{\mathbb{R}^p} + \langle \rho, BE_0\mathbf{c}\cdot M\,\nabla\rho\rangle_{\mathbb{R}^p} + \langle \rho, BE_0\mathbf{c}\cdot G\,\nabla\mu\rangle_{\mathbb{R}^p} = 0 \tag{3.71a}$$

$$\langle \mu, L\partial_t \mu\rangle_{\mathbb{R}^q} + \langle \mu, LE_1\mathbf{c}\cdot M\,\nabla\rho\rangle_{\mathbb{R}^q} + \langle \mu, LE_1\mathbf{c}\cdot G\,\nabla\mu\rangle_{\mathbb{R}^q} = -\frac{1}{\varepsilon}\langle \mu, LE_1KG\,\mu\rangle_{\mathbb{R}^q} \tag{3.71b}$$

From the definition of B we immediately see

$$BE_0 = M^*ME_0 = M^*Q = (QM)^* = M^* \tag{3.72}$$

since Q is self-adjoint.

For the product LE_1 we analogously find with $(3.66)_4$ that

$$LE_1 = G^*GE_1 = G^*S = (SG)^* = G^*. \tag{3.73}$$

Plugging in these expressions yields

$$\langle \rho, B\partial_t \rho\rangle_{\mathbb{R}^p} + \langle \rho, M^*\mathbf{c}\cdot M\,\nabla\rho\rangle_{\mathbb{R}^p} + \langle \rho, M^*\mathbf{c}\cdot G\,\nabla\mu\rangle_{\mathbb{R}^p} = 0 \tag{3.74a}$$

$$\langle \mu, L\partial_t \mu\rangle_{\mathbb{R}^q} + \langle \mu, G^*\mathbf{c}\cdot M\,\nabla\rho\rangle_{\mathbb{R}^q} + \langle \mu, G^*\mathbf{c}\cdot G\,\nabla\mu\rangle_{\mathbb{R}^q} = -\frac{1}{\varepsilon}\langle \mu, G^*KG\,\mu\rangle_{\mathbb{R}^q} \tag{3.74b}$$

After adding the equations (3.74) and using the self-adjointness of L, B and $\mathbf{c}$, we obtain

$$\begin{aligned}
&\partial_t\left(\frac{1}{2}\,\langle \rho, B\rho\rangle_{\mathbb{R}^p} + \frac{1}{2}\,\langle \mu, L\mu\rangle_{\mathbb{R}^q}\right)\\
&+\nabla\cdot\left(\frac{1}{2}\,\langle \rho, M^*\mathbf{c}M\,\rho\rangle_{\mathbb{R}^p} + \langle \rho, M^*\mathbf{c}G\,\mu\rangle_{\mathbb{R}^p} + \frac{1}{2}\,\langle \mu, G^*\mathbf{c}G\,\mu\rangle_{\mathbb{R}^q}\right)\\
&= -\frac{1}{\varepsilon}\,\langle \mu, G^*KG\,\mu\rangle_{\mathbb{R}^q}
\end{aligned} \tag{3.75}$$

which is an entropy law.

Negativity of entropy production: Since K is self-adjoint and positive semidefinite, we have $G^*KG = (\sqrt{K}G)^*\sqrt{K}G$. Hence, we can rewrite $\langle \mu, G^*KG\,\mu\rangle_{\mathbb{R}^q} = \langle \sqrt{K}G\mu, \sqrt{K}G\,\mu\rangle_V > 0$, since K is positive on the range of G. Therefore the entropy production is negative definite.

□

With Theorems 3.6.1 and 3.7.2 we have shown that the scale induced closure yields equations which are physically accurate in terms of a Knudsen number expansion as well as mathematically stable.

We have seen before that, in the new theory, the definition of the distribution function $G\,\mu$ depends on the structure of the collision operator. This is a natural outcome since the scale decompostion of the non-equilbrium phase space is induced by the collision operator. To find a symmetric projector GE_1, additional constraints on the choice of the variables μ follow. Hence, also the choice of variables μ is governed by K which is somewhat surprising. For the moments in (3.4) this results in an additional recombination of the vectors and tensors to find an appropriate basis.

3.8 Higher Order Scale Induced Closure

The orignial order-of-magnitude method developed by Struchtrup in [49] is a 3rd order approximation. In this section we sketch how to construct the scale induced closure including contributions of order up to ε^n in our linear case. This is just an outline and shall serve as starting point for future research. Many details remain unspecified.

Extending (3.30), we decompose

$$f = M\rho + \varepsilon G_1\hat{\mu}_1 + \varepsilon^2 G_2\hat{\mu}_2 + \cdots \varepsilon^n G_n\hat{\mu}_n + \varepsilon^{n+1}\hat{f}_{R_n}, \tag{3.76}$$

with moments $\hat{\mu}_k \in \mathbb{R}^{q_k}$. The corresponding operators E_1,...,E_n and G_1,...,G_n are required to fulfill

$$E_k G_k = id_{\mathbb{R}^{q_k}}, \quad E_k M = 0, \quad E_0 G_k = 0, \quad E_i G_j = 0, \tag{3.77}$$

where i, j, $k = 1, ..., n$ and $i \neq j$.

With these operators, we can construct the projections $S_k = G_k E_k$, $k = 1, ..., n$ and $R_n = P - S_1 - ... - S_n$, such that our phase space is divided as in Fig. 3.2.

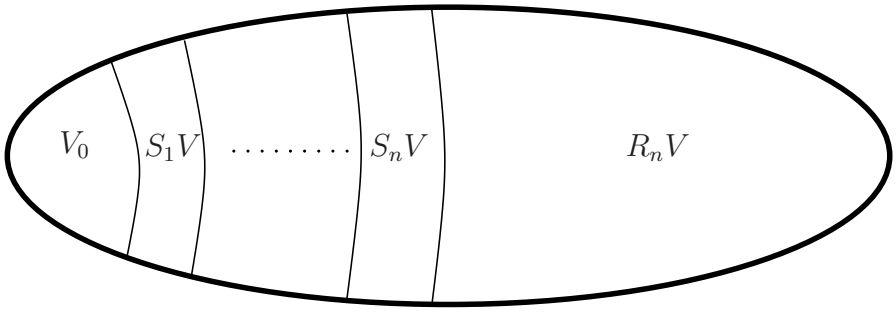


Figure 3.2: The phase space is subdivided into various higher non-equilibrium parts.

In accordance with the construction in Sect. 3.5.1, we match orders as in (3.32) and derive the equations determining the $G_k\hat{\mu}_k$ as

$$\begin{aligned} G_1\hat{\mu}_1 &= -K^\dagger \mathbf{c}\cdot M\nabla\rho \\ G_2\hat{\mu}_2 &= -K^\dagger \mathbf{c}\cdot G_1\nabla\hat{\mu}_1 \\ &\vdots \\ G_n\hat{\mu}_n &= -K^\dagger \mathbf{c}\cdot G_{n-1}\nabla\hat{\mu}_{n-1} \end{aligned} \tag{3.78}$$

Under certain conditions on the relation between the image sets of $K^\dagger \mathbf{c} \cdot G_i$ and S_j, the operators can be constructed analogously to Sect. 3.5.2 and Sect. 3.5.3.

The equations are derived by plugging (3.76) into the linear kinetic equation, applying the operators E_0,...,E_n and setting $f_{R_n} = 0$. Using scaled variables $\mu_k = \varepsilon^k \hat{\mu}_k$, $k = 1, ..., n$, they read

$$\begin{aligned}
\partial_t \rho + E_0 \mathbf{c} \cdot \nabla M\rho + E_0 \mathbf{c} \cdot \nabla G_1 \mu_1 + \cdots + E_0 \mathbf{c} \cdot \nabla G_n \mu_n &= 0 \\
\partial_t \mu_1 + E_1 \mathbf{c} \cdot \nabla M\rho + E_1 \mathbf{c} \cdot \nabla G_1 \mu_1 + \cdots + E_1 \mathbf{c} \cdot \nabla G_n \mu_n & \\
+ \frac{1}{\varepsilon} E_1 K G_1 \mu_1 + \cdots + \frac{1}{\varepsilon} E_1 K G_n \mu_n &= 0 \\
\vdots & \\
\partial_t \mu_n + E_n \mathbf{c} \cdot \nabla M\rho + E_n \mathbf{c} \cdot \nabla G_1 \mu_1 + \cdots + E_n \mathbf{c} \cdot \nabla G_n \mu_n & \\
+ \frac{1}{\varepsilon} E_n K G_1 \mu_1 + \cdots + \frac{1}{\varepsilon} E_n K G_n \mu_n &= 0
\end{aligned} \tag{3.79}$$

3.8.1 Stability for the Higher Order case

The stability analysis in the higher order case uses similar arguments as in Sect. 3.7. We need the first 3 assumptions of (3.66), assumption 4) needs to be extended to all the projectors S_1,...,S_n, and the positive semi-definiteness of K will need some extension as well. The entropy becomes

$$\eta_n = \frac{1}{2}\langle \rho, B\rho \rangle_{\mathbb{R}^p} + \frac{1}{2}\langle \mu_1, L_1 \mu_1 \rangle_{\mathbb{R}^{q_1}} + \cdots + \frac{1}{2}\langle \mu_n, L_n \mu_n \rangle_{\mathbb{R}^{q_n}}, \quad \rho \in \mathbb{R}^p, \quad \mu_k \in \mathbb{R}^{q_k}, \tag{3.80}$$

where $L_k = G_k^* G_k$ and $k = 1, ..., n$.

We proceed analogously to the proof of Thm. 3.7.2 and get the entropy flux

$$\begin{aligned}
\frac{1}{2}\langle \rho, M^* \mathbf{c} M\rho \rangle_{\mathbb{R}^p} + \sum_{k=1}^{n} \langle \rho, M^* \mathbf{c} G_k \mu_k \rangle_{\mathbb{R}^p} + \frac{1}{2}\sum_{k=1}^{n} \langle \mu_k, G_k^* \mathbf{c} G_k \mu_k \rangle_{\mathbb{R}^{q_k}} & \\
+ \sum_{j=1}^{n} \sum_{i=j+1}^{n} \langle \mu_j, G_j^* \mathbf{c} G_i \mu_i \rangle_{\mathbb{R}^{q_j}}. &
\end{aligned} \tag{3.81}$$

On the right hand side of the entropy equation, we get

$$-\frac{1}{\varepsilon}\sum_{j=1}^{n}\sum_{k=1}^{n} \langle \mu_j, G_j^* K G_k \mu_k \rangle_{\mathbb{R}^{q_j}}. \tag{3.82}$$

Since this quantity should be negative, assumption 2) in (3.66) extends to negativity not only of $-K$, but of the above combination. This is satisfied, for example, if K is self adjoint and negative definite on $V_0^\perp$ and if the subspaces S_iV are K-invariant for $i = 1, ..., n$.

If all these assumptions are met, we obtain an entropy law with negative entropy production, and therefore symmetric hyperbolic and stable equations.

3.8.2 Order Analysis

The order analysis becomes even more technical for $n \geq 2$ than it is in Sect. 3.6. A general analysis is therefore beyond the scope of this work. For examples like the following 16 discrete velocities model in Sect. 3.9.2, the easiest way to check the order of accuracy is a direct asymptotic expansion of the equations under consideration, see App. A.4.3. Note however that the 16 discrete velocities models is not complex enough to serve as a good testcase for higher orders in the scale induced closure.

3.9 Examples

As examples for the theory described above we discuss three specific cases. One displays the generalized 13-moment-case. Then we consider a model with 16 discrete velocities and show the approximation features of the various closures. The last case demonstrates the accuracy of the closure approximations in the case of a generic linear model system.

3.9.1 Generalized 13-Moment-Equations

The generalized 13-moment-equations have been derived by Struchtrup in [50] by the order-of-magnitude approach described above. In the derivation a general interaction potential has been assumed and, hence, general production terms in the moment equations have been considered. The closure approximation takes into account the structure of the production terms and the resulting coefficients could be identified with classical Burnett coefficients.

The final equations for stress tensor σ_{ij} and heat flux q_i read

$$\frac{D\sigma_{ij}}{Dt} + \sigma_{ij}\frac{\partial v_k}{\partial x_k} + 2\sigma_{k\langle i}\frac{\partial v_{j\rangle}}{\partial x_k} + \Pr\frac{4\varpi_3}{5\varpi_2}\left(\frac{\partial q_{\langle i}}{\partial x_{j\rangle}} - \omega\, q_{\langle i}\frac{\partial \ln\theta}{\partial x_{j\rangle}}\right)$$
$$+ \Pr\frac{4\varpi_4}{5\varpi_2}q_{\langle i}\frac{\partial \ln p}{\partial x_{j\rangle}} + \Pr\frac{4\varpi_5}{5\varpi_2}q_{\langle i}\frac{\partial \ln\theta}{\partial x_{j\rangle}} + \frac{\varpi_6}{\varpi_2}\sigma_{k\langle i}S_{j\rangle k} = -\frac{2p}{\varpi_2\mu}\left[\sigma_{ij} + 2\mu\frac{\partial v_{\langle i}}{\partial x_{j\rangle}}\right] \quad (3.83)$$

and

$$\frac{Dq_i}{Dt} + q_k\frac{\partial v_i}{\partial x_k} + \frac{5}{3}q_i\frac{\partial v_k}{\partial x_k} - \frac{5}{2\,\mathrm{Pr}}\sigma_{ik}\frac{\partial\theta}{\partial x_k} + \frac{5\theta_3}{4\theta_2\,\mathrm{Pr}}\sigma_{ik}\frac{\partial \ln p}{\partial x_k}$$
$$+ \frac{5\theta_4}{4\theta_2\,\mathrm{Pr}}\theta\left(\frac{\partial\sigma_{ik}}{\partial x_k} - \omega\sigma_{ik}\frac{\partial \ln\theta}{\partial x_k}\right) + \frac{15\theta_5}{4\theta_2\,\mathrm{Pr}}\sigma_{ik}\frac{\partial\theta}{\partial x_k} = -\frac{5p}{2\theta_2\,\mathrm{Pr}\,\mu}\left[q_i + \frac{5\mu}{2\,\mathrm{Pr}}\frac{\partial\theta}{\partial x_i}\right] \quad (3.84)$$

where v_i is the velocity, θ is the temperature (in energy units), p is the pressure, Pr is the Prandtl number, μ viscosity and ϖ_α, θ_α are Burnett coefficients. Interestingly, for Maxwell molecules the equations reduce to the 13-moment-system of Grad which is based on a Hermite series of the distribution function. This is a mere coincidence and related to the fact that the eigenfunctions of the linearized collision operator for Maxwell molecules are Hermite functions. Hence, Grad's equations form the accurate second order system only for Maxwell molecules while the above system is the second order accurate extension to general interaction potentials. I.e., it is a stable system that reproduces the correct general Burnett relations when expanded in Knudsen number. In that sense it is also related to the regularized Burnett equations in [29].

The system demonstrates the capabilities of the described scale-induced closure procedure.

3.9.2 Linearized 16 Discrete Velocity Model

The following example considers a linearized 16 discrete velocities model in one space and two velocity dimensions [3]. Such models are a generalization of models initially developed in [10]. They have been more thoroughly investigated in [3].

Choosing the bilinear interactions according to Fig. 2.4 in the Introduction, Sect. 2.2.4, we obtain the kinetic equations

$$\partial_t u_i(x,t) + \sum_{j=1}^{16} V_{ij}\partial_x u_j(x,t) + \frac{1}{\varepsilon}K_i^{\mathrm{nonlin}}[u] = 0, \quad (3.85)$$

with $V_{ij} = \delta_{ij}c_i^{(1)}$, $K^{\mathrm{nonlin}} = K^{\mathrm{diag}} + K^{\mathrm{straight}}$ being positive semidefinite bilinear forms (see Appendix A.4.1, equations (A.22) and (A.23)). We linearize (3.85) around a constant equilibrium $f_i^0 = 1$ by the ansatz $u_i = 1 + \varepsilon f_i$ and neglect all higher order terms. This leads again to a positive semidefinite linear map $K : V \to V$. The linear equations read

$$\partial_t f_i(x,t) + \sum_{j=1}^{16} V_{ij}\partial_x f_j + \frac{1}{\varepsilon}\sum_{j=1}^{16} K_{ij}f_j = 0, \quad (3.86)$$

with K as in (A.24).
The nullspace of K defines the equilibrium moments[2]

$$\rho = \sum_{i=1}^{16} f_i, \quad \rho v_x = \sum_{i=1}^{16} c_i^{(1)} f_i, \quad \rho v_y = \sum_{i=1}^{16} c_i^{(2)} f_i, \quad e = \sum_{i=1}^{16} c_i^2 f_i. \tag{3.87}$$

The orthogonal complement of the nullspace of K is spanned by the arbitrarily chosen vectors r_1,...,r_{12}.

For the detailed computations leading to the form of the classical equations in the following subsections we refer to A.4.2. Here we only give the results.

Euler Equations

The Euler equations (3.17) become

$$\partial_t \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{66}{5} & 0 & 0 \end{pmatrix} \partial_x \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.88}$$

Navier-Stokes-Fourier System

With the pseudoinverse of K we get the Navier Stokes Fourier equations according to (3.21)

$$\partial_t \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{66}{5} & 0 & 0 \end{pmatrix} \partial_x \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} = \varepsilon \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{4}{5} & 0 & 0 \\ 0 & 0 & \frac{289}{20} & 0 \\ -\frac{140}{5} & 0 & 0 & \frac{14}{5} \end{pmatrix} \partial_x^2 \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} \tag{3.89}$$

Burnett Equations

From (3.24), the Burnett equations turn out to be

$$\begin{aligned} \partial_t \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{66}{5} & 0 & 0 \end{pmatrix} \partial_x \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} &= \varepsilon \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{4}{5} & 0 & 0 \\ 0 & 0 & \frac{289}{20} & 0 \\ -\frac{140}{5} & 0 & 0 & \frac{14}{5} \end{pmatrix} \partial_x^2 \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} \\ &\quad - \varepsilon^2 \begin{pmatrix} 0 & 0 & 0 & 0 \\ -\frac{19}{4} & 0 & 0 & \frac{27}{40} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1354}{125} & 0 & \frac{14}{5} \end{pmatrix} \partial_x^3 \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \end{pmatrix} \end{aligned} \tag{3.90}$$

[2]Note that left and right eigenvectors are equal since K is symmetric.

We will see later in Fig. 3.3 and Fig. 3.5 that the Burnett equations lead to instabilities, as is also observed in [7].

Grad Equations

To obtain a Grad Closure, we have to choose some higher moments through the operators G and E_1, satisfying the constraints given in section 3.4.3. We will argue that the scale induced closure produces a set of 3 higher moments, so in order to have a fair comparison, we chose the same number for Grad.

We will choose these moments once arbitrarily and, to compare, also as fluxes of lower order equations.

Arbitrary Choice of Moments

Let arbitrarily $\mu_1 = E_1 r_1$, $\mu_2 = E_1 r_2$ and $\mu_3 = E_1 r_3$, where r_1, r_2 and r_3 are the first 3 basis vectors of the orthogonal complement to $\ker K$ (see comment after (3.87)). E_1 and G are chosen arbitrarily, as shown in App. A.4.2, fulfilling only the basic requirements of proper separation to E_0 and M, (3.26), but not necessarily the specific conditions for the scale induced closure, (3.33). For details of the construction of G and E_1, see App. A.4.2.

With E_1 and G, we use (3.27) and (3.28) and get the equations

$$\partial_t \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \left(\begin{array}{cccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 & -1 & 1 \\ 0 & 66/5 & 0 & 0 & 72/5 & 56/5 & -56/5 \\ \hline -30/32 & 0 & 0 & 3/32 & 9/4 & 3/4 & 3/4 \\ 5/16 & 0 & -1/20 & -1/32 & 2/5 & -6/5 & -1/5 \\ 5/8 & 0 & -1/10 & -1/16 & 4/5 & -2/5 & 3/5 \end{array}\right) \partial_x \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$
$$+\frac{1}{\varepsilon} \left(\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 4 & -7 & 7 \\ 0 & 0 & 0 & 0 & 0 & 11 & -3 \\ 0 & 0 & 0 & 0 & 2 & 1 & 7 \end{array}\right) \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.91}$$

Kinetic Fluxes as Moments

A more natural way to construct E_1 and G for the Grad equations is to consider those variables that appear in the fluxes of the equations, see also the remark in section 3.4.3. From the kinetic model, we obtain heat fluxes in x and y direction, as well as the pressure tensor

$$q_x = \sum_{i=1}^{16} c_i^{(1)} c_i^2 f_i, \quad q_y = \sum_{i=1}^{16} c_i^{(2)} c_i^2 f_i, \quad \sigma_{xy} = \sum_{i=1}^{16} c_i^{(1)} c_i^{(2)} f_i. \tag{3.92}$$

However, building E_1 and G from these vectors, we can compute that their equilibrium part is not zero, i.e. conditions (3.26) are not fullfilled in our model with 16 discrete velocities. The remedy is to chose the non-equilibrium projections

$$Pq_x, \quad Pq_y, \quad P\sigma_{12}. \tag{3.93}$$

For more details, see App. A.4.2. The resulting equations are:

$$\begin{aligned}
&\partial_t \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \left(\begin{array}{cccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 20 \\ 0 & 66/5 & 0 & 0 & 16\sqrt{34/5} & 0 & 0 \\ \hline -\sqrt{85/128} & 0 & 0 & \sqrt{1.7}/16 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2\sqrt{1.7} \\ 0 & 0 & 1/4 & 0 & 0 & 2\sqrt{1.7} & 0 \end{array}\right) \partial_x \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \\
&+ \frac{1}{\varepsilon} \left(\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 25/17 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 25/17 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 9/25 \end{array}\right) \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\end{aligned} \tag{3.94}$$

We will later show solutions to the Grad equations arbitrarily chosen non-equilibrium spaces $span\{r_1, r_2, r_3\}$, as well as with $span\{r_5, r_6, r_7\}$ and compare them to Grad solutions with kinetic fluxes as higher moments.

Scale Induced Closure

The operator $-K^\dagger \mathbf{c} \cdot M$ in (3.33) has a 3-dimensional image, its nullspace is $(1, 0, 0, 10)^T$. Therefore we get 3 higher moments μ_1, μ_2 and μ_3. We choose D as parametrization of the orthogonal complement of $(1, 0, 0, 10)^T$

$$D = \begin{pmatrix} \frac{10}{\sqrt{101}} & 0 & 0 & -\frac{1}{\sqrt{101}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad D^\dagger = \begin{pmatrix} \frac{10}{\sqrt{101}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\frac{1}{\sqrt{101}} & 0 & 0 \end{pmatrix}. \tag{3.95}$$

Constructing E_1 as the pseudoinverse $G^\dagger$ and plugging all into (3.50/3.51), we obtain the equations

$$\partial_t \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \left(\begin{array}{cccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & -\frac{4}{5} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{289}{20} \\ 0 & \frac{66}{5} & 0 & 0 & \frac{14\sqrt{101}}{5} & 0 & 0 \\ \hline -\frac{5600}{487\sqrt{101}} & 0 & 0 & \frac{560}{487\sqrt{101}} & 0 & \frac{608}{487\sqrt{101}} & 0 \\ 0 & -2 & 0 & 0 & \frac{19}{16\sqrt{101}} & 0 & 0 \\ 0 & 0 & -\frac{289}{841} & 0 & 0 & 0 & 0 \end{array}\right) \partial_x \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

$$+ \frac{1}{\varepsilon} \left(\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \frac{560}{487} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{289}{841} \end{array}\right) \begin{pmatrix} \rho \\ \rho v_x \\ \rho v_y \\ \rho e \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.96}$$

These equations are of second order accuracy since the conditions in Thm. 3.6.1 are met. Another way of checking the accuracy is through direct asymptotic expansion of (3.96). This furthermore shows that the equations are not of 3rd order (see App. A.4.3).

Comparison

In order to compare the results of the different closures, we look at the spatial Fourier transform $f_j(x,t) = \sum_{k\in\mathbb{Z}} e^{-ikx} \hat{f}_j^k(t)$. This transforms the gradients into factors of $-ik$. We apply the Fourier transform to (3.85), (3.88), (3.89), (3.91) and (3.96) and obtain ordinary differential equations with the solution

$$\partial_t \hat{f}_j^k(t) - i\sum_{j=1}^{16} V_{ij} k \hat{f}_j^k(t) + \frac{1}{\varepsilon}\sum_{j=1}^{16} K_{ij}\hat{f}_j^k(t) = 0, \qquad \hat{\mathbf{f}}^k(t) = \exp[ik\mathbf{V} - \frac{1}{\varepsilon}\mathbf{K}]\hat{\mathbf{f}}^k(0) \tag{3.97}$$

As initial condition we choose $\hat{f}_j^k(0) = 1$ for the wave number $k = 2\pi$ and for all $j = 1, ...16$, corresponding to Dirac peaks in the untransformed space. We show the results obtained with the various closures for the real part of the Fourier transformed mass density $\hat{\rho}^k = \sum_{j=1}^{16} \hat{f}_j^k(t)$ in Fig. 3.3.

For any ε, the Euler solution is oscillating without damping (no influence of the collision term). In Fig. 3.3, we see that the damping in Navier-Stokes dominates already after short time ($\varepsilon = 0.1$, $\varepsilon = 0.5$). Both, Euler and Navier-Stokes use 4 variables. For the Grad solution, we have different options of chosing the closure. We compare some of these choices in Fig. 3.4 (r_1, r_2, r_3; r_5, r_6, r_7; projected heat fluxes and pressure tensor). Clearly, the choice of heat flux and pressure tensor projected onto the non-equilibrium space (see Sect. 3.9.2), performs best for all ε.

The Grad solution, using 7 variables, shows damping, and there is a phase shift to the kinetic solution. The scale induced closure performs better with the same number of

variables for $\varepsilon = 0.01$ and $\varepsilon = 0.1$. We see hardly any difference between the scale induced closure and the kinetic equation for $\varepsilon = 0.01$, $\varepsilon = 0.1$, only at $\varepsilon = 0.5$ some deviations start to occur.

Summarizing, the choice of variables is the main point in all these methods - among the totally 16 kinetic variables, we want to choose a few linear combinations to build macroscopic variables. These should mimic the microscopic behaviour as good as possible. The scale-induced closure gives a hint how to choose the variables optimally in terms of the underlying kinetic structure.

Burnett equations (3.90) are unstable for large relaxation times $\varepsilon = 0.5$. This is due to the fact that the Burnett equations only consider an asymptotic expansion in ε, which does not necessarily become more accurate by adding higher order terms. In the scale induced closure, we are not only taking into account higher orders but additionally an enrichment of the approximation space.

To validate our results, we also show the imaginary part of the Fourier transformed velocity in x direction, $\hat{v}_x^k = \sum_{j=1}^{16} c_j^{(1)} \hat{f}_j^k(t)$ in Fig. 3.5. Density and x-velocity are non-trivial quantities in the model under consideration. The energy shows to be just a scaling of the density. Usually, higher moments are more difficult to capture, however in our case, the approximations for the velocity show the same qualities as for the density.

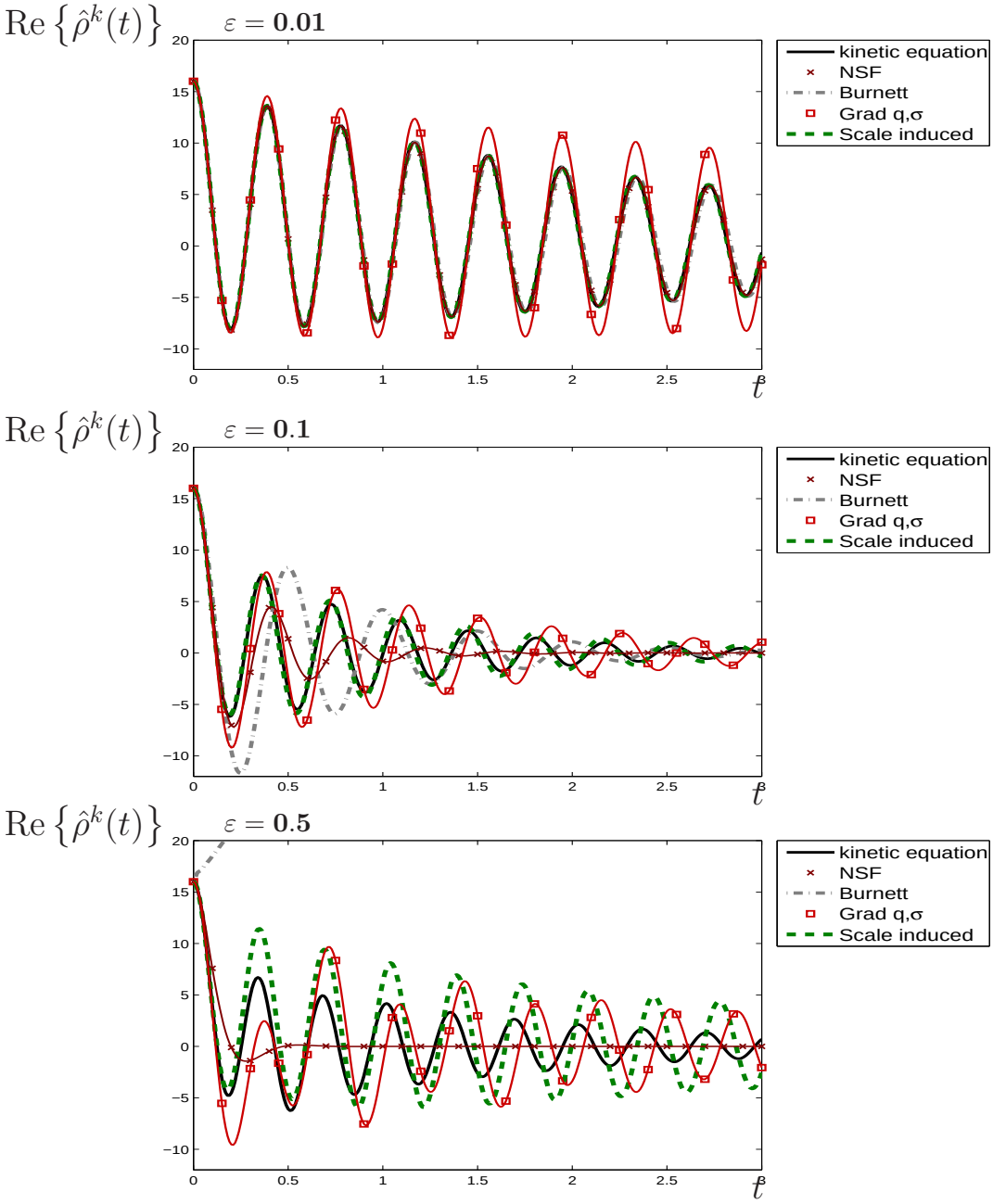


Figure 3.3: The different Closures for Fourier coefficient $k = 2\pi$ at $\varepsilon = 0.01$ (top), $\varepsilon = 0.1$ (middle) and $\varepsilon = 0.5$ (bottom).

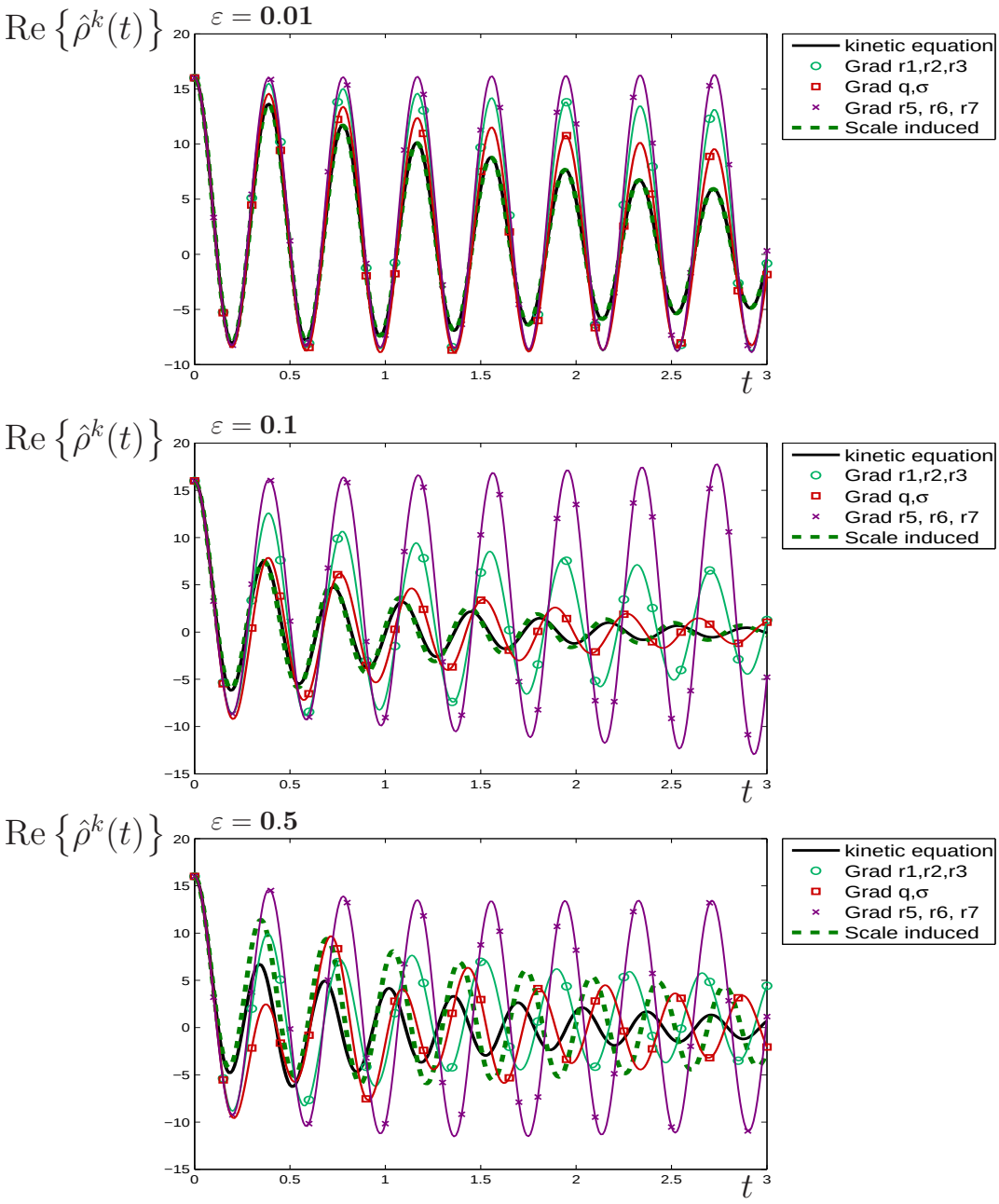


Figure 3.4: Various Grad closures for Fourier coefficient $k = 2\pi$ at $\varepsilon = 0.01$ (top), $\varepsilon = 0.1$ (middle) and $\varepsilon = 0.5$ (bottom).

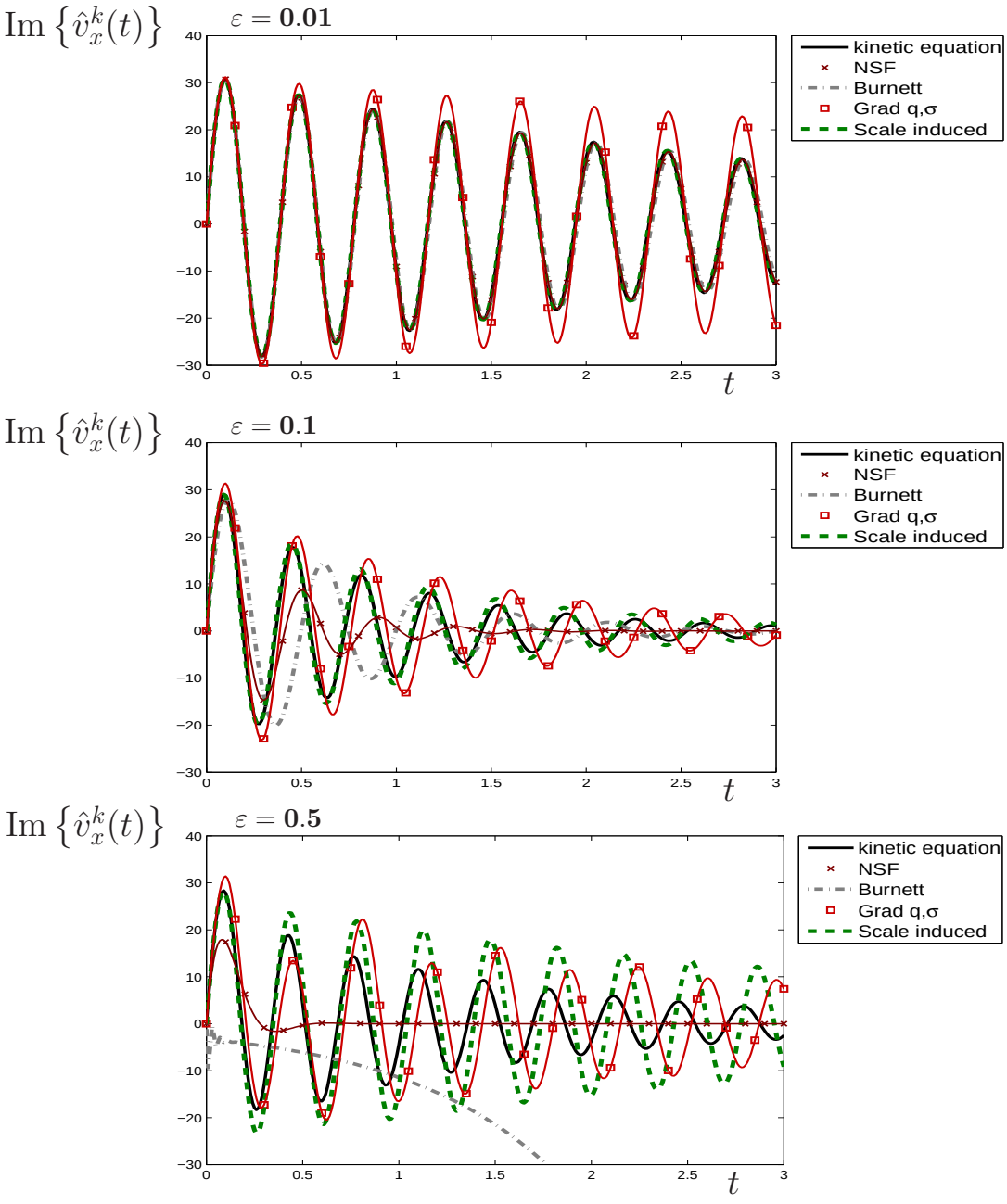


Figure 3.5: The different Closures for Fourier coefficient $k = 2\pi$ at $\varepsilon = 0.01$ (top), $\varepsilon = 0.1$ (middle) and $\varepsilon = 0.5$ (bottom).

3.9.3 Linear Matrix System

The second example is more abstract and illustrates the fundamental range of the new closure procedure. We consider a vector function $y : \mathbb{R}^+ \to \mathbb{R}^N$ satisfying an ordinary differential equation

$$\partial_t y + T\, y + \frac{1}{\varepsilon} K\, y = 0, \qquad y|_{t=0} = y^{(0)} \tag{3.98}$$

with initial conditions $y^{(0)}$. The matrix T generalizes the transport operator, while K can be viewed as collisional part. As for the kinetic model we assume that there exist vectors or matrices M and E_0 with $KM = 0$ and $E_0 K = 0$, as well as $E_0 M = id$. Equilibrium variables are given by $\rho = E_0\, y$. The whole theory derived above can be easily translated to the present case. The aim is to replace the high-dimensional system (3.98) by a lower dimensional system for ρ with high accuracy.

To check the approximation quality we consider a concrete example and take $N = 4$ and

$$K = \frac{1}{54} \begin{pmatrix} 45 & -3 & 21 & -21 \\ -3 & 65 & 31 & -31 \\ 21 & 31 & 53 & 1 \\ -21 & -31 & 1 & 53 \end{pmatrix} \tag{3.99}$$

as collision matrix. This matrix was constructed such that it exhibits the eigenvalues $\lambda_i \in \{0, 1, 1, 2\}$ and a one-dimensional kernel given by $M = (1, 1, -1, 1)^{tr}$ with $KM = 0$. In accordance with section 3.7, the equilibrium operator with $E_0 K = 0$ is given by $E_0 = (M^* M)^{-1} M^* = \frac{1}{4}(1, 1, -1, 1)$ and the equilibrium variable $\rho = E_0\, y$ is scalar. T is chosen to be

$$T = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{3.100}$$

We will solve the full system (3.98) with (3.99) and (3.100) numerically and compare the numerical results of various approximations like a Chapman-Enskog-type or the scale-induced closure to the full solution.

The kinetic variable ("distribution") satisfies $y = M\, \rho + y_1$ with a disturbance computed in (3.20)

$$y_1 = -\varepsilon\, K^\dagger T\, M\, \rho. \tag{3.101}$$

This leads to the equation (see (3.21))

$$\partial_t \rho + (E_0 T\, M)\, \rho - \varepsilon\, (E_0 T\, K^\dagger T\, M)\, \rho = 0 \tag{3.102}$$

in the sense of a first Chapman-Enskog expansion. Initial conditions are given by $\rho|_{t=0} = E_0 y^{(0)} = \rho^{(0)}$. For our example, it turns out that $E_0 T\, M = 0$ and $E_0 T\, K^\dagger T\, M = \frac{65}{54}$, so we find $\rho(t) = \rho^{(0)} \exp\left(-\varepsilon \frac{65}{54} t\right)$ as first approximation. According to the theory above,

a better approximation is given by equations for ρ coupled to a scalar higher moment $\mu = E_1 y$ with the structure (compare (3.27), (3.28))

$$\partial_t \begin{pmatrix} \rho \\ \mu \end{pmatrix} + \underbrace{\begin{pmatrix} E_0 T\, M & E_0 T\, G \\ E_1 T\, M & E_1 T\, G \end{pmatrix}}_{A} \begin{pmatrix} \rho \\ \mu \end{pmatrix} = -\frac{1}{\varepsilon} \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & E_1 K\, G \end{pmatrix}}_{B} \begin{pmatrix} \rho \\ \mu \end{pmatrix} \tag{3.103}$$

and particular choices for G and E_1.

Some of these choices are proposed through the Grad closure. As we have seen in the previous example, not every choice of G and E_1 just fulfilling (3.26) offers the same accuracy. Thus we first choose $E_1 = E_0 T$ and with it $G = (E_1 E_1^*)^{-1} E_1^*$, imitating the selection of higher moments from the kinetic equations. Luckily, conditions (3.26) are met, meaning that $E_0 T$ contains no equilibrium part.

Out of curiosity, we construct arbitrary vectors

$$G = \left(-\frac{1}{2}, 1, \frac{1}{14}, -\frac{3}{7}\right)^{tr}, \quad E_1 = \left(1, \frac{1}{2}, -1, -\frac{5}{2}\right) \tag{3.104}$$

satisfying the basic requirements $E_0 G = 0$, $E_1 M = 0$ and $E_1 G = 1$, for comparison. In Grad's approach, independent of the choice of non-equilibrium moments, the kinetic structure given through K is not fully exploit. Instead, the new scale-induced order-of-magnitude approach ($D = 1$) suggests to use

$$G = -K^\dagger T M D^\dagger, \quad E_1 = (G^* G)^{-1} G^* \tag{3.105}$$

which is adapted to the structure of the kinetic equation.

In Fig. 3.6 we compare the evolution of ρ as predicted from the full system (3.98), from the Chapman-Enskog-type result, the two Grad approaches and the order-of-magnitude equations.

The relaxation times are chosen to be $\varepsilon = 0.01$, $\varepsilon = 0.1$ and $\varepsilon = 0.5$. The CE result manages to predict a general decaying behaviour, while the random moment Grad approximation gives an initial behaviour that is qualitatively correct but fails for large times t. The Grad approximation with $E_1 = T E_0$ performs much better, however also fails in the low $\varepsilon = 0.01$ case, compared to the scale induced closure. The scale induced closure result matches the full solution in a nearly perfect way for $\varepsilon = 0.01$ and $\varepsilon = 0.1$. For $\varepsilon = 0.5$, the Grad method becomes slightly better, which is a surprising coincidence, but not more than that: the behaviour of Grad is difficult to predict for different settings of parameters and problems. The scale-induced closure behaves more steadily in its approximation quality.

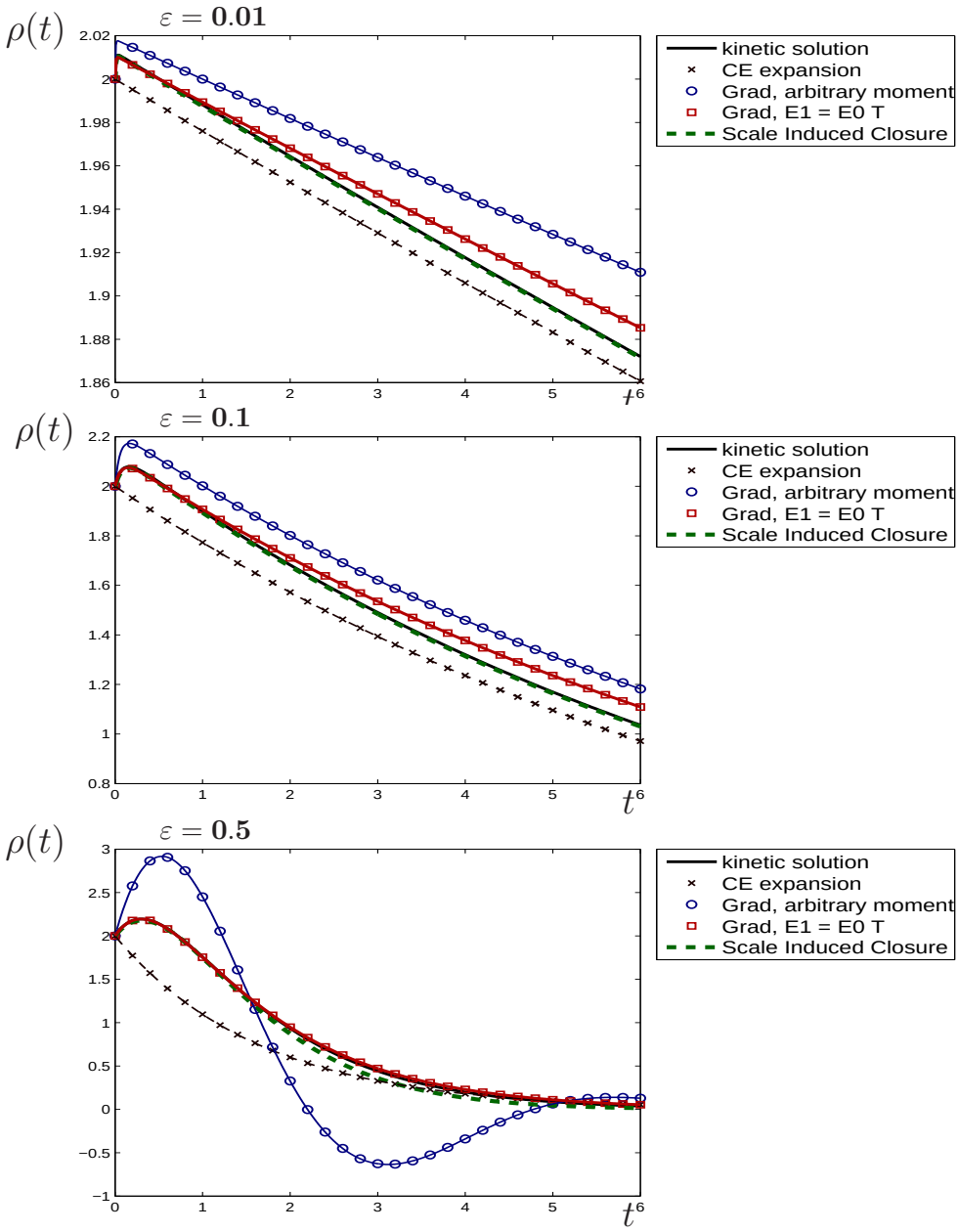


Figure 3.6: Solution of the full matrix system (3.98) and various lower dimensional approximations at $\varepsilon = 0.01$ (top), $\varepsilon = 0.1$ (middle) and $\varepsilon = 0.5$ (bottom).

We do not want to overstress this rather special example, but it indicates that the scale-induced closure may considerably improve the accuracy of lower dimensional approximations of more general equations.

For completeness we give the resulting matrices in the system (3.103) for the random moment Grad approach

$$A = \begin{pmatrix} 0 & -\frac{41}{28} \\ \frac{5}{4} & -\frac{81}{28} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 0 & \frac{55}{27} \end{pmatrix}, \tag{3.106}$$

the Grad approach with $E_1 = TE_0$

$$A = \begin{pmatrix} 0 & 1 \\ -\frac{3}{2} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 0 & \frac{113}{81} \end{pmatrix}. \tag{3.107}$$

and the scale induced closure

$$A = \begin{pmatrix} 0 & \frac{65}{57} \\ -\frac{65}{54} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 0 & \frac{65}{57} \end{pmatrix}. \tag{3.108}$$

As initial condition for the full system $y^{(0)} = (1, 4, -2, 1)$ was used which corresponds to $\rho^{(0)} = 2$.

3.10 Conclusion

This work supplements the work of Struchtrup in [49] and [50] where an order-of-magnitude closure for moment equations in kinetic gas theory was developed. Here, we generalize this approach to the level of kinetic equations and relate it to standard methods of Chapman-Enskog and Grad. The new closure obeys a scaling of the non-equilibrium phase space that is introduced by asymptotic expansion. This scaling structures the phase space and allows to formulate a distribution function based on moments respecting the asymptotic properties of the kinetic equation. In this sense, it provides a scale-induced closure. The resulting moment equations exhibit high asymptotic accuracy in a natural way.

The theory is developed in the case of a linear kinetic model equation. The final equations can be shown to possess an entropy law and to be L^2-stable. In future work the results need to be extended to the non-linear case. This should be possible since the original method was conducted on non-linear moment equations, however the necessary mathematical tools in the non-linear setting will be more sophisticated. Our example with a linearized discrete velocity model showed that the scale induced closure is performing very well in approximating the high-dimensional kinetic evolution by low dimensional equations, giving good reasons to also use it in more general settings.

4 Multi-Scale Modeling for the non-linear Boltzmann Equation

In this part, we are presenting a hybrid computational approach to the Boltzmann-BGK equation. We are modelling a physical domain of medium Knudsen number, where (near) equilibrium methods as Navier-Stokes-Fourier or Euler do not capture the particular effects and DSMC methods are too costly to be applied.

We will first derive a Galilei invariant, sound speed scaled formulation of the Boltzmann equation. Then, the Boltzmann-BGK distribution function will be non-linearly approximated through a multi-scale ansatz: we will use the equilibrium Gaussian and multiply it with a series of higher order non-linear fluctuations. These fluctuations are represented in terms of perturbation functions that have to be chosen a priori and corresponding coefficients that follow from a quasi-linear PDE system. We will discuss an efficient Galilei-invariant and scaled formulation of this PDE system and consider issues of microscopic-macroscopic compatibility. Conservation of macroscopic fields will be ensured through a coupling of the PDE system to the macroscopic balance laws. To validate these equations, we will derive Grad's five moment equations, which will also serve as test case of the numerical scheme designed to our PDE system. We will compare various choices of perturbation functions in pure approximation of a given distribution function, and then validate the PDE solutions resulting from these choices by comparison to a fine scale discrete velocity Boltzmann-BGK solver.

4.1 Key Ideas

The Boltzmann equation poses challenges for numerical solutions because in addition to time and space also the velocity variables need to be discretized. In moment equations like Grad or in the equations of Euler or Navier-Stokes-Fourier, the velocity space is replaced by a finite set of variables. The resulting equations can be viewed as an approximation to the phase space density $f(x, t, c)$ that solves the Boltzmann equation. Despite of the much smaller number of parameters these approximations are quite accurate. This is due to scale separation and built-in physical asymptotics, see [30] and Part 3.

As is well known, in rarefied gases scale separation breaks down and we need to extend the number of moments to still capture the relevant physics (see [57]). However, the

extension to higher order moments is rather complicated, as can be verified in the textbook [51]. There are successfull approaches that build upon moment equations and use a regularization procedure to get higher approximation order and stability (see [49], [50], [52],[53] and Part 3). Another promising approach that is related to moment equations is based on a parametrization with a distribution from the Pearson family instead of using the equilibrium Gaussian, see [58]. This approach exhibits highly non-linear relations between parameters and distribution approximation.

Discrete velocity schemes, as opposed to moment equations, are computationally expensive discretizations of the velocity space. If the production term is reasonably simple (e.g. in the BGK model), the accuracy of such schemes can be easily increased by adding more discretization points to the velocity grid. In the setting of moment equations it is not clear how to increase the accuracy in a simple and predictable way (see Part 3). In more than one dimension, an accurate pointwise velocity space discretization becomes very expensive. In the kinetic regime, where the Knudsen number Kn is large, this can be overcome by the 'Direct Simulation Monte Carlo' method (DSMC), as long as the Mach number is not too low (see [6]). For lower Knudsen numbers or low Mach numbers, the DSMC method fails and produces inaccurate results.

Another approach to make discrete-velocity-like schemes more feasible in higher dimensions is described in [3]. Relatively few velocity gridpoints are used, but they are distributed in a very sophisticated way such that collisions can be simulated optimally. This method is very successfull if we are interested in qualitative features, but lacks quantitative accuracy. Further development on the design of discrete velocity schemes can also be found in [38].

Approaches which allow for a hybrid treatment of the computational domain have been developed in e.g. [15] or [31]. Such methods generally rely on splitting the spatial domain into regions of various Knudsen numbers. Depending on the size of Kn, an appropriate model (Euler, DSMC, BGK, etc.) is used, based on some switching rules. Hybrid approaches clearly increase the computational efficiency and can describe different physics in the spatial domain accurately. They furthermore overcome the problems of breakdowns of discrete velocity schemes for low Knudsen numbers (see e.g. [34]). The problem of finding 'good' switching criteria between the different regions is however challenging.

Our approach focuses conceptionally onto the closure problem of the balance laws of mass, momentum and energy. In many approaches (e.g. discrete velocity schemes), the conservation of these macroscopic fields poses problems. We are closing the balance laws by computing anisotropic pressure and heat flux as moments of a model phase density $f(x, t, c)$, that approximates the Boltzmann-BGK density function.

Our model decomposes f in a setting of one space and one velocitiy dimension into an equilibrium part and into higher order perturbations. In order to do this effectively, we use a physically adaptive grid for the velocity space and then separate the equilibrium contribution from the perturbation part. We specify:

Physical adaptivity: The shape of the distribution function f varies in space and time, which is why a straight forward discretization of the velocity space will either have much redundancy or will pose tedious challenges for adaptivity in every space time point. We are using a scaled Galilei transform of the velocities

$$c \quad \mapsto \quad \xi = \frac{c - v(x,t)}{\sqrt{\theta(x,t)}}$$

with mean velocity v and temperature θ.[1] Thus we rescale f and obtain a quasi Lagrangian, moving discretization of the velocity space, where the grid automatically focusses on the support of f and therefore captures the relevant physics.

Decomposition into basis functions: Simple discretizations of f into point values do not use any information about the shape of f. However, we have quite some information about this shape, at least close to equilibrium, where f is (almost) a Maxwellian. Making use of this information, we decompose the distribution function more effectively into equilibrium and general perturbation factors,

$$\hat{\mathrm{f}}(x,t,\xi) = \underbrace{\frac{\rho(x,t)}{\sqrt{2\pi\theta(x,t)}} \exp\left(-\frac{(c - v(x,t))^2}{2\theta(x,t)}\right)}_{\text{Maxwellian } F_M} \left(1 + \sum_{\alpha=1}^{N} \kappa_\alpha(x,t)\phi_\alpha(\xi)\right).$$

Observe that this decomposition gives way to a transition between discrete velocity models (choosing Dirac functions for ϕ_α) and moment methods (choosing polynomials for ϕ_α). Typically, this ansatz is capturing the shape of f in a more appropriate and efficient way than either of the two. This shows well in the case of bimodal distribution functions, where classical moment equations are known to be inaccurate or in the case of Euler equations, where the one dimensional situation can be modelled with 3 parameters instead of hundreds of point values.

In this context, our approach yields an intrinsically hybrid scheme for the Boltzmann equation: it adapts automatically to regions with high non-equilibrium by increasing the sizes of the perturbation factors, and sets these to (almost) zero in regions with (almost) equilibrium. Like this, we can avoid the problem of finding appropriate switches, but of course we do not gain computational efficiency, since we are solving equations for all of the perturbation factors, even if they are close to zero. Unfortunately, asymptotic properties of our approach are very difficult to obtain, so it is not clear so far, in which sense the limiting scheme would indeed be a numerical approximation of the discretized Euler equations.

Summarizing, we are reducing the computational complexity from a point value discretization to a few parameters for the functions ϕ_α. The resulting PDE system for

[1] These correspond to mean and variance of f.

κ_α, we call it 'constitutive equations'[2], depends on the macroscopic density, velocity and temperature, and is thus coupled to the corresponding balance laws. Its structure is partially determined by properties of the perturbation functions ϕ_α. The choices of these perturbation functions are almost unlimited (monomials, piecewise polynomials, general wavelets), which poses some challenges in appropriate non-linear modelling.

Ideally, we can combine the advantages and accuracy of several models while avoiding the major drawbacks from either of them.

This part is organized as follows: In Sect. 4.2 we will derive the PDE system for the perturbation function coefficients κ_α in one space and one velocity dimension. There we will also consider some mathematical features of this system and will take care about compatibility between the microscopic quantities κ_α and the macroscopic fields ρ, v and θ. We will then compare various perturbation functions in Sect. 4.3. Their approximation features are examplified in a test situation of a strongly bimodal distribution function, typically occuring in shock problems. In Sect. 4.4 we will present a numerical method to solve the PDE system with coupling to the balance laws. Sect. 4.5 is dedicated to the performance analysis of our numerical scheme in the setting of Grad's 5 moment system, which is contained as a special case of our PDE system. In Sect. 4.6, we will use our approach with various choices of perturbation functions ϕ_α and compare the results to those obtained through a fine scale discrete velocity BGK-solver. We will conclude in Sect. 4.7 with an overview of what goals further research could aim at.

4.2 The Constitutive Equations

In this section, we are deriving the PDE system for the perturbation coefficients κ_α. We call this system 'constitutive equations' since it closes the balance laws with a specific model, taking into account properties of the material that is described. First we will transform the Boltzmann-BGK equations into a Galilei-invariant form (Sect. 4.2.1). In Sect. 4.2.2 we will formulate our equilibrium / non-equlibrium ansatz for the distribution function and derive (necessary and) sufficient conditions for its compatibility with the macroscopic fields. Then, we will cast the resulting equations for κ_α into weak form (Sect. 4.2.3) and couple them to the balance laws in Sect. 4.2.4.

4.2.1 Galilei-Invariant Boltzmann Equation

The Boltzmann-BGK equation (compare Part 2) reads

$$\partial_t f + c\,\partial_x f = S(f) \overset{BGK}{=} \frac{1}{\tau}\left(\frac{\rho}{\sqrt{2\pi\theta}}\exp\left(-\frac{(c-v)^2}{2\theta}\right) - f\right). \tag{4.1}$$

[2] 'Constitutive' since these equations form a closure of the balance laws by modeling the heat flux, see Sect. 4.2.4 for more details.

As we see on the left in Fig. 4.1, the domain of dependence of f on the velocity c can vary extensively in space and time, even in equilibrium (different temperatures and mean velocities). This source of inefficiency for the numerical discretization can be overcome with appropriate scaling. While (4.1) in its analytical form is Galilei-invariant, any direct discretization with a finite velocity grid is not[3]. Therefore, we consider a Galilei-invariant and sound speed scaled formulation of (4.1), which gives the proper physically adaptive scaling for the discrete case:

$$c \quad \overset{\text{Galilei-invariance}}{\longrightarrow} \quad c - v(x,t) \quad \overset{\text{scaling}}{\longrightarrow} \quad \frac{c - v(x,t)}{\sqrt{\theta(x,t)}} := \xi(x,t,c). \tag{4.2}$$

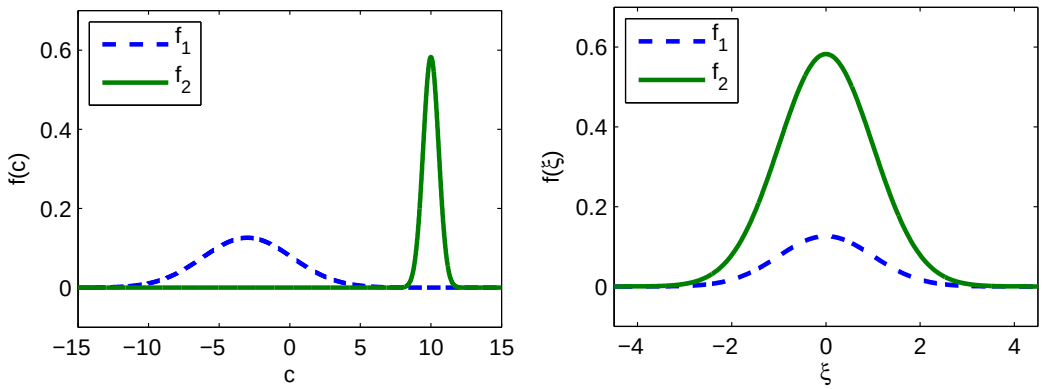


Figure 4.1: Two Gaussians, f_1 with $\rho_1 = 1$, $v_1 = -3$, $\theta_1 = 10$ and f_2 with $\rho_2 = 0.8$, $v_2 = 10$, $\theta_2 = 0.3$, in dependence of c (left), and ξ (right).

The macroscopic quantities entering here are

$$\rho = m \int f dc, \quad v = m\frac{1}{\rho} \int c f dc, \quad \theta = m\frac{1}{\rho} \int (c-v)^2 f dc. \tag{4.3}$$

We denote $\hat{f}(x,t,\xi(x,t,c)) := f(x,t,c)$. Derivatives and integrals in (4.1) expand to

$$\partial_{\mathrm{c}} f(c) = \frac{1}{\sqrt{\theta}} \partial_\xi \hat{f}(\xi), \qquad dc = \sqrt{\theta(x,t)} d\xi \tag{4.4a}$$

$$\partial_t f = \partial_t \hat{f} + \left(-\frac{\partial_t v}{\sqrt{\theta}} - \frac{1}{2} \frac{c-v}{\sqrt{\theta^3}} \partial_t \theta \right) \partial_\xi \hat{f} \tag{4.4b}$$

$$\partial_{\mathrm{x}} f = \partial_{\mathrm{x}} \hat{f} + \left(-\frac{\partial_{\mathrm{x}} v}{\sqrt{\theta}} - \frac{1}{2} \frac{c-v}{\sqrt{\theta^3}} \partial_{\mathrm{x}} \theta \right) \partial_\xi \hat{f}. \tag{4.4c}$$

[3] In a finite grid, there are minimal and maximal velocities C_{min} and C_{max}. Shifting the equation by some speed v would require a grid starting at $C_{min} + v$ and ending at $C_{max} + v$, so the grid formulation is not Galilei invariant.

With the convective time derivative $D_t := \partial_t + v\partial_x$, the Boltzmann-BGK equation (4.1) turns into

$$\begin{aligned} D_t \hat{\mathrm{f}} + \sqrt{\theta}\xi\, \partial_x \hat{\mathrm{f}} + \partial_\xi \hat{\mathrm{f}} \left\{ -\frac{1}{\sqrt{\theta}} \left(D_t v + \sqrt{\theta}\xi\, \partial_x v \right) - \frac{1}{2\theta}\xi \left(D_t\theta + \sqrt{\theta}\xi\, \partial_x \theta \right) \right\} \\ = S(\hat{\mathrm{f}}, \rho, v, \theta) \overset{BGK}{=} \frac{1}{\tau} \left(\frac{\rho}{\sqrt{2\pi\theta}} \exp\left(-\xi^2/2\right) - \hat{\mathrm{f}} \right). \end{aligned} \tag{4.5}$$

The additional terms describe 'self-forcing' to the invariant formulation. These terms are the price to pay for physical adaptivity.

In the untransformed case, f contains all information about ρ, v, θ or higher moments. From the Galilei tranformed and scaled $\hat{\mathrm{f}}$ alone we are not able to reconstruct f, since information on v and θ is missing[4]. This shows by applying the transformation to the moment definitions (4.3),

$$\rho = \sqrt{\theta} m \int \hat{\mathrm{f}}\, d\xi \tag{4.6a}$$

$$\rho v = m\sqrt{\theta} \int (\sqrt{\theta}\xi + v)\, \hat{\mathrm{f}}\, d\xi \tag{4.6b}$$

$$\rho\theta = m\sqrt{\theta}^3 \int \xi^2\, \hat{\mathrm{f}}\, d\xi. \tag{4.6c}$$

Observe here that (4.5) is not a closed system since it contains two more variables than equations[5]. We will use and interpret (4.6) as *compatibility conditions* on the distribution function $\hat{\mathrm{f}}$ and solve the problem of underdetermination in (4.5) through a coupling to the balance laws for ρ, v and θ as derived from the untransformed Boltzmann equation (4.1).

4.2.2 Ansatz and Compatibility Conditions

Knowing the equilibrium Maxwellian,

$$\hat{F}_M(x,t,\xi) = \frac{\rho(x,t)}{\sqrt{2\pi\theta(x,t)}} \exp\left(-\xi^2/2\right), \tag{4.7}$$

we expand our distribution function $\hat{\mathrm{f}}$ into an equilibrium prefactor and a finite series of perturbation factors[6],

$$\hat{\mathrm{f}} = \frac{\rho}{\sqrt{2\pi\theta}} \exp\left(-\xi^2/2\right) \left\{ 1 + \sum_{\alpha=1}^{N} \kappa_\alpha(x,t)\phi_\alpha(\xi) \right\}. \tag{4.8}$$

[4] This is why the collision operator $S(\hat{\mathrm{f}}, \rho, v, \theta)$ depends not only on $\hat{\mathrm{f}}$, but also directly on ρ, v and θ.
[5] ρ can be expressed in terms of θ through (4.6).
[6] tThe exact mathematical meaning of this formal ansatz will be clarified in Sect. 4.2.3.

The compatibility conditions (4.6) for $\hat{\mathrm{f}}$ translate into conditions on κ and ϕ as

$$\sum_{\alpha=1}^{N} \kappa_\alpha(x,t) \int_{-\infty}^{\infty} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} \exp\left(-\xi^2/2\right) \phi_\alpha(\xi) d\xi = 0. \tag{4.9}$$

The compatibility could also be imposed onto the perturbation functions by asking

$$\int_{-\infty}^{\infty} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} \exp\left(-\xi^2/2\right) \phi_\alpha(\xi) d\xi = 0, \tag{4.10}$$

which is a stronger condition than (4.9), we will see more on this in Sect. 4.3.2.

Plugging the Ansatz (4.8) into (4.5), we obtain equations for κ_α (use the Einstein sum convention, see App. A.1),

$$\begin{aligned} &\left(\frac{Dt\rho}{\rho} - \frac{D_t\theta}{2\theta}\right)(1+\kappa_\alpha\phi_\alpha) + \phi_\alpha D_t\kappa_\alpha + \sqrt{\theta}\xi\left\{\left(\frac{\partial_x\rho}{\rho} - \frac{\partial_x\theta}{2\theta}\right)(1+\kappa_\alpha\phi_\alpha) + \phi_\alpha\,\partial_x\kappa_\alpha\right\} \\ &\quad + \left\{-\xi\,(1+\kappa_\alpha\phi_\alpha) + \kappa_\alpha\,\partial_\xi\,\phi_\alpha\right\}\left\{-\frac{1}{\sqrt{\theta}}\left(D_t v + \sqrt{\theta}\xi\,\partial_x v\right) - \frac{1}{2\theta}\xi\left(D_t\theta + \sqrt{\theta}\xi\,\partial_x\theta\right)\right\} \\ &\qquad = \tilde{S}(\kappa;\rho,v,\theta) \overset{BGK}{=} -\frac{1}{\tau}\left(\kappa_\alpha\phi_\alpha\right), \end{aligned} \tag{4.11}$$

with

$$\tilde{S}(\kappa,\rho,v,\theta) = \frac{\sqrt{2\pi\theta}}{\rho}\exp\left(\xi^2/2\right) S(\kappa,\rho,v,\theta)$$

4.2.3 Weak Formulation

The strong formulation for the Galilei-invariant and sound speed scaled Boltzmann-BGK equation, (4.11), still contains the velocity variable ξ. With a weak formulation, a direct discretization of ξ will not be necessary, all the information about the velocity space can be encoded in ϕ_α and the equilibrium Maxwellian.

First, the decomposition (4.8) is to be understood in a weak L^2-sense,

$$< \phi_\beta, \hat{\mathrm{f}} >_{L^2(\xi)} = < \phi_\beta, \frac{\rho(x,t)}{\sqrt{2\pi\theta(x,t)}}\exp\left(-\xi^2/2\right)\left\{1 + \sum_{\alpha=1}^{N}\kappa_\alpha(x,t)\phi_\alpha(\xi)\right\} >_{L^2(\xi)}, \tag{4.12}$$

with $< f,g >_{L^2(\xi)} := \int_{-\infty}^{\infty} f(\xi)g(\xi)d\xi$. (4.12) corresponds to a linear system of equations for κ_α with some given data $< \phi_\beta, \hat{\mathrm{f}} >_{L^2(\xi)}$.

In order to cast (4.11) into a weak form, we choose test functions ψ_β, $\beta = 1, ..., N$ and form the scalar product $\langle \psi_\beta, (4.11) \rangle_{L^2(\xi)}$, for $\beta = 1, ..., N$. This yields (use again the Einstein sum convention)

$$
\begin{aligned}
&\left(\frac{D_t \rho}{\rho} - \frac{D_t \theta}{2\theta} \right) \left(\langle \psi_\beta, 1 \rangle + \kappa_\alpha \langle \psi_\beta, \phi_\alpha \rangle \right) + D_t \kappa_\alpha \langle \psi_\beta, \phi_\alpha \rangle \\
&+\sqrt{\theta} \left\{ \left(\frac{\partial_x \rho}{\rho} - \frac{\partial_x \theta}{2\theta} \right) \left(\langle \psi_\beta, \xi \rangle + \kappa_\alpha \langle \psi_\beta, \xi \phi_\alpha \rangle \right) + \partial_x \kappa_\alpha \langle \psi_\beta, \xi \phi_\alpha \rangle \right\} \\
&+\frac{1}{\sqrt{\theta}} \left[D_t v \langle \psi_\beta, \xi \rangle + \sqrt{\theta}\, \partial_x v \langle \psi_\beta, \xi^2 \rangle \right] + \frac{1}{2\theta} \left[D_t \theta \langle \psi_\beta, \xi^2 \rangle + \sqrt{\theta}\, \partial_x \theta \langle \psi_\beta, \xi^3 \rangle \right] \\
&+\kappa_\alpha \frac{1}{\sqrt{\theta}} \left[D_t v \langle \psi_\beta, \xi \phi_\alpha \rangle + \sqrt{\theta}\, \partial_x v \langle \psi_\beta, \xi^2 \phi_\alpha \rangle \right] \\
&+\kappa_\alpha \frac{1}{2\theta} \left[D_t \theta \langle \psi_\beta, \xi^2 \phi_\alpha \rangle + \sqrt{\theta}\, \partial_x \theta \langle \psi_\beta, \xi^3 \phi_\alpha \rangle \right] \\
&-\kappa_\alpha \frac{1}{\sqrt{\theta}} \left[D_t v \langle \psi_\beta, \partial_\xi \phi_\alpha \rangle + \sqrt{\theta}\, \partial_x v \langle \psi_\beta, \xi\, \partial_\xi \phi_\alpha \rangle \right] \\
&-\kappa_\alpha \frac{1}{2\theta} \left[D_t \theta + \langle \psi_\beta, \xi\, \partial_\xi \phi_\alpha \rangle + \sqrt{\theta}\, \partial_x \theta \langle \psi_\beta, \xi^2\, \partial_\xi \phi_\alpha \rangle \right] \\
&= \langle \psi_\beta, \tilde{S}(\kappa, \rho, v, \theta) \rangle \overset{BGK}{=} -\frac{1}{\tau} \kappa_\alpha \langle \psi_\beta, \phi_\alpha \rangle
\end{aligned}
\tag{4.13}
$$

For simplifications of (4.13), let us define

$$
\begin{aligned}
&M_{\mu\nu} := \langle \psi_\mu, \phi_\nu \rangle \\
&M^1_{\mu\nu} := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi \phi_\nu \rangle, \quad M^2_{\mu\nu} := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^2 \phi_\nu \rangle, \\
&M^3_{\mu\nu} := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^3 \phi_\nu \rangle \\
&D^0_{\mu\nu} := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \partial_\xi \phi_\nu \rangle, \quad D^1_{\mu\nu} := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi\, \partial_\xi \phi_\nu \rangle, \\
&D^2_{\mu\nu} := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^2\, \partial_\xi \phi_\nu \rangle \\
&V^0_\mu := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, 1 \rangle, \quad V^1_\nu := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi \rangle, \\
&V^2_\mu := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^2 \rangle, \quad V^3_\mu := (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^3 \rangle \\
&Q_\mu := \frac{1}{\sqrt{2\pi}} \int \xi^3 e^{-\xi^2/2} \phi_\mu d\xi \qquad \text{(heat flux, used in (4.17))}
\end{aligned}
\tag{4.14}
$$

We rename the summation index α in (4.13) to γ, multiply by $(M^{-1})_{\alpha\beta}$, and use the abbreviations (4.14) to obtain

$$
\begin{aligned}
&\left(\frac{D_t\rho}{\rho}-\frac{D_t\theta}{2\theta}\right)\left(V_\alpha^0+\kappa_\alpha\right)+D_t\kappa_\alpha\\
&+\sqrt{\theta}\left\{\left(\frac{\partial_x\rho}{\rho}-\frac{\partial_x\theta}{2\theta}\right)\left(V_\alpha^1+\kappa_\gamma M_{\alpha\gamma}^1\right)+\partial_x\,\kappa_\gamma M_{\alpha\gamma}^1\right\}\\
&+\frac{1}{\sqrt{\theta}}\left[D_t v V_\alpha^1+\sqrt{\theta}\,\partial_x\,v V_\alpha^2\right]+\frac{1}{2\theta}\left[D_t\theta V_\alpha^2+\sqrt{\theta}\,\partial_x\,\theta V_\alpha^3\right]\\
&+\kappa_\gamma\frac{1}{\sqrt{\theta}}\left[D_t v M_{\alpha\gamma}^1+\sqrt{\theta}\,\partial_x\,v M_{\alpha\gamma}^2\right]+\kappa_\gamma\frac{1}{2\theta}\left[D_t\theta M_{\alpha\gamma}^2+\sqrt{\theta}\,\partial_x\,\theta M_{\alpha\gamma}^3\right]\\
&-\kappa_\gamma\frac{1}{\sqrt{\theta}}\left[D_t v D_{\alpha\gamma}^0+\sqrt{\theta}\,\partial_x\,v D_{\alpha\gamma}^1\right]-\kappa_\gamma\frac{1}{2\theta}\left[D_t\theta D_{\alpha\gamma}^1+\sqrt{\theta}\,\partial_x\,\theta D_{\alpha\gamma}^2\right]\\
&=(M^{-1})_{\alpha\beta}\langle\psi_\beta,\tilde{S}(\kappa,\rho,v,\theta)\rangle\overset{BGK}{=}-\frac{1}{\tau}\kappa_\alpha.
\end{aligned}
\tag{4.15}
$$

We replace the D_t derivatives of ρ, v and θ in (4.15) with the help of the balance laws,[7]

$$
D_t\begin{pmatrix}\rho\\ v\\ \theta\end{pmatrix}=\begin{pmatrix}-\rho\,\partial_x\,v\\ -\frac{\theta}{\rho}\,\partial_x\,\rho-\partial_x\,\theta\\ -2\theta\,\partial_x\,v-\frac{1}{\rho}\,\partial_x\,q\end{pmatrix},
\tag{4.16}
$$

where the heat flux q is (compare Sect. 2.2.5)

$$
\begin{aligned}
q&=\int(c-v)^3 f\,dc=\int\sqrt{\theta}^3\xi^3\,\hat{\mathrm{f}}\,\sqrt{\theta}d\xi=\theta^2\frac{\rho}{\sqrt{2\pi\theta}}\left(1\cdot\int\xi^3 e^{-\xi^2/2}d\xi+\kappa_\gamma\int\xi^3 e^{-\xi^2/2}\phi_\gamma d\xi\right)\\
&=\theta^{3/2}\frac{\rho}{\sqrt{2\pi}}\kappa_\gamma\int\xi^3 e^{-\xi^2/2}\phi_\gamma d\xi,
\end{aligned}
\tag{4.17}
$$

and its derivative[8]

$$
\partial_x\,q=Q_\gamma\left(\rho\theta^{3/2}\,\partial_x\,\kappa_\gamma+\theta^{3/2}\kappa_\gamma\,\partial_x\,\rho+\rho\kappa_\gamma\frac{3}{2}\sqrt{\theta}\,\partial_x\,\theta\right).
\tag{4.18}
$$

Thus

$$
D_t\begin{pmatrix}\rho\\ v\\ \theta\end{pmatrix}=\begin{pmatrix}-\rho\,\partial_x\,v\\ -\frac{\theta}{\rho}\,\partial_x\,\rho-\partial_x\,\theta\\ -\frac{\theta^{3/2}}{\rho}Q_\gamma\kappa_\gamma\,\partial_x\,\rho-\frac{3}{2}\sqrt{\theta}Q_\gamma\kappa_\gamma\,\partial_x\,\theta-2\theta\,\partial_x\,v-\theta^{3/2}Q_\gamma\,\partial_x\,\kappa_\gamma\end{pmatrix}.
\tag{4.19}
$$

[7] The conservation laws in one space dimension can be derived analogously to the multi dimensional recipe given in Sect. 2.2.5, $\sigma=\rho\theta$.

[8] Q_γ is defined in (4.14).

With this substitution and some simplifications, (4.15) reformulates to

$$\begin{aligned}
\partial_t \kappa_\alpha + \left(v\delta_{\alpha\gamma} + \sqrt{\theta} M^1_{\alpha\gamma} - \frac{\sqrt{\theta}}{2} \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\mu M^2_{\alpha\mu} - \kappa_\mu D^1_{\alpha\mu} \right) Q_\gamma \right) \partial_x \kappa_\gamma \\
+ \partial_x \rho \left\{ \kappa_\gamma D^0_{\alpha\gamma} - \frac{1}{2} Q_\gamma \kappa_\gamma \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\gamma M^2_{\alpha\gamma} - \kappa_\gamma D^1_{\alpha\gamma} \right) \right\} \frac{\sqrt{\theta}}{\rho} \\
+ \partial_x \theta \left\{ \frac{V^3_\alpha}{2} + \kappa_\gamma \frac{M^3_{\alpha\gamma}}{2} - \kappa_\gamma \frac{D^2_{\alpha\gamma}}{2} - \frac{3}{2} V^1_\alpha - \frac{3}{2} \kappa_\gamma M^1_{\alpha\gamma} + \kappa_\gamma D^0_{\alpha\gamma} \right. \\
\left. - \frac{3}{4} Q_\gamma \kappa_\gamma \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\gamma M^2_{\alpha\gamma} - \kappa_\gamma D^1_{\alpha\gamma} \right) \right\} \frac{1}{\sqrt{\theta}} \\
= (M^{-1})_{\alpha\beta} \langle \psi_\beta, \tilde{S}(\kappa, \rho, v, \theta) \rangle \overset{BGK}{=} -\frac{1}{\tau} \kappa_\alpha.
\end{aligned} \tag{4.20}$$

Observe that (4.20) does not depend on $\partial_x v$ anymore.

In order to densify notation and make the structure of (4.20) more clearly visible, we introduce

$$W := (\rho, v, \theta)^T \in \mathbb{R}^3, \qquad \kappa = (\kappa_1, ..., \kappa_N)^T \in \mathbb{R}^N \tag{4.21}$$

and

$$B(W, \kappa) \in \mathbb{R}^{N \times N}, \qquad C(W, \kappa) \in \mathbb{R}^{N \times 3} \tag{4.22}$$

as

$$\begin{aligned}
B_{\alpha\gamma}(W, \kappa) &:= v\delta_{\alpha\gamma} + \sqrt{\theta} M^1_{\alpha\gamma} - \frac{\sqrt{\theta}}{2} \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\mu M^2_{\alpha\mu} - \kappa_\mu D^1_{\alpha\mu} \right) Q_\gamma, \\
C_{\alpha 1} &:= \left\{ \kappa_\gamma D^0_{\alpha\gamma} - \frac{1}{2} Q_\gamma \kappa_\gamma \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\gamma M^2_{\alpha\gamma} - \kappa_\gamma D^1_{\alpha\gamma} \right) \right\} \frac{\sqrt{\theta}}{\rho}, \\
C_{\alpha 2} &:= 0_\alpha \\
C_{\alpha 3} &:= \left\{ \frac{V^3_\alpha}{2} + \kappa_\gamma \frac{M^3_{\alpha\gamma}}{2} - \kappa_\gamma \frac{D^2_{\alpha\gamma}}{2} - \frac{3}{2} V^1_\alpha - \frac{3}{2} \kappa_\gamma M^1_{\alpha\gamma} + \kappa_\gamma D^0_{\alpha\gamma} \right. \\
&\qquad \left. - \frac{3}{4} Q_\gamma \kappa_\gamma \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\gamma M^2_{\alpha\gamma} - \kappa_\gamma D^1_{\alpha\gamma} \right) \right\} \frac{1}{\sqrt{\theta}}, \\
R_\alpha(W, \kappa) &:= (M^{-1})_{\alpha\beta} \langle \psi_\beta, \tilde{S}(\kappa, \rho, v, \theta) \rangle \overset{BGK}{=} -\frac{1}{\tau} \kappa_\alpha.
\end{aligned} \tag{4.23}$$

Now, (4.20) condenses to

$$\partial_t \kappa + B(W, \kappa) \partial_x \kappa + C(W, \kappa) \partial_x W = R(W, \kappa) \tag{4.24}$$

4.2.4 Coupling to Conservation Laws

The proper balance of ρ, v and θ can be ensured, if we solve the underdetermination problem in (4.24) by a heat flux coupling to the macroscopic balance laws. This way, the constitutive equations (4.24) for κ become a closure of the conservation law system – a closure that models the non-equilibrium phase space in adaptively riche scale resolution.

The coupled system of balance laws and constitutive equations in one space and one velocity dimensions then reads

$$\left.\begin{aligned} &\partial_t \rho + \partial_x(\rho v) = 0 \\ &\partial_t(\rho v) + \partial_x\left(\rho v^2 + \rho\theta\right) = 0 \\ &\partial_t\left(\tfrac{1}{2}\rho\theta + \tfrac{1}{2}\rho v^2\right) + \partial_x\left[\left(\tfrac{1}{2}\rho\theta + \tfrac{1}{2}\rho v^2\right) v + \rho\theta v + \tfrac{1}{2}q\right] = 0 \end{aligned}\right\} \quad \textbf{cons. laws}$$

$$\mathbf{q} = \theta^{\mathbf{3/2}}\rho\kappa_\gamma \mathbf{Q}_\gamma \qquad \textbf{coupling} \tag{4.25}$$

$$\partial_t \kappa + B(W,\kappa)\partial_x\kappa + C(W,\kappa)\partial_x W = R(W,\kappa) \quad \textbf{const. equations}$$

With this system, we have a formulation for any kind of perturbation functions ϕ_α. Before we consider specific choices, we have a look at the structure of (4.25).

The conservation law part of the coupled system corresponds to Euler's equations, extended by the heat flux q. For these equations, we call $U = (\rho, \rho v, \rho v^2 + \rho\theta)$ the conserved variables. There are one to one mappings from the primitive variables $W = (\rho, v, \theta)$ to U and vice versa. It is usually simpler to express the flux function in the primitive (non-conserved) variables,

$$\mathcal{F}(W,\kappa) = \begin{pmatrix} \rho v \\ \rho v^2 + \rho\theta \\ \left(\rho\theta + \rho v^2\right) v + 2\rho\theta v + q \end{pmatrix}, \tag{4.26}$$

For the constitutive equations, we can in general not define any flux functions, thus structurally, (4.25) can be written as partially conservative system

$$\partial_t \begin{pmatrix} U \\ \kappa \end{pmatrix} + \begin{pmatrix} \partial_x \mathcal{F}(W,\kappa) \\ B(W,\kappa)\partial_x\kappa + C(W,\kappa)\partial_x W \end{pmatrix} = \begin{pmatrix} 0 \\ R(W,\kappa) \end{pmatrix} \tag{4.27}$$

In the case of ϕ_α being hermite functions (see Sect. 4.3.1), there exists a conservative formulation of the constitutive equations. Knowing about its existence does unfortunately not provide us with any recipe of how to construct a flux function. For general ϕ_α, we do not even know, whether a conservative formulation exists.

It is not clear, what kind of mathematical properties we can expect for (4.27): Are there solutions and are they unique? Are the equations (locally) hyperbolic? Is there an

entropy? Is there some stability? We could intuitively argue that, the coupled system stems from the Boltzmann-BGK equation, which is hyperbolic and conservative, has an entropy, and with that a unique entropy solution. Unfortunately, a direct translation of these properties to our system seems unavailable. [9]

To summarize, a simple standard numerical scheme for hyperbolic equations will not be applicable to (4.27). We will design some mixed scheme in Sect. 4.4 and later validate hyperbolicity, existence and stability numerically through comparisons to fine scale discrete velocity BGK solutions. This validity will strongly depend on the strength of dissipation we get from the right hand side, $R(W, \kappa)$.

First, we are now considering more specific choices for the perturbation functions ϕ_α and analyse their approximation features.

4.3 Choice of Perturbation Functions

The choice of appropiate perturbation functions ϕ_α is motivated through the shape of the distribution function to approximate. One important class of distribution functions are the bimodals. Physically, they stem from shock tube problems and occur in the (possibly numerically diffused) shock region due to strong gradients in velocity and temperature. It is known that e.g. moment equations have difficulty capturing bimodalities, as we will see in more detail below.

Fig. 4.2 shows a typical bimodal distriubtion function,

$$\begin{aligned}\hat{\mathrm{f}}_{bimod}(\xi) = 0.16241 \exp\left(-0.45116(-0.8+\xi)^2\right) \\ + 0.812051 \exp\left(-6.34444(0.6+\xi)^2\right),\end{aligned} \tag{4.28}$$

and the corresponding function to be approximated by ϕ_α, namely $\exp\left(\xi^2/2\right)\left(\hat{\mathrm{f}} - 1\right)$.

The bimodal (4.28) is constructed such that it satisfies the compatibility conditions (4.9), furthermore it is normalized $\int \hat{\mathrm{f}}_{bimod}(\xi) d\xi = 1$. It will serve as a test case for the approximation features discussed in the next sections. In addition to good approximation features (mainly in the center of the plots since $\exp\left(-\xi^2/2\right)$ decays very fast in the tails), the combination of coefficients κ_α and perturbation functions ϕ_α also has to satisfy the compatibility conditions. This ensures the correct micro-macro relations between $\hat{\mathrm{f}}$ and its moments.

The discussion of approximation in the static case of $\hat{\mathrm{f}}_{bimod}$ will substantiate the approximation analysis for the (time and space dependent) PDE in Sect. 4.6.

First, we now consider two choices of ϕ_α, Hermite polynomials and splines.

[9] An overview about hyperbolicity, stability and entropy is provided in App. A.2.

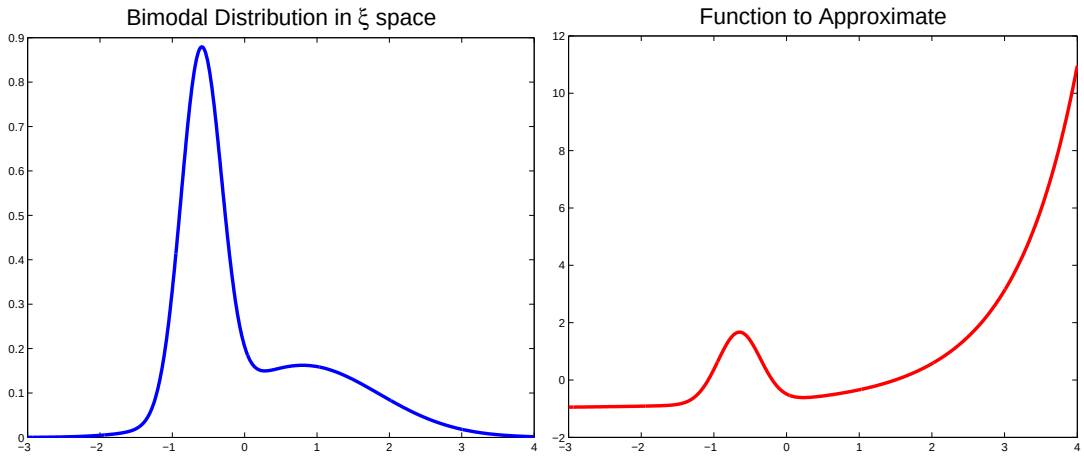


Figure 4.2: Left: a typical bimodal distribution; Right: the effective function to approximate

4.3.1 Hermite Polynomials

Hermite polynomials (see e.g. [63]) constitute a complete orthogonal set of functions in a weighted L^2 space with weight $\exp(-\xi^2/2)$. They are functions of the form

$$f(\xi) = \frac{1}{\sqrt{2\pi}} \exp(-\xi^2/2) \left(a_0 + a_1\xi + a_2\xi^2 + ... + a_N\xi^N + ...\right), \tag{4.29}$$

which (formally) corresponds to our Ansatz (4.8) with $a_0 = 1$ and $\phi_\alpha = \xi^\alpha$.

Grad (see [22]) was using a hermite series to approximate the Boltzmann distribution function, and closed his moment systems with the help of this series (see Part 3). Intuitively, the hermite functions should be a good model for gas dynamics in a polynomially disturbed equilibrium. Mathematically, they offer some nice properties (see below), but their critical drawback is that they are functions global in ξ. As such, they are not designed to model the multiscale behaviour of a bimodal distribution function.

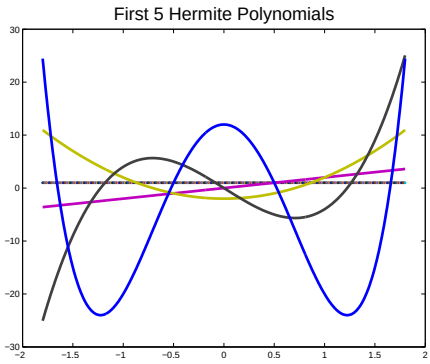
Definition and Properties

The Hermite polynomials are defined as

$$H_n(\xi) = (-1)^n \exp\left(\xi^2/2\right) \frac{d^n}{d\xi^n} \exp\left(-\xi^2/2\right), \tag{4.30}$$

which produces the first few polynomials

$$
\begin{aligned}
H_0(\xi) &= 1 \\
H_1(\xi) &= \xi \\
H_2(\xi) &= \xi^2 - 1 \\
H_3(\xi) &= \xi^3 - 3\xi \\
H_4(\xi) &= \xi^4 - 6\xi^2 + 3.
\end{aligned}
$$



One can show that the Hermite polynomials form an orthogonal set in the weighted L^2 space with gaussian weight $\exp\left(-\xi^2/2\right)$, i.e.

$$\int_{-\infty}^{\infty} H_n(\xi) H_m(\xi) \exp\left(-\xi^2/2\right) d\xi = n!\sqrt{2\pi}\delta_{nm}. \tag{4.31}$$

Indeed, they form a basis of this weighted L^2 space.

The construction of the Hermite polynomials can be done recursively,

$$H_{n+1}(\xi) = xH_n(\xi) - \frac{d}{d\xi}H_n(\xi) \tag{4.32a}$$

$$\frac{d}{d\xi}H_n(\xi) = nH_{n-1}(\xi) \tag{4.32b}$$

$$\rightarrow \quad H_{n+1}(\xi) = xH_n(\xi) - nH_{n-1}(\xi). \tag{4.32c}$$

Moments of a Hermite function $H_n(\xi)\exp(-\xi^2/2)$ can be computed analytically with the identities

$$
\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \xi^{2k} \exp\left(-\xi^2/2\right) d\xi = (2k-1)!!, \\
&\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \xi^{2k+1} \exp\left(-\xi^2/2\right) d\xi = 0, \quad k = 0, 1, 2, \ldots \quad .
\end{aligned}
\tag{4.33}
$$

Most practically, the Hermite polynomials satisfy the compatibility conditions (4.9):

Lemma 4.3.1.

$$(4.9)\forall n \geq 3: \quad \int_{-\infty}^{\infty} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} \exp\left(-\xi^2/2\right) H_n(\xi) d\xi = 0. \tag{4.34}$$

Proof. Because of orthogonality (4.31), we immediatly obtain the result for $1 = H_0$, $\xi = H_1$ and $\xi^2 = H_2 + H_0$. □

Since the first 3 Hermite functions are absorbed into the compatibility conditions, our first perturbation function is $\phi_1 = H_3$, and for higher accuracy, we take as many more Hermite functions as we wish. Since we want the integrals in (4.20) to remain finite, we take weighted monomials as test functions ψ_α. In Grad's approach, testing with the first three moments 1, ξ and ξ^2 yields equations for ρ, v and θ, formally equivalent to the conservation laws. These are however already coupled to our constitutive equations, in a convenient conservative formulation. So we choose

$$\psi_\alpha(\xi) := \frac{1}{\sqrt{2\pi}} \exp\left(-\xi^2/2\right) \xi^{\alpha+2}, \quad \phi_\alpha(\xi) := H_{\alpha+2}, \quad \alpha = 1, 2, ..., N. \tag{4.35}$$

All the matrices in (4.20) can be computed analytically with the identities (4.33).

Note that with (4.27) and (4.35), we can algorithmically construct Grad's equations to arbitrary orders - unfortunately not in conservative form.

Approximation Properties

In Fig. 4.3, we approximate our bimodal distribution (4.28) by a series of 4, 7 and 13 hermite functions. We observe a very limited capability of capturing the bimodal behaviour with 4 functions. Using more hermite functions, we get a 'better' approximation of the overall behaviour, but oscillations are present, and the peak is not well resolved.

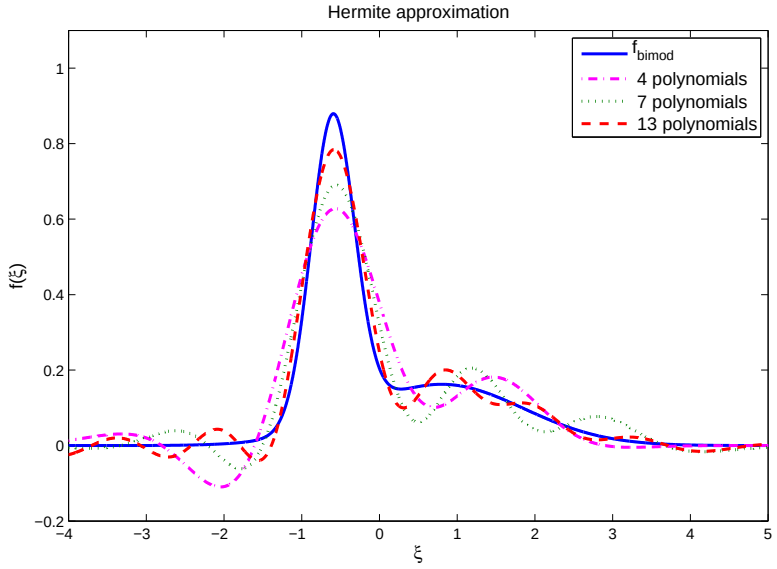


Figure 4.3: Examplified approximation features of Hermite functions.

In Fig. 4.4, we see the relative error

$$\frac{||\,\hat{\mathrm{f}}_{bimod} - \hat{\mathrm{f}}_{hermite}\,||}{||\,\hat{\mathrm{f}}_{bimod}\,||} \tag{4.36}$$

for various numbers of Hermite functions, in the L^1- and in the maximum-norm. As expected, the convergence in the maximum norm is slower than in the L^1-norm, due to the small scale oscillations and the underresolution of the peak, observable in Fig. 4.3. In the interesting cases of a few functions[10], the convergence is rather poor, and we will see that this can be overcome by a different choice of perturbation functions. Notice

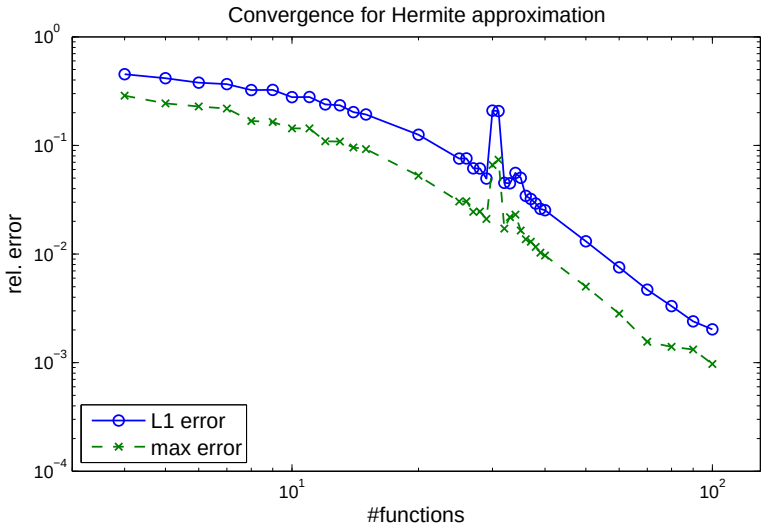


Figure 4.4: Convergence of Hermite functions.

the slightly singular behaviour around 30 Hermite functions. This is due to numerical instabilities in the computation of the L^2-projection onto the space spanned by the Hermite polynomials. The corresponding matrices have a very high condition number in the area of 30 Hermite functions, which translates into oscillations in the approximation. It is known that Hermite polynomials exhibit suboptimal numerical behaviour, as can already be guessed from their definition, yielding (uncontrollably) high values in (4.33).

[10] We are focussing on not too high dimensional PDE systems in order to remain competitive with discrete velocity solvers.

4.3.2 Splines

The Hermite functions provide a piecewise *global* approximation of our distribution function. We have seen that they converge rather poorly if we are using only a few of them (see Fig. 4.4). In this section we will consider a local, piecewise polynomial approximation through B-splines. B-splines have been developed for interpolation (see [47], Chapter 3.7 for a general introduction), which is *not* what we are doing with them: in interpolation, the task is to lay a curve through given data points. We want to model a function without any data points given.

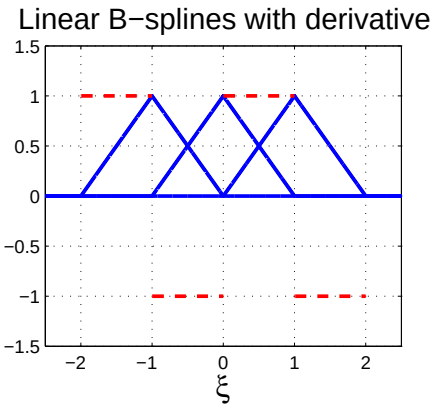
The locality of the B-splines will lead to a better convergence for only a few functions, as we will carefully analyse in Sect. 4.3.4. Furthermore, B-splines exhibit some nice mathematical features: they form a partition of unity and are (piecewise) differentiable. In contrary to the Hermite functions, B-splines do not fullfill the compatibility conditions (4.9), but this can be fixed (see following paragraphs).

The locality of the spline functions demands for one more discretization parameter: we need to choose the spread of the spline functions. Since we are choosing equidistant spline locations, spread and number of functions determine their location. The choice of the spread is crucial, if it is too small or too large, we lose relevant information of the distribution function. A reasonable spread is the interval $\xi \in [-3, 3]$. It can be computed that 99.73% of the area under the Gaussian $\frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}$ is captured over this interval[11]. With this interval, we can therefore capture the relevant information, if we are using enough spline functions. What 'enough' means will be discussed.

Definition and Properties

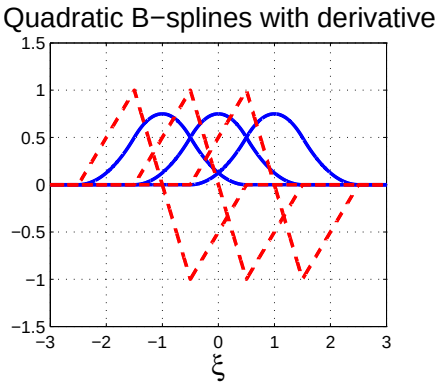
B-splines can be constructed to arbitrary polynomial order by recursion. They will always exhibit the maximally possible degree of smoothness, given the number of parameters. We will use splines of orders 1, 2 and 3. They read

$$B^{(1)}(\xi) = \begin{cases} 1 - \frac{x}{\Delta\xi} & : \quad 0 \leq \xi \leq \Delta\xi \\ 1 + \frac{x}{\Delta\xi} & : \quad -\Delta\xi \leq \xi < 0 \\ 0 & : \quad |\xi| > \Delta\xi \end{cases}$$

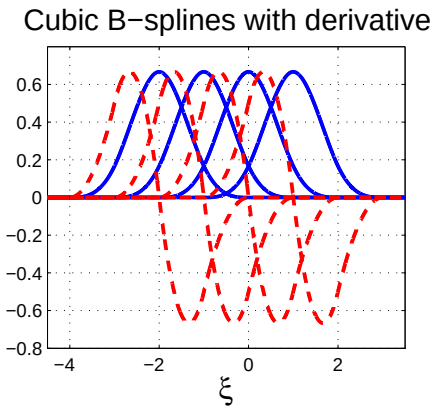


[11] In statistics, this fact is known as 3σ rule.

$$B^{(2)}(\xi) = \begin{cases} \frac{1}{8}\left(2\frac{\xi}{\Delta\xi} - 3\right)^2 & : \quad \frac{1}{2}\Delta\xi \le \xi \le \frac{3}{2}\Delta\xi \\ -\left(\frac{\xi}{\Delta\xi}\right)^2 + \frac{3}{4} & : \quad -\frac{1}{2}\Delta\xi \le \xi < \frac{1}{2}\Delta\xi \\ \frac{1}{8}\left(2\frac{\xi}{\Delta\xi} + 3\right)^2 & : \quad -\frac{3}{2}\Delta\xi \le \xi < -\frac{1}{2}\Delta\xi \\ 0 & : \quad |\xi| > \frac{3}{2}\Delta\xi \end{cases}$$



$$B^{(3)}(\xi) = \begin{cases} -\frac{1}{6\Delta\xi^3}\left(\xi - 2\Delta\xi\right)^3 & : \quad \Delta x \le \xi \le 2\Delta\xi \\ \frac{2}{3} + \xi^2\frac{\xi - 2\Delta\xi}{2\Delta\xi^3} & : \quad 0 \le \xi < \Delta\xi \\ \frac{2}{3} - \xi^2\frac{\xi + 2\Delta\xi}{2\Delta\xi^3} & : \quad -\Delta x \le \xi < 0 \\ \frac{1}{6\Delta\xi^3}\left(\xi + 2\Delta\xi\right)^3 & : \quad -2\Delta x \le \xi < -\Delta\xi \\ 0 & : \quad |\xi| > 2\Delta\xi \end{cases}$$



In the figures next to the formulae we see the spline functions (solid line) and their derivatives (dashed). For computational efficiency, it is crucial to have the invariant formulation in ξ, allowing for an invariant set of splines at all space and time points.

In Fig. 4.5 we see how our bimodal test curve (4.28) is approximated through a set of 4, 7 and 13 splines. Whereas the hermite approximations had severe problems resolving the 'peak' (compare Fig. 4.3), the splines exhibit higher accuracy there. Note that the choice of 4, 7 and 13 splines is not just arbitrary, but follows a hierarchy principle. We are using a minimum of 4 splines, a finer approximation uses the same 4 spline and fills the gaps in between them, leading to 7 splines, and so on. Without this hierarchy, we would not get any reasonable convergence behaviour for general distribution functions.

In Fig. 4.6, we compare the relative errors of approximation in numbers of splines in the maximum and in the L^1-norm. Both norms show approximately the same orders of convergence. The choice of degree has more impact onto the maximum norm, but its effect seems difficult to predict. Since the PDE approximation will be more delicate, it is not evident that higher polynomial degrees will perform better there.

A full comparison between convergence rates of hermite and various spline approximations for the PDE system can be found later in Sect. 4.6.

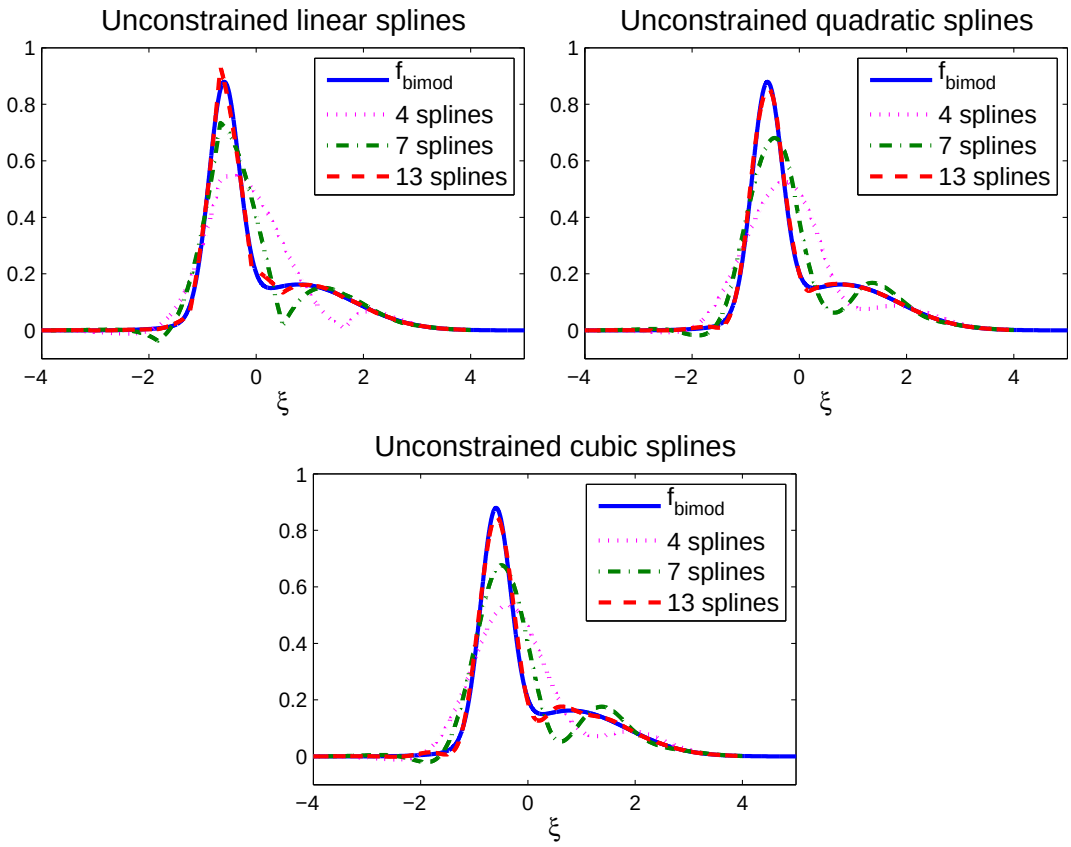


Figure 4.5: Various numbers of linear (left top), quadratic (right top) and cubic (bottom) splines approximate a bimodal test curve. The grid of spline centers ranges from $\xi = -3$ to $\xi = 4$.

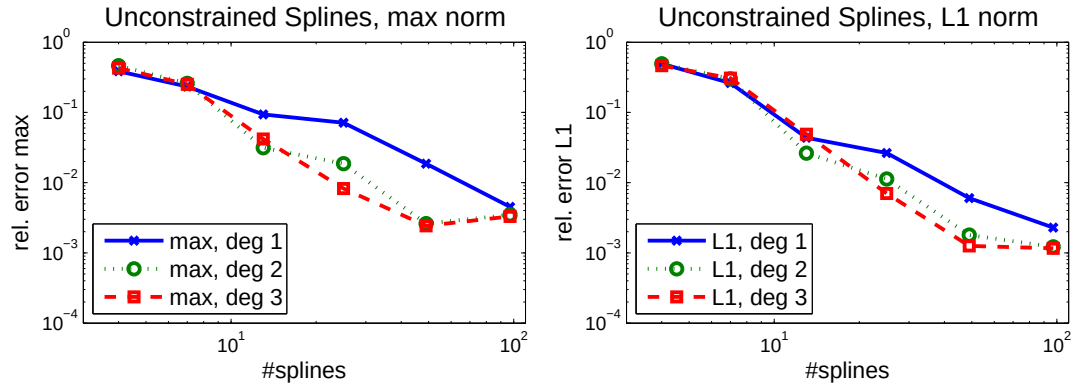


Figure 4.6: Convergence analysis for 4, 7, 13, 25, 49, 97 splines in maximum- (left) and L^1-norm (right).

4.3.3 Compatibility Conditions

So far, we have not been considering any compatibility between the spline model for $\hat{f}$ and the macroscopic fields ρ, v and θ. Such compatibility can be obtained in two ways: either by choosing a linear combination of B-splines that fullfills (4.10) (sufficient condition), or by projecting the coefficients κ_α onto the linear subspace that satisfies (4.9) (necessary and sufficient condition). We will analyze both methods in terms of their approximation features.

Compatible Linear Combination

We can make any set of perturbation functions ϕ_α compatible by considering a linear combination of four such functions,

$$\phi_\alpha^{comp} = a_0(\alpha)\phi_\alpha(\xi) + a_1(\alpha)\phi_{\alpha+1}(\xi) + a_2(\alpha)\phi_{\alpha+2}(\xi) + a_3(\alpha)\phi_{\alpha+3}(\xi), \quad \alpha = 1, ..., N-3. \tag{4.37}$$

To compute the coefficients $a_i(\alpha)$, we consider $\tilde{a}_3(\alpha) = 1$ and solve the following linear systems for $\tilde{a}_0(\alpha)$, $\tilde{a}_1(\alpha)$, $\tilde{a}_2(\alpha)$:

$$\sum_{j=0}^{2} \tilde{a}_j(\alpha) \int_{-\infty}^{\infty} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} \exp\left(-\xi^2/2\right) \phi_{\alpha+j}(\xi) d\xi = -\int_{-\infty}^{\infty} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} \exp\left(-\xi^2/2\right) \phi_{\alpha+3}(\xi) d\xi$$
$$\alpha = 1, ..., N-3 \tag{4.38}$$

These systems are well posed, if the perturbation functions ϕ_α are linearly independent and not all contained in the orthogonal complement to $(1, \xi, \xi^2)$. Had we chosen only 3 coefficients in (4.37), the resulting system would either be homogenous (3 unknown coefficients and zero right hand side) and yield only the trivial solution, or it would be non-quadratic (2 unknown coefficients, 1 right hand side) and only two of the three compatibility conditions could be satisfied.

After determinig $\tilde{a}_0(\alpha)$, $\tilde{a}_1(\alpha)$ and $\tilde{a}_2(\alpha)$ through (4.38), we balance

$$\begin{aligned} a_0(\alpha) &= \frac{\tilde{a}_0(\alpha)}{\sqrt{\tilde{a}_0(\alpha)^2 + \tilde{a}_1(\alpha)^2 + \tilde{a}_2(\alpha)^2 + 1}}, \quad a_1(\alpha) = \frac{\tilde{a}_1(\alpha)}{\sqrt{\tilde{a}_0(\alpha)^2 + \tilde{a}_1(\alpha)^2 + \tilde{a}_2(\alpha)^2 + 1}} \\ a_2(\alpha) &= \frac{\tilde{a}_2(\alpha)}{\sqrt{\tilde{a}_0(\alpha)^2 + \tilde{a}_1(\alpha)^2 + \tilde{a}_2(\alpha)^2 + 1}} \quad a_3(\alpha) = \frac{1}{\sqrt{\tilde{a}_0(\alpha)^2 + \tilde{a}_1(\alpha)^2 + \tilde{a}_2(\alpha)^2 + 1}}. \end{aligned} \tag{4.39}$$

Determining $a_i(\alpha)$ is quite some effort that, however, needs to be paid only once in every (PDE) computation: due to the invariant formulation, ϕ_α is independent of space and time, and the linear combinations can be computed once and for all in the beginning.

In Fig. 4.7, we see some typical shapes of the functions ϕ_α^{comp} for various α. Note that any effects occuring at the boundaries of the computational domain do not influence

our results by much since the Gaussian $e^{-\xi^2/2}$ exhibits a very fast decay. The loss of information through using only $N-3$ functions ϕ^{comp} compared to N functions ϕ is not relevant for high N, but can make a difference for small N.

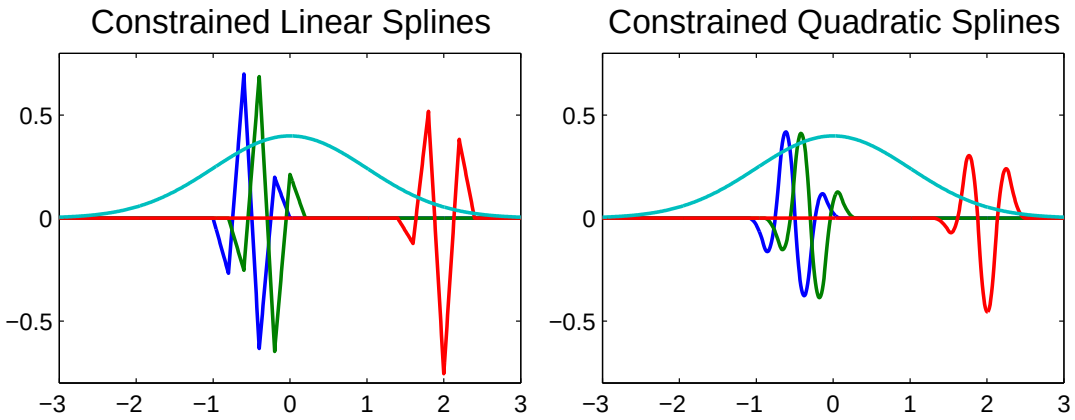


Figure 4.7: Linear (left) and quadratic (right) compatible B-splines

Subspace Projection

We can also directly impose (4.9) on the coefficients κ_α. This yields a necessary and sufficient condition for compatibility.

Mathematically, we satisfy (4.9) by orthogonally projecting the non-compatible parts of the coefficients onto the kernel of the $3 \times N$-matrix

$$A_{\cdot,\alpha} = \int_{-\infty}^{\infty} \begin{pmatrix} 1 \\ \xi \\ \xi^2 \end{pmatrix} \exp\left(-\xi^2/2\right) \phi_\alpha(\xi) d\xi. \tag{4.40}$$

Technically, we do a singular value decomposition $A = USV^T$, with orthogonal U and V (see e.g. [42]). Since A will have rank 3 if the ϕ_α are linearly independent and not all L^2-orthogonal to $(1, \xi, \xi^2)$, the columns 4 till N of V are an orthogonal basis of the kernel of A, and we can easily project our coefficients κ_α onto that space ($\mu = 1, ..., N-3$),

$$\kappa_\alpha^{proj} = V_{\gamma,\mu+3} \kappa_\gamma V_{\alpha,\mu+3}. \tag{4.41}$$

The projector P onto the kernel of A can thus be written as

$$P_{\alpha\gamma} := V_{\alpha,\mu+3} V^T_{\mu+3,\gamma}, \tag{4.42}$$

and $\kappa^{proj} = P\kappa$.[12]

[12] Remark that we could also construct a projector via the pseudo inverse of A, as done in Sect. 3.3. This does not necessarily lead to the same projector as the above construction.

Note, that the singular value decomposition of A is again done once in the beginning of a PDE computation since the matrix A is space and time independent. The projection, which is a matrix vector multiplication, needs to be done at every space-point and every time step.

Fig. 4.8 shows the relative difference of the projected and unprojected coefficients for spline degrees 1, 2 and 3. The compatible subspace grows richer and richer with the number of perturbation functions since the number of constraints remains the same. Therefore, we can expect that the difference between projected and unconstrained κ_α will decrease if we are using more perturbation functions. Despite that the difference is small in this setting of approximating one bimodal, it is essential to remain compatible in the PDE system where the errors add up in every time step.

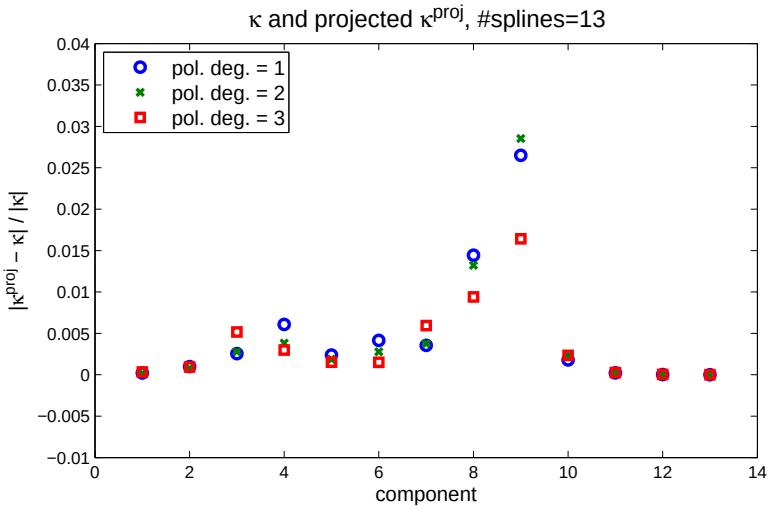


Figure 4.8: Relative error of unconstrained and projected κ, approximating f_{bimod} with 13 spline functions.

4.3.4 Discussion

Which Spline Approach?

In Fig. 4.9, we see that there is no significant difference between the 3 approaches (unconstrained spline, projected spline and constrained spline) up to the use of 25 perturbation functions. There is no clear advantage in terms of approximation of the projected (necessary and sufficient) compatibility and the (sufficient) compatibility enforced on the single perturbation functions ϕ_α. Neither is there any significant difference in the polynomial degree for the shown L^1-errors. Given these figures, it seems reasonable to choose the most simple version, i.e. the projected splines of polynomial degree 1.

There are two more features that are nicely visible in Fig. 4.9. One is the saturation effect that becomes more and more pronounced with higher polynomial degrees for the unconstrained splines. This is due to two effects at the boundary: around the boundary of the spline domain, the splines are not a partition of unity anymore (we do not consider fractional splines). Cubic splines need an overlap of four functions to yield unity on some interval, quadratic splines need three and linear splines need two functions. Therefore, for higher polynomial degrees, the area at the boundary without the partition of unity property is larger, this causes a part of the stagnation in convergence at high numbers of splines. The more significant part comes from the tail contribution of f_{bimod}. We only have a finite intervall with splines, and therefore do always have some approximation errors due to the non-zero (but exponentially decaying) tails.

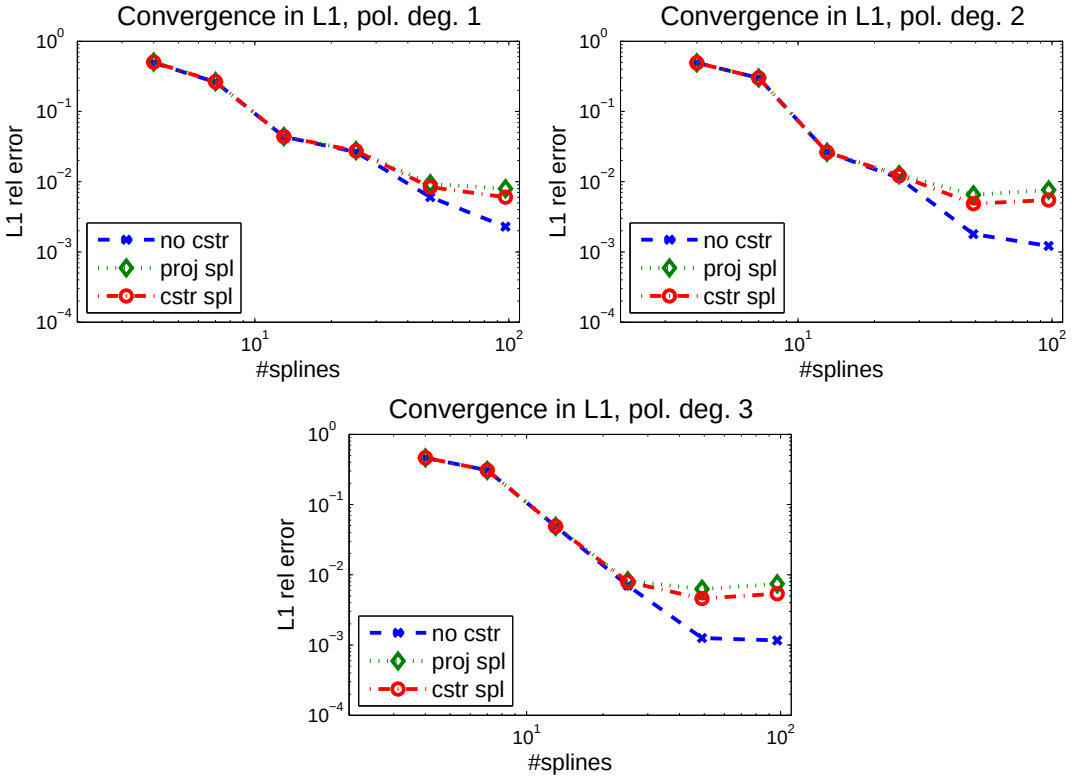


Figure 4.9: Convergence comparison for 4, 7, 13, 25, 49, 97 splines at polynomial degrees 1 (top left), 2 (top right) and 3 (bottom), unconstrained, projected and constrained versions.

The other effect is the opening gap between constrained and unconstrained approximations. This gap becomes visible for higher numbers of splines due to the increased accuracy (we see a loglog plot). One could wonder why there is a difference at all between the three methods since f_{bimod} is compatible. When we project f_{bimod} to the spline functions (sub)space, this hapens L^2-optimally in terms of approximation accu-

racy. The compatibility projection now applies to the reduced spline space, which does not necessarily gain approximation quality towards the full (compatible) function space this way. Compatibility on the full space does not yield exactly the same restrictions as compatibility on the subspace. The same argument holds for the compatible ϕ_α's: we again ensure compatibility on a subspace, which does not necessarily improve this subspace's approximation features in view of the full (compatible) space. But as we see in Fig. 4.9, the difference is negligible in comparison to the conceptional advantage we gain with compatible approximation.

Hermite or Splines?

Mathematically, Hermite functions offer very nice properties and seem a natural choice because they fulfill the compatibility conditions without further ado. They also lead to a very nice set of PDE's, the Grad equations. Conceptionally, splines have the advantage of being local approximations. This makes them much more flexible to adapt to multiscale phenomena, as we have it in our bimodal distribution function. Mathematically they may be inferior to the concept of Hermite functions, but modelwise they clearly are the method of choice.

The numerical convergence analysis in Fig. 4.10 complies with these arguments: The L^1-convergence is significantly better for the splines in the interesting order of spline numbers (10 to 20). The maximum-norm convergence depends on the polynomial degree of the splines. For polynomial degree 1, we have no significant difference between the splines and Hermite (again in the interesting regime); for polynomial degree 2 a small difference in favour of the splines is visible also for the maximum-norm error.

It would be venturesome to conclude from this pure approximation analysis to a specific choice of perturbation functions, namely B-splines – the PDE will most likely exhibit features that are not captured here. The more realistic analysis of the PDE results in Sect. 4.6 will indeed suggest, that splines exhibit good approximation features for shock tube problems, but that Hermite functions on the other hand are more accurate for cases with smooth solutions.

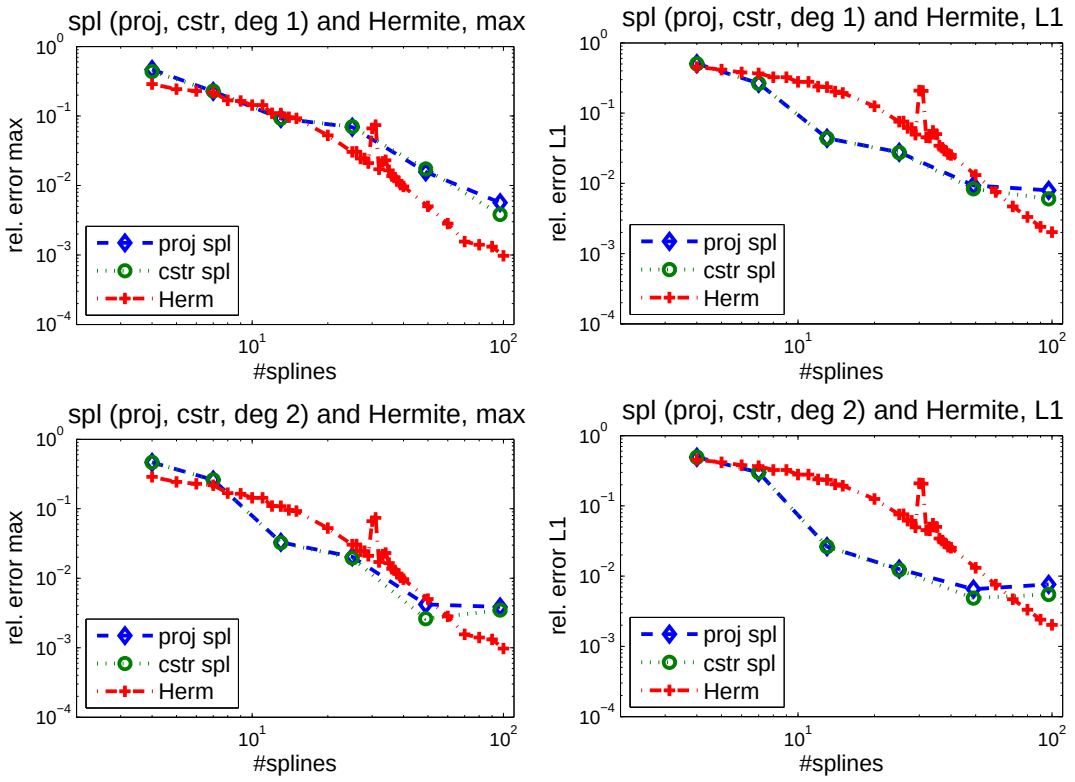


Figure 4.10: Convergence comparison for 4, 7, 13, 25, 49, 97 splines at polynomial degrees 1 (top), and 2 (bottom) with Hermite functions.

4.4 Numerical Methods

In this section, we will review some basics for numerical schemes to solve hyperbolic conservation laws. Focusing on approximate Riemann solvers (Sect. 4.4.1), we will extend the Rusanov scheme (Sect. 4.4.2) to yield solutions of our partially conservative system (4.27). We will motivate this extension through a linear stability analysis in Sect. 4.4.3.

4.4.1 Basic Definitions

In order to motivate a discretization for our system (4.27), we first consider quasilinear systems of hyperbolic conservation laws

$$\partial_t u + \partial_x f(u) = 0,\ u(x,0) = u_0(x) \tag{4.43}$$

with $\mathbb{R} \times \mathbb{R}_0^+ \ni (x,t) \mapsto u(x,t) \in \mathbb{R}^m$ a possibly discontinuous solution and $\mathbb{R}^m \ni u \mapsto f(u) \in \mathbb{R}^m$ the flux function. Hyperbolicity means that the eigenvalues of the Jacobian $Df(u)$ are real, quasilinearity means that f depends on (x,t) only through u. Derivatives are understood in the weak sense.

Solutions to (4.43) are not uniquely determined by initial data and flux function f. We select so called vanishing viscosity solutions, which are limits as $\varepsilon \to 0^+$ of the viscous system

$$\partial_t u + \partial_x f(u) = \varepsilon \partial_{xx} u. \tag{4.44}$$

Such solutions satisfy an entropy inequality,

$$\partial_t \eta(u) + \partial_x h(u) \leq 0, \tag{4.45}$$

with a convex entropy function $\mathbb{R}^m \ni u \mapsto \eta(u) \in \mathbb{R}$. $\mathbb{R}^m \ni u \mapsto h(u) \in \mathbb{R}$ is called entropy flux.

For the discretization of (4.43), we consider 3-point explicit finite volume schemes

$$u_j^{n+1} = u_j^n - \frac{\Delta \mathrm{t}}{\Delta \mathrm{x}} \left(F_{j+1/2} - F_{j-1/2} \right), \tag{4.46}$$

where $F_{j+1/2} = F(u_j^n, u_{j+1}^n)$, $F_{j-1/2} = F(u_{j-1}^n, u_j^n)$ is a *numerical* flux function.

Space and time are discretized by means of lattices $\{x_0 + j\,\Delta \mathrm{x}, j \in \{0,1,...,J\}\}$ and $\{n\,\Delta \mathrm{t}, n \in \{0,1,...,N\}\}$, and u_j^n is a cell average,

$$u_j^n = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} u(\tilde{x}, n\,\Delta \mathrm{t}) d\tilde{x}, \tag{4.47}$$

with $x_{j \pm 1/2} = x_0 + \left(j \pm \frac{1}{2} \right) \Delta \mathrm{x}$.

Definition 4.4.1. *(Consistency)*
F is called consistent with f if $F(u,u) = f(u)$.

Every consistent numerical flux F can be written as

$$F(u,v) = \frac{f(u)+f(v)}{2} + d(u,v), \qquad d(u,u) = 0. \tag{4.48}$$

Definition 4.4.2. *(Stability)*
An arbitrary explicit numerical time stepping scheme of the type

$$u_j^{n+1} = H_{\Delta t}(u^n;)$$

is called stable (in the sense of Lax-Richtmeyer, see [35]), if for n evaluations of $H_{\Delta t}$, $H^n_{\Delta t}$, we have

$$||H^n_{\Delta t}(u)|| < C_{LR}||u||.$$

Here, $||.||$ is some norm and $C_{LR} \in \mathbb{R}$ a constant independent of n..

A theorem by Lax and Wendroff (see [27]), states that limits of the scheme (4.46) solve the weak form of (4.43) if the numerical flux is Lipschitz continuous, consistent with the continuous flux and the scheme produces approximations with finite total variation in space. Furthermore the discrete values converge to a *vanishing viscosity solution* of (4.43), if there is a discrete entropy flux consistent with the continuous entropy flux and a corresponding discrete version of (4.45).

Scalar linear Advection

To construct specific schemes, let us consider a scalar linear advection equation with constant velocity a,

$$\partial_t u + \partial_x(au) = 0, \quad u(x,0) = u_0(x). \tag{4.49}$$

The solution to this equation reads $u(x,t) = u_0(x-at)$ and describes the advection of a profile $u_0(x)$ at (signal) velocity a. The solution at time t in the point x is influenced by the values of u_0 in the (analytic) *domain of dependence* $[x-at, x]$, $[x, x-at]$ respectively for $a > 0$ and $a < 0$.

A necessary condition for convergence of any numerical scheme is that the numerical domain of dependence is included in the analytical domain of dependence. A useful tool to verify this property is the CFL number:

Definition 4.4.3. *(CFL number)*
The CFL (Courant-Friedrich-Lewy) number of a scheme is

$$CFL = a\frac{\Delta t}{\Delta x}. \tag{4.50}$$

If $CFL < 1$, the inclusion property is satisfied and within one time step Δt, the profile does not travel further than the size of a space cell Δx. Note that the CFL condition gives us a maximal possible time step depending on Δx and a for 3-stencil schemes like (4.46). The faster the signal velocity or the finer the space grid, the smaller we must choose Δt.

A straight forward discretization of (4.49) can be achieved through a central difference scheme in space,

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0. \tag{4.51}$$

A Taylor expansion in space of $u_{j+1}^n - u_{j-1}^n$ around x_j yields the modified equation

$$\partial_t u + a\partial_x u = -a\frac{1}{6}\Delta x^2 \partial_x^3 u. \tag{4.52}$$

Apart from other drawbacks, the discretization through central differences does not lead to a vanishing viscosity solution, there is no diffusive term $\varepsilon\partial_{xx}$ involved in (4.52). We therefore correct the central difference scheme (4.56) in order to obtain such a contribution,

$$\frac{|a|}{\Delta x}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) \quad \longrightarrow \quad |a|\Delta x \partial_{xx} u \tag{4.53}$$

and obtain

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{a}{2}\frac{u_{j+1}^n - u_{ji-1}^n}{\Delta x} = \frac{1}{2}|a|\Delta x \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x}. \tag{4.54}$$

This scheme is called 'upwind scheme' and can be written as

$$u_j^{n+1} = u_j^n - \frac{\Delta \mathrm{t}}{\Delta \mathrm{x}}\left[a^-\left(u_{j+1}^n - u_j^n\right) + a^+\left(u_j^n - u_{j-1}^n\right)\right], \tag{4.55}$$

with $a^+ = \max(a,0) = \frac{1}{2}(a+|a|)$ and $a^- = \min(a,0) = \frac{1}{2}(a-|a|)$. Depending on the sign of a, this corresponds to using a forward difference $u_{j+1}^n - u_j^n$ (upwind) or a backward difference $u_j^n - u_{j-1}^n$ (downwind).

With the diffusive correction, we obtain the desired form of the modified equation,

$$\partial_t u + a\partial_x u = |a|\Delta x \partial_x^2 u + \mathcal{O}(\Delta x^3). \tag{4.56}$$

Diffusion has a stabilizing effect on a numerical scheme, as can be seen through a von Neumann analysis. For this, consider the Fourier transform $u(x,t) = \int e^{ikx}\hat{u}(k,t)dk$ and derive an ordinary differential equation in time for $\hat{u}$. For the upwind scheme (4.54), this yields a stable ordinary differential equation,

$$\partial_t \hat{u}(k,t) = -aik\hat{u}(k,t) - \Delta x k^2 \hat{u}(k,t). \tag{4.57}$$

For the central difference scheme (4.56), we would obtain

$$\partial_t \hat{u}(k,t) = -aik\hat{u}(k,t) + i\Delta x k^3 \hat{u}(k,t), \tag{4.58}$$

which causes oscillations without stabilization.

The equation (4.54) can be written in flux-form (4.46) with the numerical flux function

$$F(u,v) = \frac{a}{2}(u+v) + \frac{1}{2}|a|\,(u-v)\,. \tag{4.59}$$

System of Linear Advection Equations

With the form (4.59), we can generalize the upwind scheme for the linear scalar case (4.49) to a linear system

$$\partial_t u + \partial_x (Au) = 0. \tag{4.60}$$

Now we have $u \in \mathbb{R}^m$ and a constant matrix $A \in \mathbb{R}^{m\times m}$. We can solve this system analytically by decoupling it into m equations according to the eigendecomposition $A = T\Lambda T^{-1}$. Λ is the diagonal matrix of the eigenvalues $\lambda_1, ..., \lambda_m$ of A and T the corresponding transformation[13]. With the definitions $|A| = T|\Lambda|T^{-1}$, we can construct a consistent numerical (upwind) flux

$$F(u,v) = \frac{1}{2}A\,(u+v) - \frac{1}{2}|A|\,(v-u)\,. \tag{4.61}$$

The CFL condition now translates into a relation between the largest absolute eigenvalue of A and the space-time discretization.

Non-linear Systems

Generalizations of the numerical schemes to the non-linear case (4.43) require more sophisticated ideas. In [21], S. K. Godunov designed a scheme that considers non-linear Riemann problems with the analytic flux function f between all space cell interfaces,

$$\begin{aligned} \partial_t u + \partial_x f(u) = 0, \quad u(x,0) &= \begin{cases} u_j^n & x < x_j + 1/2\Delta x \\ u_{j+1}^n & x > x_j + 1/2\Delta x \end{cases} \\ &\Downarrow \\ u_{j+1/2}^{Riemann}&(x, t_n + \Delta t). \end{aligned} \tag{4.62}$$

At every time step, these intercell Riemann problems are solved, and new cell averages are computed. The time step has to be chosen such that the Riemann problems do not interact between more than one cell, which corresponds to $CFL < 1$.

Solving all these non-linear Riemann problems exactly is very costly and not always mathematically easy: depending on the structure of f, we get a solution that is a combination of shock waves, rarefaction waves and contact discontinuities, see [16]. In

[13] Note that hyperbolicity requires diagonalizability of A.

the case where f is a linear function, $f(u) = Au$, this simplifies to a combination of discontinuities according to the eigendecomposition of A, and the Godunov scheme reduces to the upwind scheme (4.59).

Since cell averageing smears out a lot of details of the exact Riemann problem solutions, we want to design *approximate Riemann solvers*. For this, we consider the integral consistency condition (see [27])

$$\begin{aligned}\int_{x_j}^{x_{j+1}} w_{j+1/2}^{appr}(x, t+\Delta t)dx &\stackrel{!}{=} \int_{x_j}^{x_{j+1}} u_{j+1/2}^{Riemann}(x, t+\Delta t)dx \\ &= \frac{\Delta x}{2}\left(u_j^n + u_{j+1}^n\right) - \Delta t f(u_{j+1}^n) + \Delta t f(u_j^n),\end{aligned} \tag{4.63}$$

where the second equation follows from integration of (4.43) over $t \in [t, t+\Delta t]$ and $x \in [x_j, X_{j+1}]$.

One specific approximate Riemann solver is proposed in [27] by Harten, Lax and van Leer (HLL scheme). There, we need lower and upper bounds a_L and a_R for the largest signal velocities and approximate the exact Riemann solution by one intermediate state,

$$w_{j+1/2}^{appr}(x/t, u_j^n, u_{j+1}^n) = \begin{cases} u_j & : x/t < a_L(u_j^n) \\ \tilde{u} & : a_L(u_j^n) < x/t < a_R(u_{j+1}^n) \\ u_{j+1} & : a_R(u_{j+1}^n < x/t \end{cases} \tag{4.64}$$

The value of $\tilde{u}$ follows from the consistency condition (4.63) and leads to a numerical flux function

$$F^{HLL}(u_L, u_R) = \frac{1+\alpha}{2} f(u_L) + \frac{1-\alpha}{2} f(u_R) + \frac{\beta}{2}\left(u_R - u_L\right), \tag{4.65}$$

with

$$\alpha = \frac{|a_L| - |a_R|}{a_L - a_R}, \quad \beta = \frac{a_R|a_L| - a_L|a_R|}{a_L - a_R}. \tag{4.66}$$

The HLL scheme can be simplified by taking $a = \max(|a_L|, |a_R|)$ and then setting $a_L = -a$ and $a_R = a$. This yields the Rusanov scheme (see e.g. [35]) with the numerical flux function

$$F^{RUS}(u_L, u_R) = \frac{1}{2}\left(f(u_L) + f(u_R)\right) - \frac{a}{2}\left(u_R - u_L\right), \tag{4.67}$$

In full, the Rusanov scheme reads

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{f(u_{j+1}^n) - f(u_{j-1}^n)}{2\Delta x} = \frac{1}{2} a \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x}. \tag{4.68}$$

This exactly corresponds to the upwind scheme in the scalar, linear advection case (4.54). So the Rusanov scheme is again a diffusion stabilized central difference approximation of the space derivatives, combined with an explicit Euler time update.

The diffusion operator for a (non-linear) system is simply approximated through the largest signal speed. The application of this approximate Riemann solver is much simpler than using a very complicated operator derived from the flux function f or using some full matrix diagonalization of the flux jacobian as examplified in the linear case of (4.61).

4.4.2 Scheme for the Coupled System

We will now use the ideas behind the Rusanov scheme to construct a diffusion stabilized central difference scheme for our equations (4.27). We had

$$\partial_t \begin{pmatrix} U \\ \kappa \end{pmatrix} + \begin{pmatrix} \partial_x \mathcal{F}(W,\kappa) \\ B(W,\kappa)\partial_x \kappa + C(W,\kappa)\partial_x W \end{pmatrix} = \begin{pmatrix} 0 \\ R(W,\kappa) \end{pmatrix},$$

with

$$\mathcal{F}(W,\kappa) = \begin{pmatrix} \rho v \\ \rho v^2 + \rho\theta \\ \left(\rho\theta + \rho v^2\right) v + 2\rho\theta v + q \end{pmatrix}.$$

We suggest the following mixed scheme

$$\begin{aligned} U_j^{n+1} &= U_j^n + \frac{\Delta t}{\Delta x}\left(F^{RUS}(W_{j-1}^n, W_j^n; \kappa_{j-1}^n, \kappa_j^n) - F^{RUS}(W_j^n, W_{j+1}^n; \kappa_j^n, \kappa_{j+1}^n)\right) \\ \kappa_j^{n+1} &= \kappa_j^n + B(W_j^{n+1}, \kappa_j^n)\frac{\Delta t}{2\Delta x}\left(\kappa_{j-1}^n - \kappa_{j+1}^n\right) + \frac{1}{2}s^{(n)}\left(\kappa_{j-1}^n - 2\kappa_j^n + \kappa_{j+1}^n\right) \\ &\quad + C(W_j^{n+1}, \kappa_j^n)\frac{\Delta t}{2\Delta x}\left(W_{j-1}^{n+1} - W_{j+1}^{n+1}\right) + R(W_j^{n+1}, \kappa_j^n) \end{aligned} \tag{4.69}$$

with

$$\begin{aligned} F^{RUS}(W_l, W_r; \kappa_l, \kappa_r) &= \frac{1}{2}\left(\mathcal{F}(W_l, \kappa_l) + \mathcal{F}(W_r, \kappa_r)\right) \\ &\quad - \frac{s(W_l, W_r)}{2}\left(U(W_r) - U(W_l)\right). \end{aligned} \tag{4.70}$$

As in Sect. 4.2.4, we use the primitive variables $W = (\rho, v, \theta)^T$ and the conservative ones $U = (\rho, \rho v, \rho v^2 + \rho\theta)^T$.

The maximum signal velocity $a(W_l, W_r)$ can be obtained through the primitive formulation of Euler's equations, which are the conservation laws without the heat flux q,

$$\partial_t \begin{pmatrix} \rho \\ v \\ \theta \end{pmatrix} + \underbrace{\begin{pmatrix} v & \rho & 0 \\ \theta/\rho & v & 1 \\ 0 & 2\theta & v \end{pmatrix}}_{A} \partial_x \begin{pmatrix} \rho \\ v \\ \theta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{4.71}$$

The eigenvalues of the matrix A are v, $v + \sqrt{3\theta}$, $v - \sqrt{3\theta}$, with the sound speed $\sqrt{3\theta}$. If we are just choosing the Euler speeds for s, we will in general underestimate the effective signal velocities, that are also influenced by non-linear terms. We are choosing a safety factor of $\sqrt{3}$ for the conservation law update,

$$s(W_l, W_r) = \max\left(|v_r + 3\sqrt{\theta_r}|, |v_l - 3\sqrt{\theta_l}|\right). \tag{4.72}$$

The value of $s^{(n)}$ in the diffusive part for the κ update is chosen as maximum signal velocity of all space points,

$$s^{(n)} = \max_{j \in \{0,1,\ldots,J-1\}}\left(s(W_j^n, W_{j+1}^n)\right). \tag{4.73}$$

The time step Δt can then be chosen adaptively according to the given CFL-condition and the computed value of $s^{(n)}$. If $s^{(n)}$ has been underestimated, this will show through instabilities in the numerical solution, in such a case, the CFL-condition can be adapted to a smaller value or the safety factor in (4.72) can be increased.

System (4.69) mimics the construction of the Rusanov scheme: The balance law part is discretized through the Rusanov scheme, additionally we put the heat flux $q(W, \kappa)$ inside the physical flux function $\mathcal{F}$. This corresponds to a direct central difference discretization of q.

The equation for the perturbation coefficients κ is discretized through a central difference in the κ and W variables. Adding diffusion only to the κ part and not to the W variables is the same as adding diffusion to the diagonal only, which is what happens in the Rusanov scheme.

A stability analysis for a linear model is done next in Sect. 4.4.3 and will motivate the choice of W^{n+1} in the time-update of κ. Numerical justification for our scheme will come from the results in Sect. 4.5 and Sect. 4.6.

Physically, the additional diffusion term is motivated by the regime of moderate Knudsen numbers that we are interested in. In this regime, shocks are smoothed by physical diffusion from the right hand side term, so the problems caused by strong discontinuities can be essentially avoided.

There are approaches to more generally deal with non-conservative systems, however it seems that there, a lot of additional information about the underlying physical structure of the system (dissipation mechanisms) becomes necessary (see [18]).

4.4.3 Stability Analysis in a Linear Model

In this section, we are considering a stability analysis for the scheme derived in Sect. 4.4.2 in a linear setting,

$$\partial_t \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} a & b \\ c & d \end{pmatrix} \partial_x \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{\tau} v \end{pmatrix}, \qquad a, b, c, d > 0. \tag{4.74}$$

Adding diffusion to one spatial derivative corresponds to using a linear upwind discretization instead of a central difference approximation for that derivative.

Solving (4.74) with the scheme $(a, b, c, d > 0)$

$$\begin{aligned} u_i^{n+1} &= u_i^n + \frac{\Delta t}{\Delta x} a(u_{i-1}^n - u_i^n) + \frac{\Delta t}{2\Delta x} b(v_{i-1}^n - v_{i+1}^n) \\ v_i^{n+1} &= v_i^n + \frac{\Delta t}{2\Delta x} c(u_{i-1}^{n+1} - u_{i+1}^{n+1}) + \frac{\Delta t}{\Delta x} d(v_{i-1}^n - v_i^n) - \frac{1}{\tau} v_i^n \end{aligned} \tag{4.75}$$

thus corresponds to solving our full equation with the scheme described in Sect. 4.4.2.

For the stability analysis of this scheme, we consider the two parameters

$$\nu := a\frac{\Delta t}{\Delta x}, \quad r := \frac{\Delta t}{\tau} \tag{4.76}$$

ν and r describe the freedom in the choice of discretization (Δx, Δt) and relaxation time τ. A von Neumann analysis allows us to compute the L^2-norm of the update operator and plot the combinations of ν and r where this norm is less than one. We see the domain of stability for the above scheme on top in Fig. 4.11 (parameters $a = c = d = 1$, $b = 2$).

In order to motivate the choice of (4.75), we compare this scheme to two slightly varied schemes. For a first variation, we replace the values u^{n+1} in the equation for v by u^n, which yields the scheme

$$\begin{aligned} u_i^{n+1} &= u_i^n + \frac{\Delta t}{\Delta x}a(u_{i-1}^n - u_i^n) + \frac{\Delta t}{2\Delta x}b(v_{i-1}^n - v_{i+1}^n) \\ v_i^{n+1} &= v_i^n + \frac{\Delta t}{2\Delta x}c(u_{i-1}^n - u_{i+1}^n) + \frac{\Delta t}{\Delta x}d(v_{i-1}^n - v_i^n) - \frac{1}{\tau}v_i^n \end{aligned} \tag{4.77}$$

Another version can be obtained by leaving the time updates as in (4.75), but using a diffusionless central difference approximation for the v-equation:

$$\begin{aligned} u_i^{n+1} &= u_i^n + \frac{\Delta t}{\Delta x}a(u_{i-1}^n - u_i^n) + \frac{\Delta t}{2\Delta x}b(v_{i-1}^n - v_{i+1}^n) \\ v_i^{n+1} &= v_i^n + \frac{\Delta t}{2\Delta x}c(u_{i-1}^{n+1} - u_{i+1}^{n+1}) + \frac{\Delta t}{2\Delta x}d(v_{i-1}^n - v_{i+1}^n) - \frac{1}{\tau}v_i^n \end{aligned} \tag{4.78}$$

In order to compare the 3 different schemes, we choose the parameters $a = c = d = 1$ and $b = 2$ and plot the domain of stability in dependence of r and ν in Fig. 4.11. Scheme (4.75) (top) allows us to use the highest CFL for very high τ values (small r values), and is therefore the scheme of our choice. Scheme (4.78) (middle) performs better for smaller τ (large r values), but that is not our primary interest. The performance of scheme (4.77) (bottom) is worse than the one of the other schemes.

The analysis presented in this section motivates the choice of our scheme (4.69) in a linear setting. Even though there are highly non-linear effects in our coupled system (4.27), the linear parts still contribute an important part to stability considerations. In this sense, the above results for the linear analysis are extended to the non-linear case.

4.4.4 Second and Higher Order

The HLL or Rusanov schemes as presented in Sect. 4.4.1 are of first order in space and time. An extension to second order is usually done with so called 'flux limiters' (see e.g. [35]). Our system, that is not in conservation form, can also be discretized to second order. We simply take the scheme (4.69), keep the central difference terms (which are already of second order) and do a second order reconstruction for the diagonal diffusive

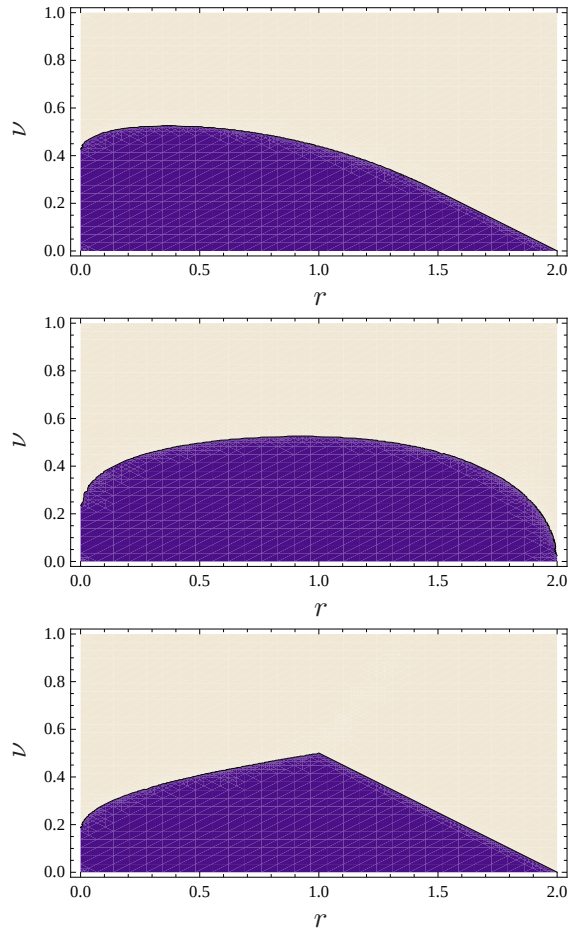


Figure 4.11: Domains of stability for the schemes (4.75) on top, (4.78) in the middle and (4.77) at the bottom. The parameters are $a = c = d = 1$ and $b = 2$.

part. The time update can then be done with the Heun method or other second order (Runge-Kutta) schemes, see [47]. Since modeling, and not numerics is our main focus, we will not consider second order extensions in the present work.

4.5 Grad5 - A Test Problem

In this section we will validate that the coupled system (4.25) contains Grad's equations for 5 moments if we choose the perturbation functions ϕ_α as Hermite functions. However, the resulting Grad system will not be in conservative form.

Grad's equations can be derived in conservative form directly from the Boltzmann equation (see [22]). There are various numerical solvers that work very well with these equations, among others the Rusanov and the HLL scheme from Sect. 4.4.

We will use the scheme (4.69) as developed in Sect. 4.4 and compare the results to Rusanov and HLL computations on the conservative formulation of Grad. Our test cases are a periodic system with smooth initial conditions and a shock tube problem. This way, we can get an idea of the scheme induced numerical error.

4.5.1 Grad Equations for 5 Moments

The derivation of Grad's equations directly out of the Boltzmann equation is straight forward, but needs some calculative efforts, especially in higher dimensions or for complicated collision kernels (see [22]). We can obtain a reformulation of Grad's equations through our PDE system (4.25), if we consider Hermite functions as perturbation functions ϕ_α and weighted monomials $\frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}\xi^\beta$ as test functions ψ_β. We will now explicitly do this for the 5 moment case in one space and one velocity dimension with a BGK collision kernel.

For simplification, we are using normalized Hermite functions

$$H_3(\xi) = \frac{1}{\sqrt{6}}\left(\xi^3 - 3\xi\right), \quad H_4(\xi) = \frac{1}{\sqrt{24}}\left(\xi^4 - 6\xi^2 + 3\right). \tag{4.79}$$

and define

$$\begin{aligned} \psi_1(\xi) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\xi^2/2\right)\xi^3, \qquad & \phi_1(\xi) &= \frac{1}{\sqrt{6}}\left(\xi^3 - 3\xi\right), \\ \psi_2(\xi) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\xi^2/2\right)\xi^4, \qquad & \phi_2(\xi) &= \frac{1}{\sqrt{24}}\left(\xi^4 - 6\xi^2 + 3\right). \end{aligned} \tag{4.80}$$

With the help of a computer algebra programm like Mathematica, it is easy to analyti-

cally compute the corresponding matrices (4.14) and obtain the equations[14],

$$\begin{aligned}
&\partial_t\kappa_1 + \left(v - 3\sqrt{\frac{3}{2}}\sqrt{\theta}\kappa_1\right)\partial_x\kappa_1 + 2\sqrt{\theta}\partial_x\kappa_2 + \frac{\sqrt{\theta}}{\rho}\left(-3\sqrt{\frac{3}{2}}\kappa_1^2 + 2\kappa_2\right)\partial_x\rho \\
&\quad + \frac{1}{\sqrt{\theta}}\left(\sqrt{\frac{3}{2}} - \frac{9}{2}\sqrt{\frac{3}{2}}\kappa_1^2 + 4\kappa_2\right)\partial_x\theta = -\frac{1}{\tau}\kappa_1 \\
&\partial_t\kappa_2 + \left(5\sqrt{\theta} - \sqrt{\frac{3}{2}}\sqrt{\theta}\left(\sqrt{6} + 4\kappa_2\right)\right)\partial_x\kappa_1 + v\partial_x\kappa_2 + \frac{\sqrt{\theta}}{\rho}\left(3\kappa_1 - \sqrt{\frac{3}{2}}\kappa_1\left(\sqrt{6} + 4\kappa_2\right)\right)\partial_x\rho \\
&\quad + \frac{1}{\sqrt{\theta}}\left(\frac{21}{2}\kappa_1 - \frac{3}{2}\sqrt{\frac{3}{2}}\kappa_1\left(\sqrt{6} + 4\kappa_2\right)\right)\partial_x\theta = -\frac{1}{\tau}\kappa_2.
\end{aligned} \tag{4.81}$$

Considering the full set of coupled equations, (4.25), we obtain a primitive formulation of Grad's 5-moment-equations, if we combine (4.81) with the conservation laws[15] in primitive form,

$$\begin{aligned}
\partial_t\rho + v\,\partial_x\rho + \rho\,\partial_x v &= 0 \\
\partial_t v + v\,\partial_x v + \frac{\theta}{\rho}\partial_x\rho + \partial_x\theta &= 0 \\
\partial_t\theta + v\,\partial_x\theta + 2\theta\,\partial_x v + \underbrace{\sqrt{6}\theta^{3/2}\partial_x\kappa_1 + \frac{3\sqrt{6}}{2}\sqrt{\theta}\kappa_1\partial_x\theta + \frac{\sqrt{6}}{\rho}\theta^{3/2}\kappa_1\partial_x\rho}_{\frac{1}{\rho}\partial_x q} &= 0.
\end{aligned} \tag{4.82}$$

A direct derivation of Grad's equations in conservative form for the BGK model leads to (compare e.g. [58])

$$\begin{aligned}
\partial_t\rho + \partial_x\rho v &= 0 \\
\partial_t\rho v + \partial_x\left(\rho v^2 + \rho\theta\right) &= 0 \\
\partial_t\left(\rho v^2 + \rho\theta\right) + \partial_x\left(\rho v^3 + 3\rho\theta\, v + q\right) &= 0 \\
\partial_t\left(\rho v^3 + 3\rho\theta v + q\right) + \partial_x\left(\rho v^4 + 6\rho\theta\, v^2 + 4q\, v + \Delta + 3\rho\theta^2\right) &= -\frac{1}{\tau}q \\
\partial_t\left(\rho v^4 + 6\rho\theta\, v^2 + 4q\, v + \Delta + 3\rho\theta^2\right) & \\
+\partial_x\left(\rho v^5 + 10\rho\theta\, v^3 + 10q\, v^2 + 5(\Delta + 3\rho\theta^2)\, v + 10\,\theta\, q\right) &= -\frac{1}{\tau}\left(\Delta + 4v\, q\right),
\end{aligned} \tag{4.83}$$

[14] For some intermediate steps, see App. A.5

[15] with the replacement of q in terms of κ according to (4.85)

and in primitive form

$$\begin{aligned}
\partial_t \rho + v\,\partial_x \rho + \rho\,\partial_x v &= 0 \\
\partial_t v + v\,\partial_x v + \frac{\theta}{\rho}\partial_x \rho + \partial_x \theta &= 0 \\
\partial_t \theta + v\,\partial_x \theta + 2\theta\,\partial_x v + \frac{1}{\rho}\partial_x q &= 0 \\
\partial_t q + v\,\partial_x q + 3\rho\,\theta\,\partial_x \theta + 4q\,\partial_x v + \partial_x \Delta &= -\frac{1}{\tau} q \\
\partial_t \Delta + v\,\partial_x \Delta + 6q\,\partial_x \theta + 5\Delta\,\partial_x v + 4\theta\,\partial_x q - 4\frac{q\,\theta}{\rho}\partial_x \rho &= -\frac{1}{\tau}\Delta.
\end{aligned} \tag{4.84}$$

One can compute that these variables relate to the κ variables in (4.81) through

$$q = \int_\infty^\infty (c-v)^3 f\,dc = \sqrt{6}\rho\theta^{3/2}\kappa_1, \quad \Delta = \int_\infty^\infty (c-v)^4 f\,dc - 3\rho\theta^2 = 2\sqrt{6}\rho\theta^2\kappa_2, \tag{4.85}$$

where q is the physical heat flux. The fourth order moment Δ has no direct physical meaning.

With the relations (4.85) and the help of the balance laws for replacing time derivatives of ρ, v, θ, we can indeed show the equivalence of (4.84) to (4.81) combined with (4.82).

Note here that with the general formulation of (4.25), we can derive Grad's equations algorithmically to arbitrary order. This can be useful for a numerical convergence analysis in number of functions of the Hermite series ansatz, which we will do in Sect. 4.6.2. The resulting PDE system will not be in conservative form which, as we will see in the next section, increases the discretization error.

4.5.2 Numerical Comparison for Grad5

We will now numerically compare schemes for the 5-moment-system of Grad.

We are considering the HLL scheme (see (4.64) and (4.65)) for the conservative form (4.83) and check whether the slightly more simple Rusanov scheme (see (4.68)) performs similarly well. We then compare these two numerical solutions for the conservative formulation to the mixed scheme that we derived in Sect. 4.4.2, (4.69).

This comparison will yield an idea of the error induced through the different formulations of Grad solved with different schemes, before we apply the mixed scheme (4.69) to more general Hermite and spline discretizations in Sect. 4.6.

Systematically, we will observe two sources of errors:

1) The equations for the perturbation coefficients κ are different to the conservative system of Grad, they relate as $q = \sqrt{6}\rho\theta^{3/2}\kappa_1$, $\Delta = 2\sqrt{6}\rho\theta^2\kappa_2$. This formulations are not equivalent if we consider general weak solutions. We expect that the difference of the two schemes varies with the relaxation time τ.

2) Discretization error (finite Δx, finite Δt). Also this error can vary with τ.

Error source 1) will combine with source 2). To analyze this combination, we look at comparisons between the schemes in a sequence of various discretizations Δx. As soon as we get a staable plateau in this sequence, we have a clear picture of 1). We will analyze 2) separately with convergence studies in Δx.

More precisely, for the convergence analysis we define relative discretization errors (de_r), i.e. errors of the type 2) as

$$de_r(\Delta x) := \frac{||f_{\Delta x} - f_{exact}||}{||f_{exact}||} \tag{4.86}$$

Since f_{exact} is usually not known, we use the finest scale discretization as a reference solution.

We define the relative model and discretization error mde_r as a combination of errors of types 1) and 2) between functions $f^{(a)}$ and $f^{(b)}$ from different models as

$$mde_r(\Delta x, a, b) := \frac{||f^{(a)}_{\Delta x} - f^{(b)}_{\Delta x}||}{||f^{(b)}_{\Delta x}||}. \tag{4.87}$$

In order to estimate the model errors, we will look at sequences of decreasing Δx. If we observe stationarity after a certain level of Δx, then this stationary picture reflects the error caused by the different models a and b. If there is no stationary behaviour in this sequence, the (main) source of error can be both, modeling or numerical discretization.

We will see that the difference in approximation quality of all the three approaches (conserved HLL, conserved Rusanov and mixed scheme) varies with the problem that we discretize. We will examine this for the two cases of a periodic boundary value problem with smooth initial conditions and a shock tube problem with open boundaries and discontinuous initial conditions.

If nothing else is indicated, the space variable x will be in the intervall $[-1, 1]$ and the time point is $T_{end} = 0.2$.

Smooth Initial Data

For $x \in [-1, 1]$, we consider periodic boundary conditions and smooth initial data,

$$\rho_0 = 2 + \frac{1}{2}\cos(\pi x)$$
$$v_0 = 1 + \frac{1}{2}\sin(\pi x)$$
$$p_0 = 1, \theta_0 = \frac{p_0}{\rho_0}$$
$$\text{higher moments} = 0.$$

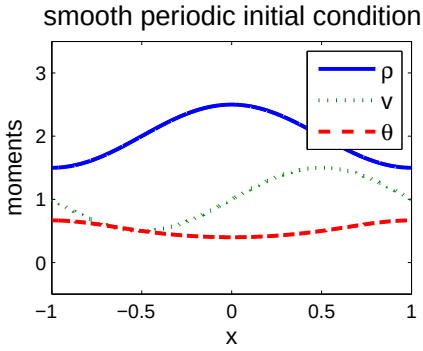


Fig. 4.12 shows the solution for $\tau = 0.1$ in ρ, v, θ, q and Δ at time $T = 0.2$ (Rusanov and HLL scheme[16]). With a CFL number of 0.9, this takes approximately 1600 time steps at $\Delta x = 0.0005$. For τ in the transition regime, we have physical damping of the initial periodic waves. Since we are using schemes at first order, there is additional numerical damping in the solution[17].

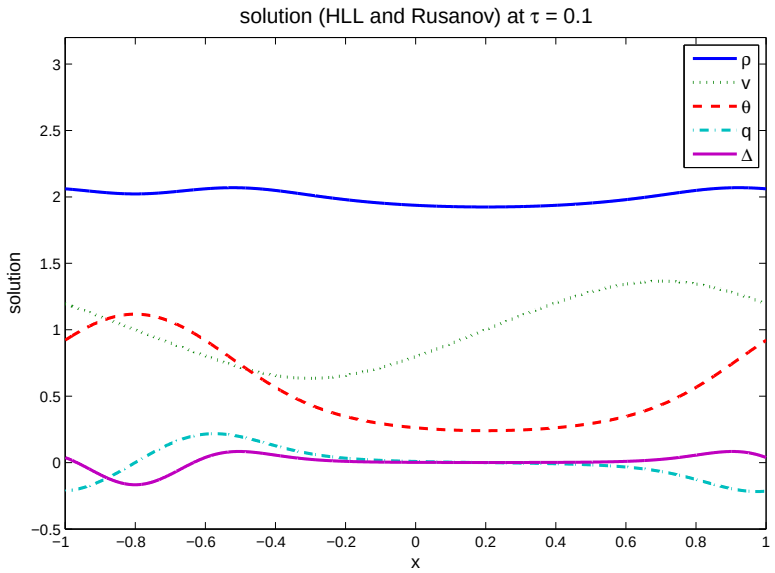


Figure 4.12: The solution (ρ, v, θ, q, Δ) at $\tau = 0.1$, $\Delta x = 0.0005$, $CFL = 0.9$.

In Fig. 4.13, we consider the convergence rate of the HLL- and of the Rusanov scheme for a given value of $\tau = 0.1$ in ρ. The quantity plotted is the relative discretization error, as described in (4.86). We use the discretization at $\Delta x = 0.0005$ as reference solution. We observe a fast rate in both, L^1- and maximum-norm and no significant difference between the schemes. Our studies revealed that indeed, the convergence behaviour is independent of τ, as long as we stay in the transition regime.

[16]The Rusanov and HLL schemes optically produce the same picture at this discretization level.
[17]The damping is not visible in this case here.

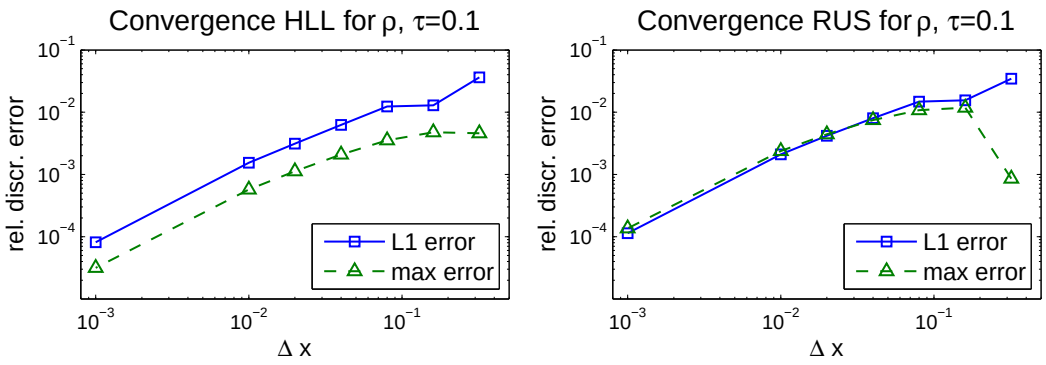


Figure 4.13: HLL (left) and Rusanov (right) scheme in ρ at $\tau = 0.1$ in various Δx (smooth case). $CFL = 0.9$. We consider the L^1-norm (solid) and the max-norm (dashed). The reference solution is approximated with $\Delta x = 0.0005$

Fig. 4.14 shows the convergence of the mixed scheme. On the right side, we see again the relative error at a given $\tau = 0.1$ varying in Δx. There is almost no difference to the behaviour of the Rusanov scheme. On the left, we show the last point in the convergence diagram for several τ (e.g. 10^{-4} for $\tau = 0.1$ as shown on the right). In this plot, we see very clearly that the convergence is not depending on τ for $\tau \geq 0.01$. For very small τ (0.001 and below), the numerical scheme experiences difficulties. This complies with the need for more sophisticated methods in the small τ regime, see [34].[18] A second fact that is well visible on the right plot of Fig. 4.14 is that accuracy decreases with higher moments. For ρ, the convergence rate is around two magnitudes better than the one for the heat flux q. This is expected since higher moments depend in more and more complicated, non-linear ways on the lower moments and thus numerical errors as well as modeling errors can build up (compare also to the figures in Sect. 3.9.2).

In Fig. 4.15, we consider the combination of modeling and discretization error between the Rusanov scheme and our mixed scheme, $mde(\Delta x, \text{Rusanov}, \text{mixed})$. We consider a sequence of $\Delta x = 0.001$ (left) and $\Delta x = 0.0005$ (right). We see that in the transition regime there is no significant difference between the two discretization levels, so the errors observed are essentially due to the differences in the schemes and not due to numerical discretization.

We see that the modeling error in Fig. 4.15 is τ-dependent: in ρ, we observe an increase towards $\tau = 0.1$ and then a decrease for larger τ. This τ-dpendence means that the models work differently on varying scales, something that is indeed expected.

We again note that the heat flux q is captured several orders of magnitudes less accurately. Also in q we observe a strong dependence on τ.

[18] We will see instabilities occuring in the small τ regime also in future calcualtions and will not comment them there anymore.

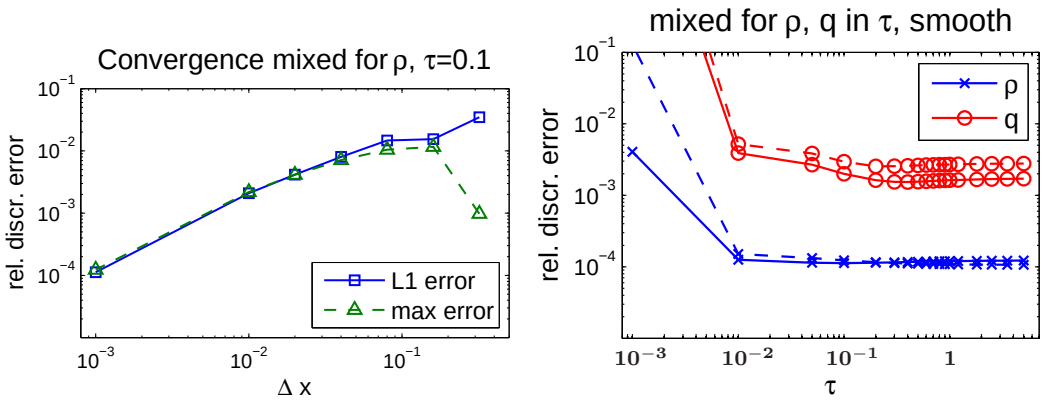


Figure 4.14: Convergence for the mixed scheme at $CFL = 0.9$. On the left for various Δx at fixed $\tau = 0.1$, on the right for various τ, $\Delta x = 0.001$. In both plots, we consider the L^1-norm (solid) and the max-norm (dashed), the reference solution is approximated with $\Delta x = 0.0005$

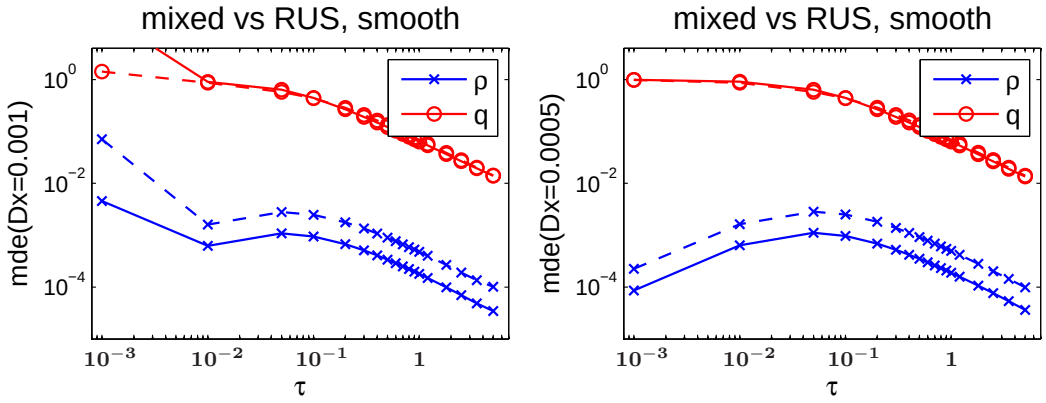
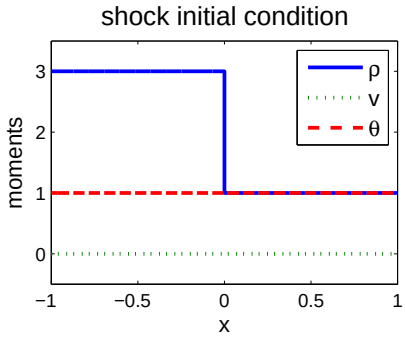


Figure 4.15: Modeling and discretization error (mde, see (4.87)) of the Rusanov and the mixed scheme in ρ and q for various τ, in L^1-norm (solid) and max-norm (dashed) at $\Delta x = 0.001$ (left), $\Delta x = 0.0005$ (right). $CFL = 0.9$.

Shock Tube Problem

We have seen an excellent match between the mixed scheme and the schemes for the conservative formulation for the smooth problem. Now, let us pose a more delicate non-smooth problem: For $x \in [-1, 1]$, we propose shock initial conditions[19] with open boundaries,

$$\rho_0 = \begin{cases} 3 & : x < 0 \\ 1 & : x > 0 \end{cases}$$
$$v_0 = 0$$
$$\theta_0 = 1$$
$$\text{higher moments} = 0.$$



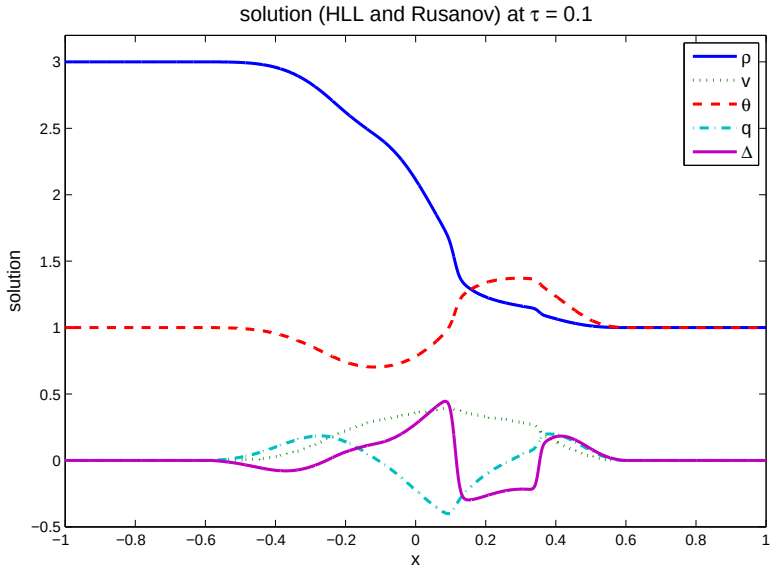


Figure 4.16: The solution (ρ, v, θ, q, Δ) at $\tau = 0.1$, $\Delta x = 0.0005$, $CFL = 0.9$.

Fig. 4.16 shows the solution for $\tau = 0.1$ in ρ, v, θ, q and Δ at time $T = 0.2$ with the same parameters as in Fig. 4.12. Here, physical (and numerical) damping are better visible, the initial shock structure is smoothed. In Fig. 4.17, we consider the convergence rate of the HLL- and of the Rusanov scheme for a given value of $\tau = 0.1$ in ρ. The quantity plotted is the relative discretization error, as described in (4.86). We approximate the exact solution with $\Delta x = 0.0005$. This figure compares to Fig. 4.13 for the smooth case.

[19] The shock sizes are in accordance with the Rankine-Hugonito conditions, see [16].

There, the L^1-convergence was one magnitude better (10^{-4}), and the maximum-norm error was even below the L^1-error. Now, typically for shock tube problems, the L^1-error is smaller than the maximum-error, this can be due to a slight miscatch of the shock speeds, leading to relatively high deviations of solutions at different discretization levels. The relative error of 10^{-3} at the lowest level is still quite accurate. Between the two schemes, HLL and Rusanov, there is no significant difference.

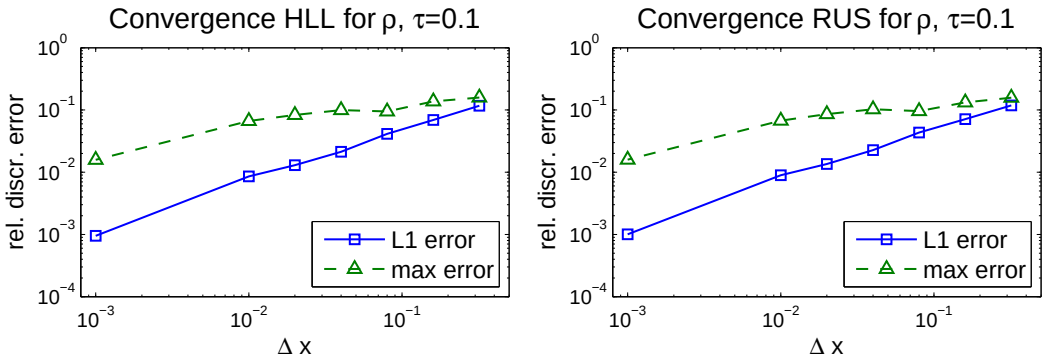


Figure 4.17: HLL (left) and Rusanov (right) scheme in ρ at $\tau = 0.1$ in various Δx, shock tube problem. $CFL = 0.9$. We consider the L^1-norm (solid) and the max-norm (dashed). The reference solution is approximated with $\Delta x = 0.0005$

Fig. 4.18 shows the convergence of the mixed scheme for the shock tube case. On the left side, we see again the relative error at a given $\tau = 0.1$ varying in Δx. Again, there is almost no difference to the behaviour of the Rusanov scheme. On the right, we show the last point in the convergence diagram for several τ (e.g. 10^{-3} for $\tau = 0.1$ as shown in the left plot). In this plot, we see a minor dependence of the convergence on τ in the regime of our interest. For very small τ (0.01 and below), the numerical scheme experiences difficulties as before in the smooth case. The accuracy decrease for higher moments compares to Fig. 4.14 in the smooth case: for ρ, the convergence rate is again around two magnitudes better than the one for the heat flux q.

In Fig. 4.19, we consider the combination of modeling and discretization error between the Rusanov scheme and our mixed scheme, $mde(\Delta x, \text{Rusanov}, \text{mixed})$. This figure compares to Fig. 4.15 in the smooth case. Like there, we consider a sequence of $\Delta x = 0.001$ (left) and $\Delta x = 0.0005$ (right). We see that in the transition regime there is no significant difference between the two discretization levels, so the errors observed are due to the differences in the schemes and not due to numerical discretization.

In analogy to the smooth case, we see that the modeling error in Fig. 4.19 is τ-dependent, and again, the heat flux q is resolved less accurately than ρ.

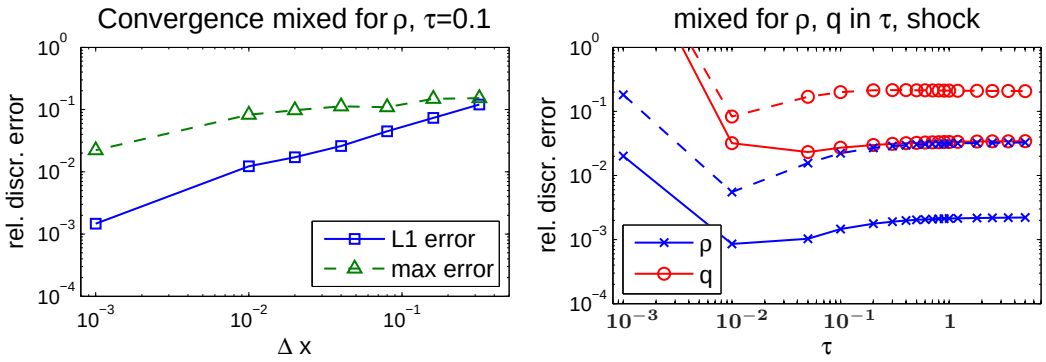


Figure 4.18: Convergence for the mixed scheme at $CFL = 0.9$. On the left for various Δx at fixed $\tau = 0.1$, on the right for various τ, $\Delta x = 0.001$. In both plots, we consider the L^1-norm (solid) and the max-norm (dashed), the reference solution is approximated with $\Delta x = 0.0005$

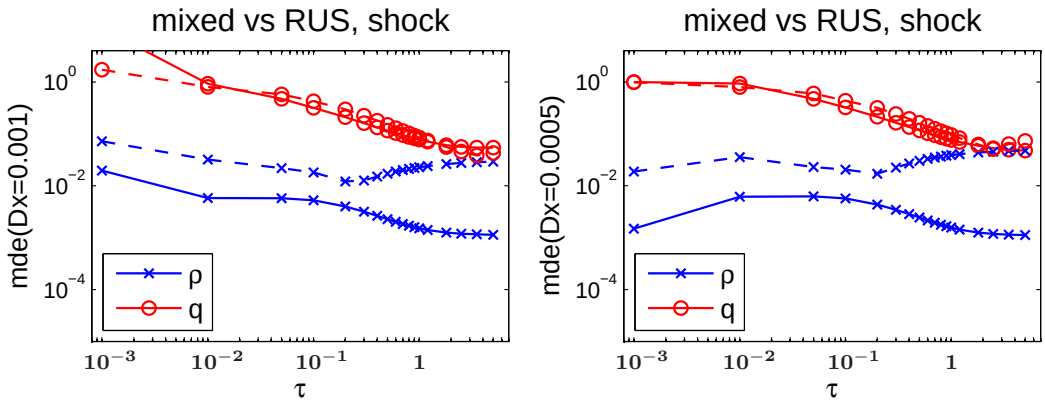


Figure 4.19: Modeling and discretization error (mde, see (4.87)) of the Rusanov and the mixed scheme in ρ and q for various τ, in L^1-norm (solid) and max-norm (dashed) at $\Delta x = 0.001$ (left), $\Delta x = 0.0005$ (right). $CFL = 0.9$.

4.6 Assessing the Model Quality

In this section we apply the mixed numerical scheme (4.69) to equations with several choices of perturbation functions (splines, Hermite functions), and compare the results to a fine grid discrete velocity BGK-solver.

The errors involved can be split into 3 contributions:

1) Modeling error of the perturbation function ansatz.
2) Numerical discretization error of the perturbation function schemes.
3) Numerical discretization error of the discrete velocity scheme.

The limiting accuracy is given through the errors 2) and 3), we can estimate their size through comparisons of the respective numerical results for various Δx and extend this to analyze 1) in terms of sequences in Δx, as done before in Sect. 4.5.2. We will consider the same periodic and shock tube problems as in that section.

If nothing else is indicated, the space variable x is again in the intervall $[-1, 1]$ and the time point is $T_{end} = 0.2$.

4.6.1 Discrete Velocity Solver for BGK

The results of our perturbation function approach will be compared to a fine scale discrete velocity solution of the Boltzmann-BGK equations (see (2.71)),

$$\partial_t f + c\partial_x f = \frac{1}{\tau}(F_M - f) \tag{4.88}$$

The numerics for such a solver are accurately manageable (at not too low τ, see [34]) since the advection is linear. For very precise results, numerical errors in the collision term would have to be compensated by the solver (see [38]). This leads too far here since the perturbation function scheme will not be that accurate.

To solve (4.88), we are using a Strang splitting (see [48]) between the advection part $\partial_t f + c\partial_x f = 0$ and the interaction process $\partial_t f = \frac{1}{\tau}(F_M - f)$.

In detail, we consider a discretization of the distribution function f on space and time as in Sect. 4.4.1, where $x_j \in \{x_0 + j\Delta x, j \in \{0, 1, ..., J\}\}$ and $t_n \in \{n\Delta t, n \in \{0, 1, ..., N\}\}$. In addition, we need a discretization of the velocity space through

$$c_k \in \{c_0 + k\Delta c, k \in \{0, 1, ..., K\}\}. \tag{4.89}$$

Thus the discrete distribution will carry 3 indices,

$$f^n_{j,k} := f(x_j, t_n, c_k). \tag{4.90}$$

The first step of the Strang splitting uses $\frac{\Delta t}{2}$ for the advection part,

$$f_{j,k}^{(n,1)} = f_{j,k}^{n} - \frac{1}{2}\frac{\Delta t}{\Delta x}\left(F_{j+1/2,k}^{n} - F_{j-1/2,k}^{n}\right), \tag{4.91}$$

with $F_{j\pm1/2,k}^{n}$ a (second order) minmod limited Lax-Wendroff flux (see e.g. [35]). The next step adds the interaction through a Heun update,

$$\begin{aligned} \tilde{f}_{j,k}^{(n,2)} &= f_{j,k}^{(n,1)} + \frac{\Delta t}{\tau}\underbrace{\left(F_M[f^{(n,1)}]_{j,k} - f_{j,k}^{(n,1)}\right)}_{A_{j,k}} \\ f_{j,k}^{(n,2)} &= \tilde{f}_{j,k}^{(n,2)} + \frac{1}{2}\frac{\Delta t}{\tau}\left(A_{j,k} + F_M[\tilde{f}^{(n,2)}]_{j,k} - \tilde{f}_{j,k}^{(n,2)}\right). \end{aligned} \tag{4.92}$$

The interaction update is non-local in c, since the Maxwellian F_M depends on ρ, v and θ, which are computed through numerical quadrature (in our case the trapeze rule, see e.g. [47]) from the discrete distribution values[20].

The final update for one time step consists of another advection step with $\frac{\Delta t}{2}$,

$$f_{j,k}^{n+1} = f_{j,k}^{(n,2)} - \frac{1}{2}\frac{\Delta t}{\Delta x}\left(F_{j+1/2,k}^{(n,2)} - F_{j-1/2,k}^{(n,2)}\right). \tag{4.93}$$

All the ingredients are of second order in space and time, the structure of the Strang splitting makes the splitting error also second order, so in total, we have a second order scheme.

Note that in one space and one velocity dimension this discrete velocity scheme for the Boltzmann-BGK equations works in reasonable CPU time. In higher dimensions, a fine velocity grid becomes computationally very expensive.

Discrete Velocity Convergence

We consider a velocity grid with $\Delta c = 0.001$ between $[-8, 8]$. This grid has to be chosen large enough such that for all space grid points in all time, all the essential parts of the distribution function can be resolved. As CFL number we choose 0.9. Since the advection part is linear, even diagonal, the maximal signal velocities are directly available, and stability should not be an issue at CFL numbers close to one for appropriate sizes of τ.

A typical solution for shock and smooth initial conditions at $\tau = 0.1$, $T_{end} = 0.2$, computed with $\Delta x = 0.001$ on an intervall of $x \in [-1, 1]$ is shown in Fig. 4.20. We see a strong damping in the shock tube problem (right figure) due to the high value of τ.

In Fig. 4.21, we see the relative discretization errors in ρ for the smooth and the shock initial conditions. We compare all the levels of Δx to the reference solution at $\Delta x =$

[20] Due to the non-locality in c, it is essential for efficiency that the operator $A_{j,k}$ is computed just once.

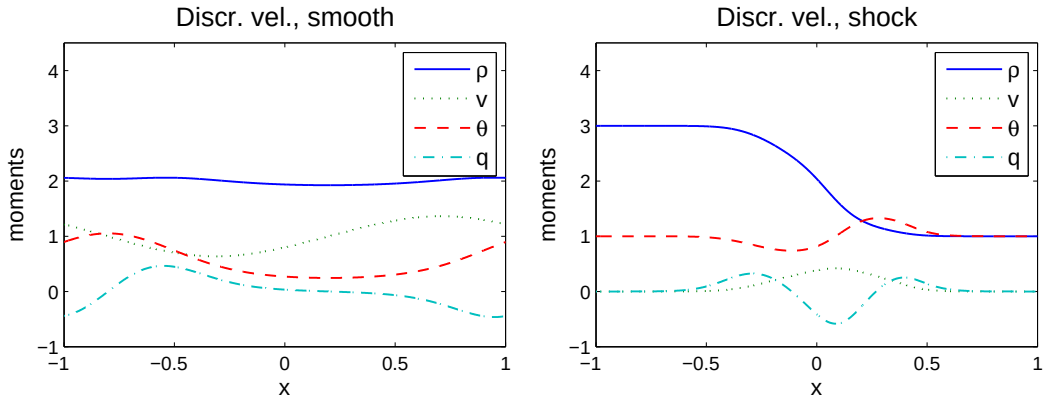


Figure 4.20: Smooth (left) and shock (right) solutions at $\tau = 0.1$, computed with $\Delta x = 0.001$, $CFL = 0.9$, $\Delta c = 0.001$ between $[-8, 8]$

0.001. The plots show a stable convergence order. We observe that the accuracy in the shock case is worse for the maximum norm, while it remains comparable for the L^1 norm. As explained in Sect. 4.5.2, this is to be expected. Similarly, it is expected that the accuracy becomes worse for higher order moments (not shown in the figure), which indeed happens also for the discrete velocity case. For q we observed maximal accuracies of approximately 10^{-2} for the same setting as in Fig. 4.21, in the smooth and schock case.

In the regime that interests us, the results for the L^1 norm do not depend on τ, the results for $\tau = 0.1$, $\tau = 0.4$ and $\tau = 0.5$ are essentially identical.

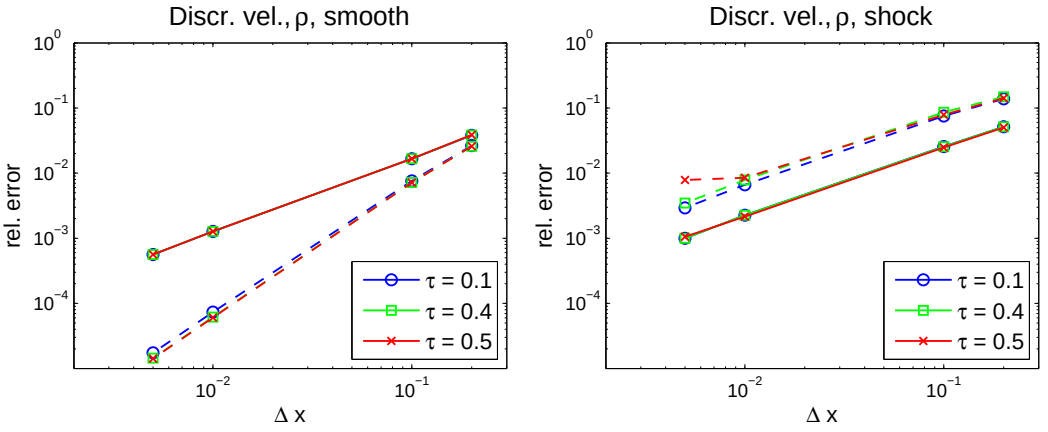


Figure 4.21: Rel. discretization errors of smooth (left) and shock (right) solutions in Δx as compared to a fine scale reference solution at $\Delta x = 0.001$. $CFL = 0.9$, $\Delta c = 0.001$ between $[-8, 8]$, various τ.

With these results, we can now proceed to analyse the model errors of the perturbation function schemes with confidence in the accuracy of the discrete velocity reference solution.

4.6.2 Perturbation through Hermite Polynomials

In this section, we are comparing our perturbation function scheme with Hermite polynomials to the results obtained from the discrete velocity calculations in Sect 4.6.1. The parameters are the same as before.

In Fig. 4.22, we analyze the purely numerical convergence of the mixed scheme with Hermite perturbation functions in dependence of τ. We use 13 perturbation functions and observe that the convergence rate does not depend significantly on τ in the regime that we are interested in. As before, we observe that the maximum-norm error is higher in the shock tube case (right plot). The finest scale relative error for both, smooth and shock initial conditions is in the same order of magnitude of 10^{-3}. The convergence rates as examplified in Fig. 4.22 look similar for all the numbers of perturbation functions (not shown).

Note that we are using $CFL = 0.4$ in the case of Hermite perturbation functions. Computations revealed that this setting is rather sensitive on the maximum signal velocities and that the safety factor of $\sqrt{3}$ in (4.72) is not large enough. This shows more and more clearly as the number of perturbation functions is increased. For 13 perturbation functions, the shock case could not be resolved with $CFL = 0.4$, but only with $CFL = 0.2$. For more functions, the CFL condition becomes even more restrictive. We will see in Sect. 4.6.4 that splines offer significant advantages in this view.

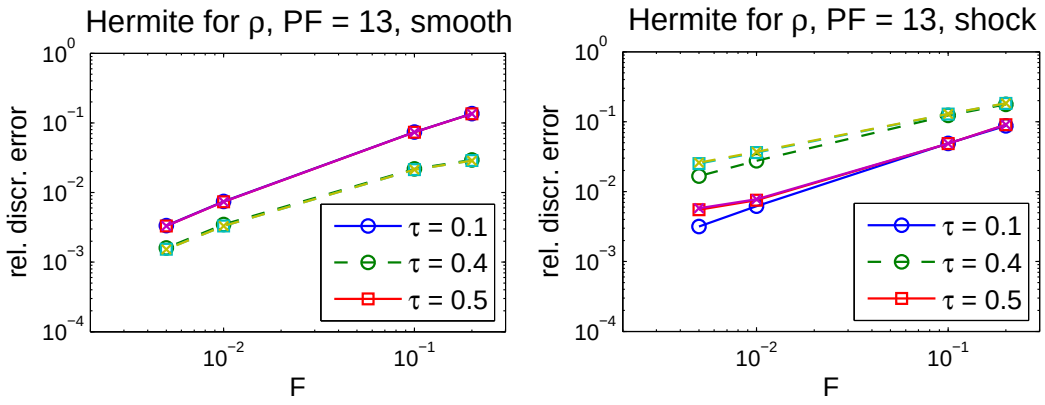


Figure 4.22: Rel. errors (maximum, dashed and L^1, solid) of smooth (left) and shock (right) solutions between $\Delta x = 0.005$ and $\Delta x = 0.001$, $CFL = 0.4$ (smooth) and $CFL = 0.2$ (shock), 13 perturbation functions.

In Fig. 4.23 we compare ρ as computed with the mixed Hermite scheme to the discrete velocity result at $\Delta x = 0.001$. We consider relative model errors as described in (4.87). An analysis of the sequence of $\Delta x \to 0.001$ (not shown) reveals that we do not yet observe a stable picture in the smooth case, but the shock tube case shows the same behaviour at $\Delta x = 0.005$ as the one we see in Fig. 4.23 at $\Delta x = 0.001$. This indicates, that for the shock case we indeed see an effective modelling error, whereas the smooth case might still be (significantly) influenced by discretization errors at this level of Δx.

The plots in Fig. 4.23 look as expected, the modelling error decreases with the number of perturbation functions used[21], and the smooth case is approximated much more accurately in terms of the modeling error. We expect that this is due to the non-locality of the Hermite functions. As explained in Sect. 4.5.2, the non-locality of the Hermite functions is more disadvantageous if we have shocks with strong bimodalities or irregular shapes of the distribution function. The disadvantage of the non-local approach also shows in the maximum-norm modeling error, which is larger than the L^1 error for the smooth and the shock case.

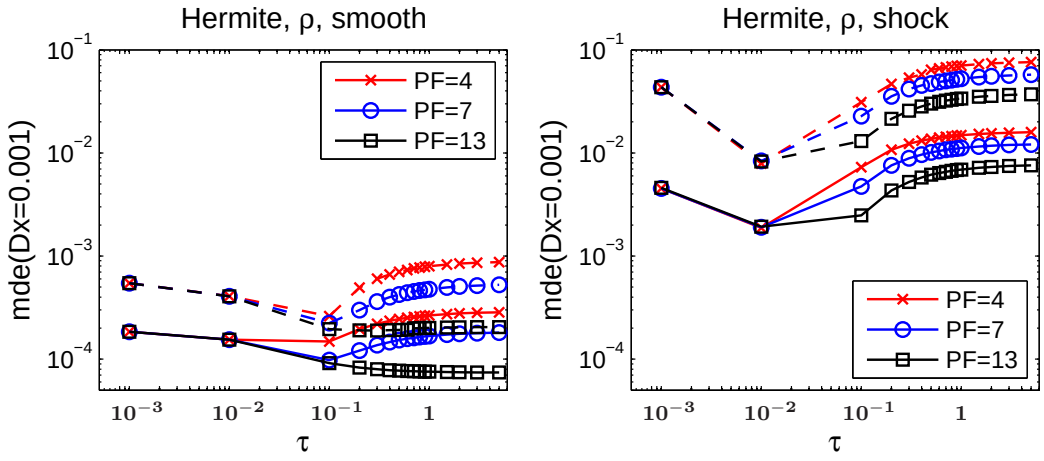


Figure 4.23: Rel. model errors (maximum, dashed and L^1, solid) of Hermite mixed scheme for smooth (left) and shock solutions (right) at $\Delta x = 0.001$ and $CFL = 0.4$ for 4 and 7 functions, $CFL = 0.2$ for 13 functions. Discrete velocity as reference solution (same Δx).

We will come back to the Hermite results in Sect. 4.6.4, where we will compare them to the spline results.

[21] The specific number of Hermite functions corresponds to the number of splines we will use in Sect. 4.6.3.

4.6.3 Perturbation through splines

We have intuitively argued in Sect. 4.3 that the globality of Hermite polynomials can be a disadvantage. This indeed showed through stability problems at already moderate CFL numbers. In this section, we are considering splines. We will first give evidence that degree one splines are a good choice. Then we will consider the compatibility conditions, as examplified in Sect. 4.3. We will not see significant differences between compatible splines and spline combinations projected onto a corresponding compatible subspace. Generally, the splines will allow for higher CFL-numbers than the Hermite functions and do not experience severe stability problems for higher numbers of perturbation functions.

The size of the spread for the splines has already been discussed in Sect. 4.3.2. Like there, we are using the interval $[-3, 3]$ for the equidistant spline centers.

Which order of Spline Functions?

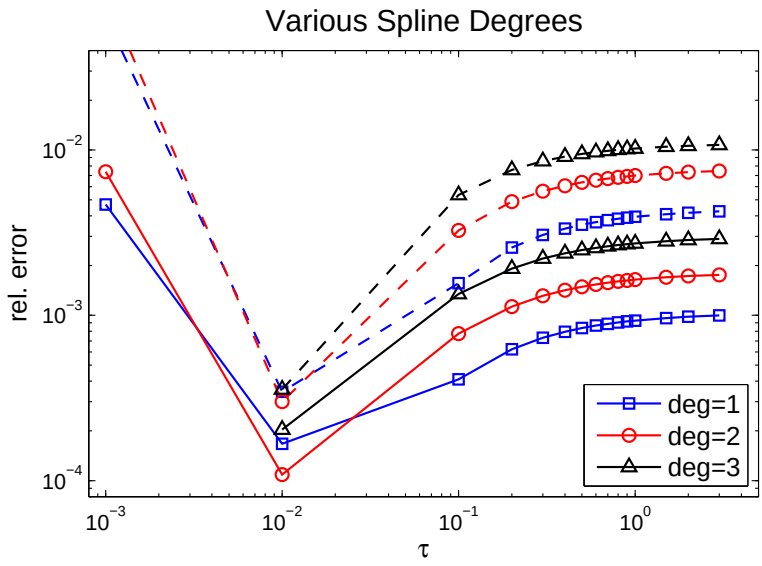


Figure 4.24: Relative errors (dashed = maximum norm, solid = L^1 norm) of spline approximation (compatible by projection) to discrete velocity solution. $\Delta x = 0.001$, 7 spline functions (equally spaced in $[-3, 3]$), smooth initial data. Degree 1 performs best over the range of interesting τ.

The aproximation features as shown earlier in Fig. 4.5 suggested that increasing the spline order from linear to quadratic or cubic is not really worth the effort. The same suggestion manifests itself in the PDE setting. Fig. 4.24 shows approximations of the discrete velocity density ρ with various spline degrees (for the case of smooth initial

data), in dependence on τ with 7 spline functions (equally centered between $\xi = -3$ and $\xi = 3$). The best approximation is given by the linear spline. During the computations, we could observe that choosing higher order spline functions can even lead to numerical instabilities. Note here again, that we are not interpolating a set of points, for which cubic splines would fulfill some optimality conditions (see [47]).

We can expect that an increase of polynomial degree would be more worthwhile in the smooth case than in the shock case. Strong bimodalities or even discontinuities, as they can appear in shock tube problems, will be resolved worse by higher polynomial degrees. In this sense it is sufficient to consider the smooth case, where higher order polynomials would have a chance to work better. From Fig. 4.24 we can thus conclude that higher degree splines should not be used.

Projected Splines

From now on, all splines will be of degree one.

We introduced the concept of projected splines ('PS') in Sect. 4.3.2. We will discuss the convergence of these with Fig. 4.25 and then consider the model error in Fig. 4.26 and Fig. 4.27.

For Fig. 4.25, we have chosen 13 perturbation functions to illustrate the convergence of the projected splines in Δx. The convergence for other numbers of perturbation functions looks very similar. We observe that also for splines, the relative discretization errors for the shock tube problem are higher than those for the smooth case. Whereas the difference is minor in the L^1 norm, it shows significantly in the maximum norm.

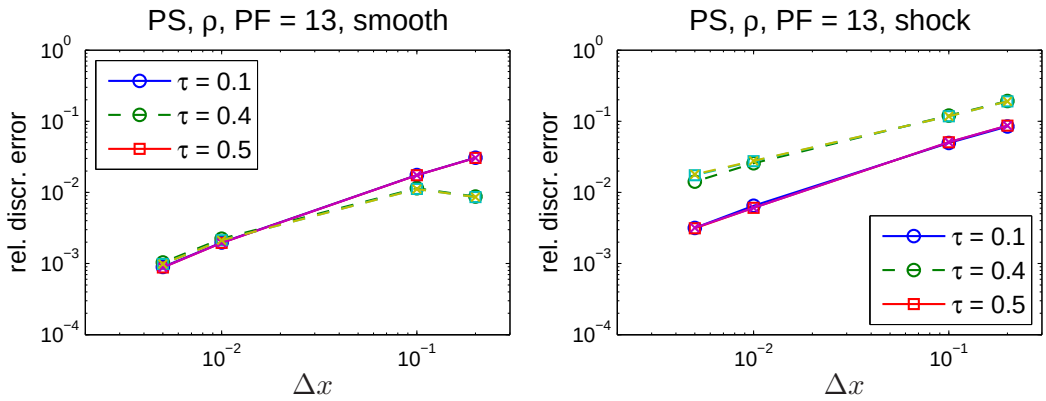


Figure 4.25: Rel. discretization errors (dashed = max.-norm, solid = L^1-norm) of smooth (left) and shock (right) solutions with projected splines (PS), 13 perturbation functions. $CFL = 0.9$, final time $T = 0.2$. Reference solution at $\Delta x = 0.001$.

Next, we consider the combination of modeling and discretization error (4.87). In Fig. 4.26 (smooth case) and Fig. 4.27 (shock tube problem), we compare various numbers of perturbation functions with the discrete velocity results. The perturbation functions are chosen in the same hierarchical way as in Sect. 4.3.2. We show the relative errors at $\Delta x = 0.005$ (left sides) and $\Delta x = 0.001$ (right sides) in the L^1-norm. The maximum-norm is omitted for the sake of legibility of the figures. In both, smooth and shock cases, we observe a stable, comparable picture for both space discretizations (for τ in the transition regime), indicating that the modeling error dominates over the discretization error in the compared schemes. Comparing the two figures, we observe that the model accuracies in the shock case are slightly lower than in the smooth case.

In the shock, as well as in the smooth case, an approximation with 13 perturbation functions is optimal in terms of accuracy. Between 4 and 7 functions there is a significant difference, 7 and 13 functions produce almost the same results. Most interestingly, 25 functions produce results of again lower accuracy. We think that this is due to the increasing complexity of the system of PDE's and corresponding non-linear effects in the errors.

Notice that for 4 perturbation functions in the smooth case in Fig. 4.26, we get an instability on the lowest discretization level $\Delta x = 0.001$, around $\tau = 0.6$. We observed the build up of this instability in the maximum-norm for τ values approaching $\tau = 0.6$. This instability could be flattened with the use of lower CFL numbers. It is surprising, that the approximation with 4 perturbation functions works at all, be it in the smooth or in the shock case, because the resolution of the non-equilibrium distribution space is rather poor. So it is not unexpected that we observe problems in regimes far away from equilibrium.

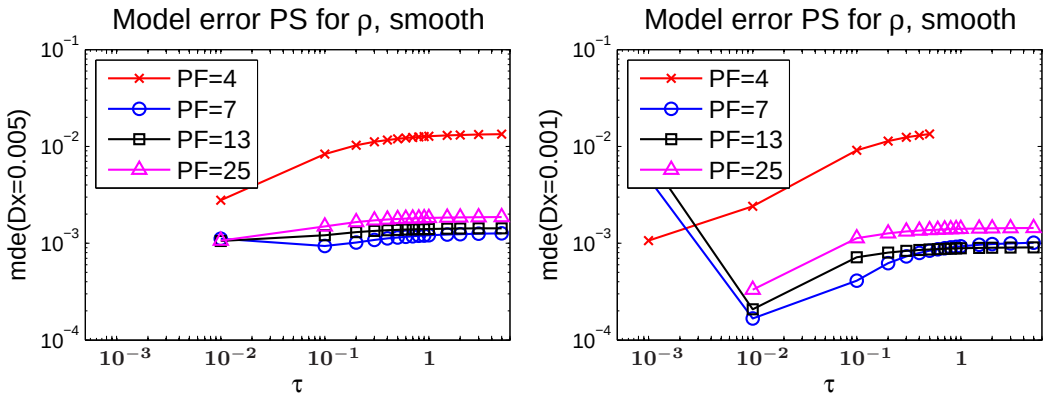


Figure 4.26: Rel. model errors (L^1 norm) between projected splines and discrete velocity scheme for smooth solutions at $\Delta x = 0.005$ and $\Delta x = 0.001$. $CFL = 0.9$, equally spaced spline centers in $[-3, 3]$.

One more noticeable problem occurs with the use of 7 perturbation functions in the

shock tube problem at the lowest scale $\Delta x = 0.001$: in Fig. 4.27 we observe a stability problem that is not resolvable with choosing lower CFL numbers. Such problems can occur in highly non-linear equations, as we have them, and are due to adverse non-linear effects on the discretization errors. We found a remedy to this problem through adjusting the spline spread from the interval $[-3, 3]$ to $[-2.4, 2.4]$ for the setting of 7 perturbation functions. Evidently, this means that we violate the hierachy of splines for this case.

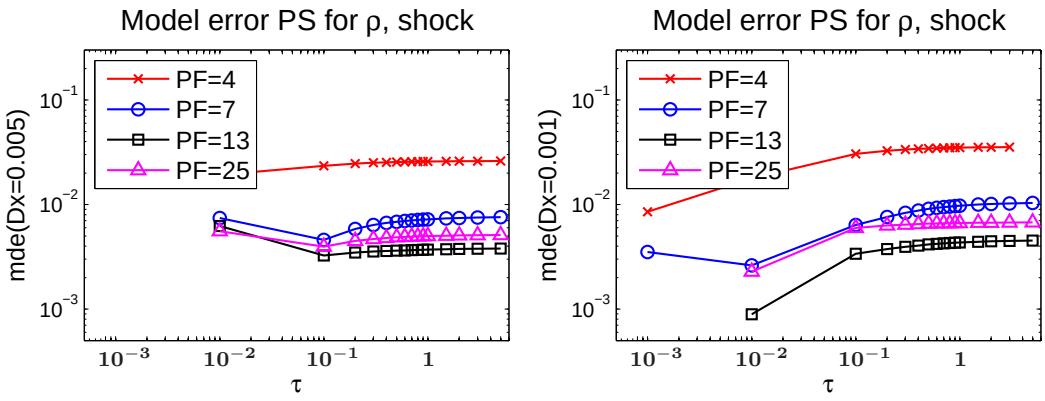


Figure 4.27: Rel. model errors (L^1 norm) between projected splines (PS) and discrete velocity scheme for the shcok tube problem at $\Delta x = 0.005$ and $\Delta x = 0.001$. $CFL = 0.9$, equally spaced spline centers in $[-3, 3]$ (4,13,25 PF), $[-2.4, 2.4]$ (7 PF).

Compatible Splines

In Sect. 4.3.2 we have seen how to impose the compatibility conditions (4.9) directly onto the perturbation functions, yielding compatible splines ('CS'). In that same section, we have examined the pure approximation properties of projected spline combinations and compatible splines. There were no significant differences. The same shows in the full PDE case. If we are using the same spline hierarchy as before, namely we put the spline centers equally spaced into the intervall $[-3, 3]$ with 7, 13 and 25 spline functions, we can form quartetts of splines, leading to total numbers of 4, 10 and 22 perturbation functions.[22]

In terms of discretization and model errors, the results are equivalent to the ones for projected splines. However here, the compatible splines have the disadvantage of being less flexibel for the interesting settings of only very few perturbation functions (we will always need at least 4 splines to form the quartetts).

[22]The case with 4 splines would yield only one quartett and is not considered here.

In terms of CPU time, the compatible splines are a bit more effective since the projection after each time step is not necessary.

We will not repeat the analysis already done for the projected splines, since the figures look the same. Also the stability problem with 7 projected splines, corresponding to 4 compatible spline quartetts occurs analogously. Some results of the compatible splines are shown below in Sect. 4.6.4, where we directly compare the modeling errors of splines and Hermite functions.

4.6.4 Direct Comparison of Hermite and Splines

Finally, we want to directly compare the performance of the various strategies discussed above. For this, we consider the most accurate approximations for projected splines ('PS'), compatible splines ('CS') and Hermite functions at $\Delta x = 0.001$ in the same settings as in the previous sections. We have seen that for the projected splines this optimal choice was 13 perturbation functions, with Hermite functions, it was also 13. For Hermite functions, we did not consider using more functions, since the CFL restriction is getting more and more severe, leading to an increase of CPU time. Splines did not exhibit this strong restrictions.

In this section, we will not only consider the approximation of ρ, but also of the heat flux q. The two figures, Fig. 4.28 and Fig. 4.29 show the same settings for ρ and q, once

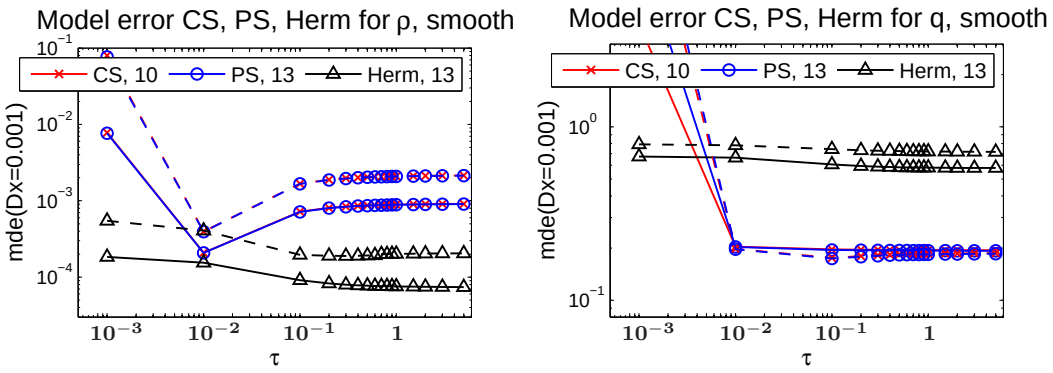


Figure 4.28: Rel. model errors in ρ and q (maximum, dashed and L^1, solid) between spline mixed scheme (projected and compatible), Hermite mixed scheme and discrete velocity scheme for smooth solutions at $\Delta x = 0.001$. $CFL = 0.9$ (splines, discrete velocity) and $CFL = 0.4$ (Hermite).

for the smooth initial conditions, and once for the shock tube problem. In both cases, it is evident that the heat flux is approximated with far less accuracy than ρ. This general fact has been observed several times before, see Sect. 4.5.2 and Sect. 3.9.2. The Hermite functions completely fail to capture the heat flux q in both, smooth and shock case. This

indicates that their global approach is much less adapted to capture higher moments of the distribution function, as we would have expected.

We also observe that indeed the results for the projected and the compatible splines do not differ, as was mentioned in the above section.

In the smooth case of Fig. 4.28, ρ (on the left) is best approximated with Hermite functions. They are one order of mangitude more accurate than the splines, both in L^1- and maximum-norm.

In the shock case, Fig. 4.28, as expected, splines outperform the Hermite approach in both, ρ and q. Note again that for stability reasons, the mixed scheme with Hermite functions is used at $CFL = 0.4$. Splines did not show stability problems depending on the CFL number, $CFL = 0.9$ is used in the plots. Therefore, the CPU time to obtain the spline solutions is considerably shorter than the one for the Hermite computations in this plot and the results of the splines are still more accurate, except for a small dip at $\tau = 0.1$.

Also in the shock case, the superior quality of the splines becomes even more evident if we consider the heat flux q. The accuracy of the spline solution is one order of magnitude better than the one of the Hermite functions.

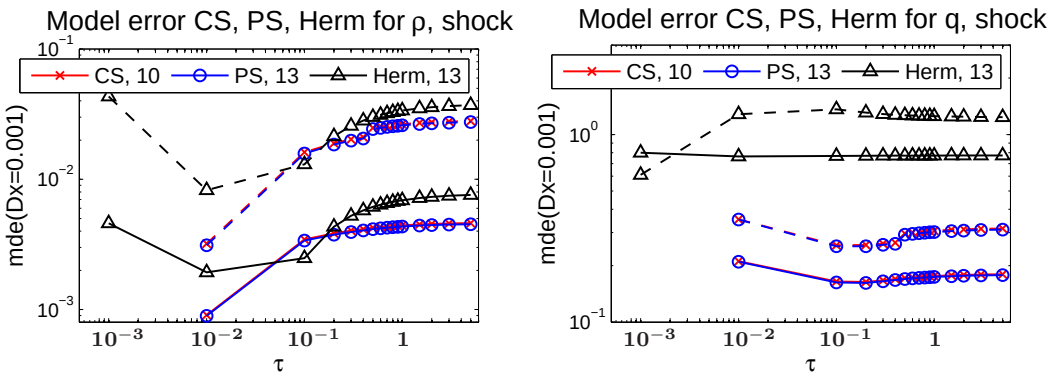


Figure 4.29: Rel. model errors in ρ and q (maximum, dashed and L^1, solid) between spline mixed scheme (projected and compatible), Hermite mixed scheme and discrete velocity scheme for the shock tube problem at $\Delta x = 0.001$. $CFL = 0.9$ (splines, discrete velocity) and $CFL = 0.4$ (Hermite).

We can conclude from the figures in this section that our mixed scheme with spline functions approximates the discrete velocity solutions with rather high accuracy. Given computational time that is several orders shorter than the time necessary fo discrete velocity schemes, our mixed schemes offer a significant improvement in efficiency.

From the analysis in this Section, we cannot make a final conclusion on some optimal choices for ϕ_α. We have seen that approximation quanlity is case dependent and that our approach works well if we choose the perturbation functions adequately.

4.7 Conclusion

We have derived a very general weak formulation of the Boltzmann equation. In this formulation, we have chosen combinations of functions to approximate equilibrium and non-equilibrium perturbations. For the resulting equations, we have developed a stable numerical scheme and shown convergence results for different physical problems.

Further research should aim in two main directions:

- Are there more optimal sets of perturbation functions than Hermite functions and splines?
- How can the current ideas be generalized to 2 and 3 dimensions?

The first point requires transdisciplinary experience in approximation theory, kinetic theory and numerical analysis. It will be challenging to construct such functions, not to mention to prove some kind of optimality theorems. We expect that some space- and velocity-adaptive choices of function hierarchies could be an interesting idea. Such a hierarchy would allow us to use more functions for parts of the distribution function with more details and save storage for parts that are very close to the equilibrium Gaussian. On the other hand, such constructions would involve space and time dependent perturbation functions ϕ_α, making the corresponding equations more complicated. Other possibilities could aim for different sets of (possibly local) orthognal polynomials. It could also be interesting to consider a non-linear ansatz with a different structure, as e.g. presented in [58]. There, the equilibrium Gaussian is replaced by a more flexible distribution function of the Pearson family.

Concerning the second point, an effective generalization of our numerical scheme to 2 and 3 dimensions will be highly relevant in competition to 'classical' (extended) moment equations and solvers for the full Boltzmann equation. It is promising that moment equations manage to capture high dimensional behaviour without using full tensorization of the one dimensional setting. In the one dimensional case, they manage to capture a significant part of the physical behaviour with extremely few parameters and this one dimensional advantage is strongly amplified for higher dimensions. A thorough understanding of the physical effects and symmetries allowing for this low number of parameters is a starting point to develop similarly effective generalizations to our scheme. Another very effective way of generalizing one dimensional problems to higher dimensions is tensorization with sparse grids. This approach is numerically motivated and hase been successfully used in [64] for (stationary) radiative transfer.

Even though our computations are one dimensional, they form a solid basis to further develop the perturbation function approach. An improvement of the ansatz for the approximation of the Boltzmann distribution functions would be more conceptionally interesting, whereas an extension to higher dimensions, including more complex geometries, could offer significant advantages over existing methods in the trasition regime.

5 Future Projects

We have seen several approximations to the Boltzmann equation. The case of linear collision operators was extensively studied in Part 3 and in Part 4, we have seen a computational approach. The results presented there lead to several open questions, generalizations and applications. We want to summarize some of them here.

5.1 Linear Collision Operators

One of the most important questions in the analysis of the linear collision operator concerns the generalization to (certain classes of) non-linear operators. Henning Struchtrup developed an 'order of magnitude' approach for non-linear moment equations (see [49]), a rigorous mathematical generalization, especially of convergence and stability results would underline the success of this method.

Another interesting mathematical question concerns the approximation order of the scale induced closure in the linear setting. We have seen conditions for Navier-Stokes-Fourier and also for Burnett order in Thm. 3.6.1. The conditions for Burnett order, (3.52) seem very technical. Could they be reformulated into 'more physical' statements? And how does the technicality translate into the increased setting, including higher non-equilibrium spaces (Sect. 3.8)? Are there intuitive answers to this question? And how can they be proven mathematically?

The approximation order is also related to 'regularizations', see Sect. 3.6.3. We have only shortly discussed this topic, due to many unclear points. Also here, a rigorous mathematical analysis would validate existing approaches (see [59]).

A practically very relevant question concerns the generalization of the scale induced closure to other differential equations. Stiff relaxation terms are occuring in various fields and are usually treated by means of asymptotic expansion (see e.g. [46]). Given the results in Sect. 3.9.2, applications to other fields are promising.

A next question that is interesting in relation to our closure is the construction of 'asymptotic preserving schemes'. The Boltzmann equation and Euler's equations are related through the limit $\tau \to 0$. It has been shown that numerical schemes for BGK at moderate τ fail if $\tau \to 0$. There are remedies to this problem, ensuring that the discrete scheme for the Boltzmann equation converges to a stable scheme for the Euler equations as $\tau \to 0$, see [34]. It would be very interesting to develop schemes that preserve the

asymptotics of the scale induced closure. A mathematical analysis, first in the toy model of a 16 discrete velocity model, could reveal more insights into the asymptotic behaviour of various numerical schemes for the Boltzmann-BGK equations.

Another, huge field of possibilities lies with hybrid schemes. There are very interesting approaches, where the computational domain is split into a DSMC part, a generalized BGK part and an equilibrium part (Euler equations), see [44]. There, computational cost could be saved by using higher order non-equilibirum approaches in between Euler and the BGK part. Whereas the switches between Euler, BGK and DSMC seem to work well, corresponding switches between higher order methods would have to be invented and tested.

5.2 Perturbation Function Approximation

We have implemented our weak formulation of the Boltzmann equation for splines of various orders and for Hermite functions. The most interesting question is, whether there are more optimal choices for the perturbation functions and whether optimality of some choices could be physically motivated or even mathematically proven.

Then, we would like to use the perturbation functions in practice, meaning in more than one dimensions. Whereas Grad's approach generalizes extremely efficiently (but also loses a lot of approximation quality), it remains to be seen how the spline perturbation functions have to be extended to model 2− and 3−dimensional velocity distributions. Simple tensor products are not the method of choice due to the decrease in computational efficiency. Ideally, some symmetries could be used to reduce the necessary amount of functions in higher dimensions, but maybe this will be very case dependent. A very effective way of reducing the dimensionality for tensor products is the use of sparse grids, see e.g. [64]. It is however not clear, whether this approach is also interesting for cases with only few functions per dimension.

Another interesting direction is the generalization of our method to more complicated collision kernels. In principle, it is straight forward to insert the ansatz for the distribution function into the corresponding kernel. Concretely, it will be advisable to 'adapt' the structure of our ansatz to the structure of the kernel, as we have done for the BGK kernel.

Of course, any mathematically provable statements about hyperbolicity, stability, or other features of the equations developed (or even just for the numerical schemes) would be very desirable. Ideally, the general form with any kind of perturbation functions could be analysed, realistically, this will only be possible for some specific choices of perturbation functions.

Hybrid schemes, as mentioned in the last paragraph of Sect. 5.1, can be direct competitors to our perturbation function approach, or they can be used together. Hybrid schemes

rely on switches between the parts of different discretizations, our method switches automatically by modeling higher order non-equilibrium with more and more non-zero coefficients for the perturbation functions, but of course this way, we do not reduce storage, as real hybrid schemes do. Also here, some combination may be possible, our method could e.g. be used as intermediate scheme between full DSMC and equilibrium Euler.

A Appendix

A.1 Notation

Vectorial quantities are typed in boldface $\mathbf{x}$ or in index notation, x_k. If there are several vectorial quantities, we use boldface variables with an index, $\mathbf{x}_i$. If not stated differently, x describes a scalar quantity.

We apply Einstein's sum convention, meaning that we sum over indices appearing twice in a product, $x_{ijk}a_{is}b_{jl} := \sum_{i,j} x_{ijk}a_{is}b_{jl}$. We also extend this to $x_i^2 := x_i x_i = \sum_i (x_i^2)$. Derivatives are denoted as $\partial_t := \frac{\partial}{\partial t}$, $\partial_k := \partial_{x_k}$, or as $\nabla := (\partial_1, ..., \partial_n)^t$. If there are variables involved that do not enter the derivative, we specify the derivatives, e.g. $\nabla_{\mathbf{x_1}}$. We will sometimes denote time derivatives with super dots, $\ddot{x}_i = \frac{\partial^2}{\partial t^2}$. Total derivatives are denoted as $\frac{d}{dt}$, and the convective derivative along a velocity field v as $D_t := \partial_t + v\partial_x$.

A.2 Appendix: Relations between Entropy, Hyperbolicity and Stability

Thm. 3.7.2 classifies a first order linear system of PDE's to be symmetric hyperbolic because there is a convex entropy with associated negative definite entropy flux. In this section we want to give a short overview of relations between hyperbolicity, entropy and stability.[1]

For this, consider the first order quasilinear system of partial differential equations

$$A(u)\partial_t u + B(u)\partial_x u = 0, \tag{A.1}$$

with $\mathbb{R} \times \mathbb{R}_0^+ \ni (x,t) \mapsto u(x,t) \in \mathbb{R}^m$ the (classical) solution and matrices $\mathbb{R}^m \ni u \mapsto A(u), B(u) \in \mathbb{R}^{m\times m}$.

A.2.1 Features of a Quasilinear System

Balance Law form

The system (A.1) is said to have a balance law form / conservative form, if it can be written as

$$\partial_t v + \partial_x F(v) = 0, \quad v = T(u) \tag{A.2}$$

with a flux function $F : \mathbb{R}^m \to \mathbb{R}^m$ and a variable transformation $T : \mathbb{R}^m \ni u \mapsto v \in \mathbb{R}^m$.

Hyperbolicity

The system (A.1) is said to be globally hyperbolic, if the generalized eigenvalue problem

$$\det\left(A(u) - \lambda B(u)\right) = 0 \tag{A.3}$$

has only (finite) real solutions λ for all $u \in \mathbb{R}^m$, and the corresponding eigenvectors form a basis of $\mathbb{R}^m$. Hyperbolicity is independent of possible variable transforms of (A.1).

The (generalized) eigenvalues λ are physical 'information' velocities (see [16], p. 51).

If for all $u \in \mathbb{R}^m$, A and B are symmetric matrices and A is positive definite, then the system is called symmetric hyperbolic, and A is called symmetrizer. The (generalized) eigenvalues in this case are automatically real (see [16], p. 51).

The system (A.1) is said to be locally hyperbolic, if the generalized eigenvalue problem has finite real solutions for *some* $u \in \mathbb{R}^m$.

[1] The author acknowledges valuable inputs from [39].

Convex Entropy

An entropy of the system (A.1) is a scalar, convex function $\mathbb{R}^m \ni u \mapsto \eta(u) \in \mathbb{R}$, that satisfies an equation

$$\partial_t \eta(u) + \partial_x h(u) = 0, \tag{A.4}$$

with an appropriate entropy flux $\mathbb{R}^m \ni u \mapsto h(u) \in \mathbb{R}$ (see [16]). For weak solutions, we generalize (A.5) to an entropy inequality,

$$\partial_t \eta(u) + \partial_x h(u) \leq 0. \tag{A.5}$$

Stability

Stability manifests itself in various (non-equivalent) definitions. Intuitively, a system is stable if for (possibly all) times t, a certain (space-)norm of the solution remains finite. There are various theorems for the L^1-norm of scalar Cauchy-problems (see [16]). A more physical understanding of stability is defined through associated quantities like energy or entropy that remain finite as time evolves, possibly translating to L^2-bounds of the solution,

$$||u(\cdot, t)||_{L^2(\mathbb{R}} < C \in \mathbb{R} \text{ for all } t \in \mathbb{R}_+. \tag{A.6}$$

A.2.2 Relations

The Godunov-Mock Theorem

The Godunov-Mock theorem ([20]) tells us that the existence of a convex entropy is equivalent to (global) symmetric hyperbolicity if our system is in conservation law form. In terms of Fig. A.1, this corresponds to the intersection of the convex entropy and symmetric hyperbolic set. In the general case of systems that are *not* in conservation law form, we may have convex entropies without the system being symmetric hyperbolic, or vice-versa.

Entropy and Stability

Thm. A.2.1 tells us that a convex entropy implies the existence of an L^2 bound for the solution.

Theorem A.2.1 (L^2 bound)**.**
Let u be a weak solution of (A.1) *and $\eta(u)$ a convex entropy with corresponding entropy flux $h(u)$, fulfilling* (A.5)*. Then, under very weak assumptions, $||u(\cdot, t)||_{L^2(\mathbb{R})}$ is bounded.*

Proof. The function $\bar{\eta}(u) := \eta(u) + \beta \cdot (u - u_0)$ is also a convex entropy for (A.1) with some corresponding entropy flux $\bar{h}(u)$. The parameter β is a vector in $\mathbb{R}^m$ (to be chosen later), and u_0 is a state close to u with $\left|\int \eta(u_0)dx\right| < \infty$. From (A.5), it follows with minor restrictions on u, namely integrability of $\eta(u(x,t))$ in x, that

$$\partial_t \int_{\mathbb{R}} \bar{\eta}(u(x,t))dx \leq C_1, \tag{A.7}$$

with $C_1 \in \mathbb{R}$ independent of t. With an elementary estimate, we can find $C_2(T)$ such that for all $t \in [0,T]$, $T \in \mathbb{R}$

$$\int \bar{\eta}(u(x,t))dx = \int \eta(u(x,t))dx + \beta \cdot \int (u(x,t) - u_0(x,t))dx \leq C_2. \tag{A.8}$$

Now, we do a Taylor expansion of η around u_0,

$$\eta(u) = \eta(u_0) + (u - u_0) \cdot \nabla\eta(u_0) + \frac{1}{2}(u - u_0)^T \cdot H\eta(\Theta) \cdot (u - u_0), \tag{A.9}$$

with the Gradient vector $\nabla\eta$ and the Hesse matrix $H\eta$. Θ is some intermediate state between u and u_0, its existence follows from the mean value theorem (see [32], p. 284). Since we assume η to be convex, the Hesse matrix $H\eta$ is positive definite. We can use the Taylor expansion (A.9) in (A.8) and obtain

$$\begin{aligned} C_2 &\geq \int \eta(u(x,t))dx + \beta \cdot \int (u(x,t) - u_0(x,t))dx \\ &= \int \eta(u_0(x,t))dx + \beta \cdot \int (u(x,t) - u_0(x,t))dx + \int \nabla\eta(u_0) \cdot (u - u_0)dx \\ &\quad + \frac{1}{2}\int (u - u_0)^T \cdot H\eta(\Theta) \cdot (u - u_0)dx \end{aligned} \tag{A.10}$$

Now, we choose the components of β such that they compensate $\int \nabla\eta(u_0) \cdot (u - u_0)dx$.

Since $H\eta$ is symmetric, it can be orthogonally diagonalized into $H\eta(\Theta) = T(\Theta)^T D(\Theta)T(\Theta)$. The eigenvalues are uniformly positive, and with

$$\lambda_{min} := \min_x \{\lambda \in \text{spectrum}(H\eta(\Theta))\}, \tag{A.11}$$

we obtain that

$$\begin{aligned} (u - u_0)^T \cdot H\eta(\Theta) \cdot (u - u_0) &= (u - u_0)^T \cdot T(\Theta)^T D(\Theta)T(\Theta) \cdot (u - u_0) \\ &\geq \lambda_{min}(u - u_0)^T(u - u_0). \end{aligned} \tag{A.12}$$

Alltogether, the estimate (A.10) becomes

$$C_2 - \int \eta(u_0(x,t))dx \geq \frac{1}{2}\int (u-u_0)^T \cdot H\eta(\Theta) \cdot (u-u_0)dx \geq \lambda_{min}\int (u-u_0)^T \cdot (u-u_0)dx. \tag{A.13}$$

Since $||u||_{L^2} = ||u - u_0 + u_0||_{L^2} \leq ||u - u_0||_{L^2} + ||u_0||_{L^2}$, we obtain

$$||u||_{L^2} \leq C_3, \tag{A.14}$$

with $C_3 = \frac{1}{\lambda_{min}} \left(C_2(T) - \min_{t\in[0,T]} \int \eta(u_0(x,t))dx \right) + ||u_0||_{L^2}$ □

Note that in this proof, we have not been using anything but the existence of a convex entropy (and some very weak conditions on the solutions[2]). Most noticeable is that we do not need any hyperbolicity or conservation law form for the L^2-bound.

In the case of a linear system of advection equations, we can directly obtain an L^2 bound, as done in the proof of Thm. 3.7.2.

Consider

$$\partial_t u + B\partial_x u = 0. \tag{A.15}$$

A convex entropy / entropy flux pair for this system is $\eta(u) := u^T u$ and $h(u) := u^T Bu$.[3] The entropy (in)equality, integrated in space, then reads

$$\partial_t ||u||_{L^2} \leq -\partial_x \int u^T Bu dx. \tag{A.16}$$

If B is symmetric positive definite, meaning that (A.15) is globally symmetric hyperbolic, the right hand side stays negative and we directly get an L^2-bound for u and all times $t \in [0, \infty]$.

In Thm. 3.7.2, we also have a term on the right hand side. In the proof there, we take this term into account, it yields a negative definite entropy production. For this case, Thm. A.2.1 can be directly generalized to the following

Corollary A.2.2 (L^2-bound with right hand side)**.**
Let u be a weak solution of the system

$$A(u)\partial_t u + B(u)\partial_x u = K(u, x, t).$$

If we can find a triplet of a convex entropy $\eta(u)$ with corresponding entropy flux $h(u)$ and a negative semidefinite entropy production P_η fulfilling[4]

$$\partial_t \eta(u) + \partial_x h(u) = P_\eta(u, x, t),$$

then, under the same very weak assumptions on u as in Thm. A.2.1, we get that $||u(\cdot, t)||_{L^2(\mathbb{R})}$ is bounded.

[2] It may be possible to find rather special counter examples based on these assumptions, but they will be practically irrelevant.
[3] If B is not symmetric, we choose $\frac{1}{2}(B + B^T)$ instead of B for the entropy flux.
[4] The only, but quite strong assumption on $K(u, x, t)$ is that we can indeed find such a triplet.

Proof. We can simply extend the constant C_1 in (A.7) by the entropy production. □

If the right hand side does not have any specific structure, we cannot conclude on stability from the existence of a convex entropy / entropy flux pair.

A.2.3 Summary

We summarize the connections between hyperbolicity, stability and entropy in Fig. A.1. A general quasilinear system of partial differential equations does not have any generic

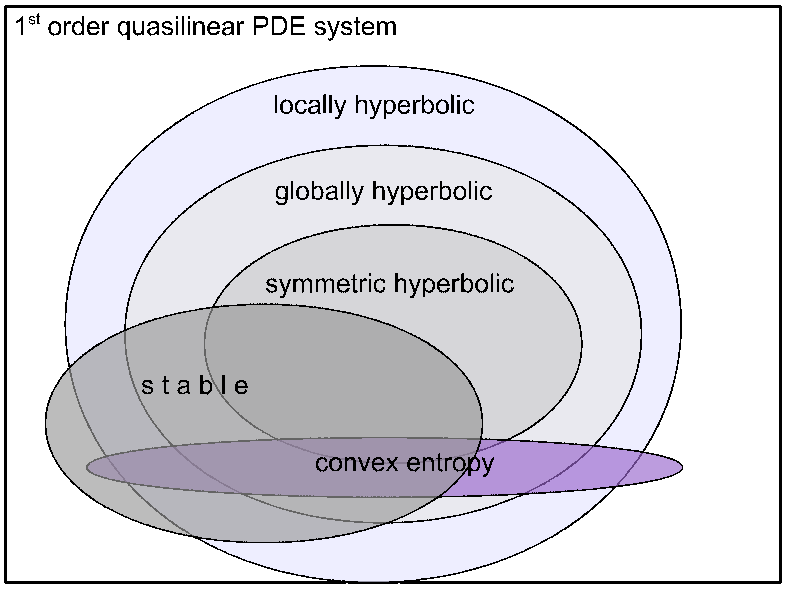


Figure A.1: Relations between hyperbolicity, entropy and stability.

properties. It can be locally hyperbolic (e.g. the Grad system) or globally hyperbolic (e.g. Euler for positive temperatures). We call a system symmetric hyperbolic, if it is globally hyperbolic and the flux matrix is symmetric. For symmetric hyperbolic equations that possess a balance law form, the Godunov-Mock theorem provides us with the existence of a convex entropy / entropy flux pair, and with that we get an L^2 bound. On the other hand, if we have the conservative form and a convex entropy / entropy flux pair, the same theorem yields symmetric hyperbolicity. If we do not have a conservative form, symmetric hyperbolicity does not necessaryly imply the existence of a convex entropy / entropy flux pair.

We can also imagine systems that are symmetric hyperbolic and have a convex entropy but are not stable due to production terms. In the linear case, the only sets that coincide in Fig. A.1 are those of local and global hyperbolicity.

There are many stable systems that are not hyperbolic at all, or that are hyperbolic but do not posses a conservation law form or a convex entropy / entropy flux pair. On the other hand, a system can also posses an entropy / entropy flux pair without being stable, e.g. if the right hand side is inappropriate.

Providing all sorts of counter examples to justify the non-overlaps in Fig. A.1 exceeds the frame of this work and is left to the enthusiastic reader.

A.3 Appendix: Proof of $E_1 = G^\dagger$, Sect. 3.5.3

We proove that $E_1 = G^\dagger$ given symmetry of $S = GE_1$ and conditions (3.31). Denote $N := dim(V) < \infty$.

A singular value decomposition of G yields

$$G = U \left(\begin{array}{c} \mathbf{\Sigma} \in \mathbb{R}^{q \times q} \\ \hline \mathbf{0} \in \mathbb{R}^{(N-q) \times q} \end{array} \right) W^*, \tag{A.17}$$

with $U \in \mathbb{R}^{N \times N}$ and $W \in \mathbb{R}^{q \times q}$ orthogonal, and $\Sigma \in \mathbb{R}^{q \times q}$ the diagonal matrix containing the non zero singular values of G. Any left inverse E_1 of G is of the form

$$E_1 = W \left(\begin{array}{c|c} \Sigma^{-1} & \mathbf{C} \in \mathbb{R}^{q \times (N-q)} \end{array} \right) U^*, \tag{A.18}$$

with $\mathbf{C}$ an arbitrary matrix. Now

$$GE_1 = U \left(\begin{array}{c|c} \mathbf{id} \in \mathbb{R}^{q \times q} & \mathbf{\Sigma} \cdot \mathbf{C} \\ \hline \mathbf{0} \cdot \mathbf{\Sigma}^{-1} & \mathbf{0} \cdot \mathbf{C} \end{array} \right) U^*. \tag{A.19}$$

This can only be symmetric if $\mathbf{C} = \mathbf{0}$, which yields the Moore-Penrose-inverse $G^\dagger$ in (A.18). For the case of a general Hilbertspace, we refer to Theorem 9.1.3 in [62].

A.4 Appendix: Details for the 16 Velocities Model

For reference we give the detailed expressions for the distribution functions and the moment operator for different closures of the 16-velocity model.

A.4.1 Matrices

The velocities are ordered by

$$\begin{array}{cccc} c_1 = (-3,3), & c_2 = (-1,3), & c_3 = (1,3), & c_4 = (3,3), \\ c_5 = (-3,1), & c_6 = (-1,1), & c_7 = (1,1), & c_8 = (3,1), \\ c_9 = (-3,-1), & c_{10} = (-1,-1), & c_{11} = (1,-1), & c_{12} = (3,-1), \\ c_{13} = (-3,-3), & c_{14} = (-1,-3), & c_{15} = (1,-3), & c_{16} = (3,-3). \end{array} \tag{A.20}$$

which gives

$$V = \mathrm{Diag}\,(-3,-1,1,3,-3,-1,1,3,-3,-1,1,3,-3,-1,1,3) \tag{A.21}$$

for the advection operator. The diagonal interactions are defined by

$$K^{\mathrm{diag}}[u] = -\begin{pmatrix} u_2u_5 - u_1u_6 \\ u_1u_6 + u_3u_6 - u_2u_5 - u_2u_7 \\ u_2u_7 + u_4u_7 - u_3u_6 - u_3u_8 \\ u_3u_8 - u_4u_7 \\ u_1u_6 + u_6u_9 - u_2u_5 - u_5u_{10} \\ u_2u_5 + u_2u_7 + u_5u_{10} + u_7u_{10} - u_1u_6 - u_3u_6 - u_6u_9 - u_6u_{11} \\ u_3u_6 + u_3u_8 + u_6u_{11} + u_8u_{11} - u_2u_7 - u_4u_7 - u_7u_{10} - u_7u_{12} \\ u_4u_7 + u_7u_{12} - u_3u_8 - u_8u_{11} \\ u_5u_{10} + u_{10}u_{13} - u_6u_9 - u_9u_{14} \\ u_6u_9 + u_6u_{11} + u_9u_{14} + u_{11}u_{14} - u_5u_{10} - u_7u_{10} - u_{10}u_{13} - u_{10}u_{15} \\ u_7u_{10} + u_7u_{12} + u_{10}u_{15} + u_{12}u_{15} - u_6u_{11} - u_8u_{11} - u_{11}u_{14} - u_{11}u_{16} \\ u_8u_{11} + u_{11}u_{16} - u_7u_{12} - u_{12}u_{15} \\ u_9u_{14} - u_{10}u_{13} \\ u_{10}u_{13} + u_{10}u_{15} - u_9u_{14} - u_{11}u_{14} \\ u_{11}u_{14} + u_{11}u_{16} - u_{10}u_{15} - u_{12}u_{15} \\ u_{12}u_{15} - u_{11}u_{16} \end{pmatrix} \tag{A.22}$$

and the straight interactions are taken to be

$$K^{\mathrm{straight}}[u] = -\begin{pmatrix} 0 \\ u_5u_7 - u_2u_{10} \\ u_6u_8 - u_3u_{11} \\ 0 \\ -(u_5u_7 - u_2u_{10}) \\ -(u_6u_8 - u_3u_{11}) - (u_6u_{14} - u_9u_{11}) \\ -(u_5u_7 - u_2u_{10}) - (u_7u_{15} - u_{10}u_{12}) \\ -(u_6u_8 - u_3u_{11}) \\ u_6u_{14} - u_9u_{11} \\ u_5u_7 - u_2u_{10} + u_7u_{15} - u_{10}u_{12} \\ u_6u_8 - u_3u_{11} - (u_9u_{11} - u_6u_{14}) \\ u_7u_{15} - u_{10}u_{12} \\ 0 \\ u_9u_{11} - u_6u_{14} \\ u_{10}u_{12} - u_7u_{15} \\ 0 \end{pmatrix} \tag{A.23}$$

Fig. 2.4 displays the interactions in the velocity grid. The linearized collision operator becomes

$$K = - \begin{pmatrix}
-1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -3 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -3 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & -3 & 2 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 2 & 0 & 0 & 2 & -6 & 2 & -1 & 0 & 2 & 1 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 2 & -1 & -1 & 2 & -6 & 2 & 0 & 1 & 2 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & 2 & -3 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -3 & 2 & -1 & 0 & 1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 2 & 1 & 0 & 2 & -6 & 2 & -1 & -1 & 2 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 1 & 2 & 0 & -1 & 2 & -6 & 2 & 0 & 0 & 2 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 2 & -3 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 2 & 0 & 0 & 1 & -3 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 2 & 0 & 0 & 1 & -3 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1
\end{pmatrix} \tag{A.24}$$

which is a symmetric positive definite matrix in $\mathbb{R}^{16\times 16}$.

A.4.2 Construction of the Operators for the Classical Closures:

The orthogonal complement of $\ker(K)$ is spanned by vectors r_1,...,r_{12}. The matrix M_0 consisting of equilibrium and r_1,...,r_{12} (see also (3.87)) gives an equivalent formulation of (3.86) in terms of moments

$$\partial_t M_0 f(x,t) + M_0 V M_0^{-1} \partial_x M_0 f + \frac{1}{\varepsilon} M_0 K M_0^{-1} M_0 f = 0. \tag{A.25}$$

The complete moment operator M_0 can be computed to be

$$M_0 = \left(\begin{array}{c} M_0^{(1)} \\ \hline M_0^{(2,1)} \;\big|\; id \end{array} \right) \in \mathbb{R}^{16\times 16} \tag{A.26}$$

with submatrices

$$M_0^{(1)} = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 \\
3 & 3 & 3 & 3 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -3 & -3 & -3 & -3 \\
18 & 10 & 10 & 18 & 10 & 2 & 2 & 10 & 10 & 2 & 2 & 10 & 18 & 10 & 10 & 18 \\
-1 & 3 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} \tag{A.27}$$

and

$$M_0^{(2,1)} = \begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & -1 \\
-1 & 0 & 3 & 2 & 1 & 2 & 5 & 6 & 5 & 6 & 9 \\
0 & -1 & -3 & -1 & -1 & -2 & -4 & -3 & -3 & -4 & -6 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & -1 & -1 & -2 & -2 & -2 & -2 & -3 & -3 & -3 & -3
\end{pmatrix}^T \tag{A.28}$$

and $id \in \mathbb{R}^{11\times 11}$. In the complete moment representation the production term then becomes

$$M_0 K M_0^{-1} = - \left(\begin{array}{c|cccccccccccc} \mathbf{0} \in \mathbb{R}^{4\times 4} & \multicolumn{12}{c}{\mathbf{0} \in \mathbb{R}^{4\times 12}} \\ \hline & -4 & 7 & -7 & 1 & 0 & -3 & 3 & 0 & 0 & 0 & 0 & 0 \\ & 0 & -11 & 3 & -1 & -1 & 3 & 1 & 0 & 0 & -1 & 0 & 0 \\ & -2 & -1 & -7 & 2 & -1 & 1 & 3 & 0 & 0 & 0 & -1 & 0 \\ & -2 & 3 & -3 & -3 & -1 & -3 & 3 & 1 & 0 & 0 & 0 & 0 \\ & -1 & 0 & 0 & 0 & -5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{0} \in \mathbb{R}^{12\times 4} & -1 & -1 & 1 & 0 & 0 & -7 & 3 & -1 & -1 & 2 & 0 & 0 \\ & -2 & 0 & 0 & 0 & -3 & 0 & -4 & 2 & 0 & 0 & 2 & -1 \\ & -4 & 6 & -6 & 1 & -2 & -6 & 6 & -3 & 0 & 0 & 0 & 1 \\ & -3 & 7 & -3 & 0 & -2 & -7 & 3 & 0 & -1 & 1 & 0 & 0 \\ & -3 & 3 & -3 & 0 & -3 & -3 & 3 & 0 & 1 & -3 & 1 & 0 \\ & -4 & 6 & -6 & 0 & -3 & -6 & 6 & 0 & 0 & 1 & -3 & 1 \\ & -6 & 13 & -9 & 0 & -3 & -9 & 5 & 1 & 0 & 0 & 1 & -1 \end{array}\right) \tag{A.29}$$

We now are using M_0 to construct the operators for the various closures.. The equilibrium distribution $M\rho$ is parametrised by the four equilibrium moments only

$$M\rho = M_0^{-1} \left(\begin{array}{c|c} id \in \mathbb{R}^{4\times 4} & \mathbf{0} \in \mathbb{R}^{4\times 12} \end{array}\right)^T \rho = \left(\tfrac{1}{80}\tilde{M}\right)^T \rho \tag{A.30}$$

with

$$\tilde{M} = \begin{pmatrix} -\frac{3}{2} & 5 & 5 & -\frac{3}{2} & 5 & \frac{7}{2} & \frac{7}{2} & 5 & 5 & \frac{7}{2} & \frac{7}{2} & 5 & -\frac{3}{2} & 5 & 5 & -\frac{3}{2} \\ -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 \\ 3 & 3 & 3 & 3 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -3 & -3 & -3 & -3 \\ \frac{5}{4} & 0 & 0 & \frac{5}{4} & 0 & -\frac{5}{4} & -\frac{5}{4} & 0 & 0 & -\frac{5}{4} & -\frac{5}{4} & 0 & \frac{5}{4} & 0 & 0 & \frac{5}{4} \end{pmatrix} \tag{A.31}$$

Correspondingly, we construct the equilibrium operator as

$$E_0 f = \left(\begin{array}{c|c} id \in \mathbb{R}^{4\times 4} & \mathbf{0} \in \mathbb{R}^{4\times 12} \end{array}\right) M_0 f, \tag{A.32}$$

This operator turns out to be

$$E_0 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 & -3 & -1 & 1 & 3 \\ 3 & 3 & 3 & 3 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -3 & -3 & -3 & -3 \\ 18 & 10 & 10 & 18 & 10 & 2 & 2 & 10 & 10 & 2 & 2 & 10 & 18 & 10 & 10 & 18 \end{pmatrix}. \tag{A.33}$$

Grad

Arbitrary Moments

For the higher moments in Grad's closure, we arbitrarily chose $\mu_1 = E_1 r_1$, $\mu_2 = E_1 r_2$ and $\mu_3 = E_1 r_3$, with the again arbitrary choices of G and E_1 as

$$G = M_0^{-1} \left(\begin{array}{c|c|c} \mathbf{0} \in \mathbb{R}^{3\times 4} & id \in \mathbb{R}^{3\times 3} & \mathbf{0} \in \mathbb{R}^{3\times 9} \end{array}\right)^T = M_0^{-1} \left(\tfrac{1}{80}\tilde{G}\right)^T \tag{A.34}$$

with

$$\tilde{G} = \begin{pmatrix} -15 & -11 & -17 & 47 & 1 & 5 & -1 & -17 & 7 & 11 & 5 & -11 & 3 & 7 & 1 & -15 \\ -1 & -9 & -7 & 5 & -9 & 63 & -15 & -3 & -7 & -15 & -13 & -1 & 5 & -3 & -1 & 11 \\ 5 & -7 & -9 & -1 & -3 & -15 & 63 & -9 & -1 & -13 & -15 & -7 & 11 & -1 & -3 & 5 \end{pmatrix} \tag{A.35}$$

and

$$E_1 = \left(\mathbf{0} \in \mathbb{R}^{3\times 4} \;\middle|\; id \in \mathbb{R}^{3\times 3} \;\middle|\; \mathbf{0} \in \mathbb{R}^{3\times 9} \right) M_0 \tag{A.36}$$

$$= \begin{pmatrix} -1 & 3 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{A.37}$$

As mentioned in 3.9.2, these operators fullfill the requirements (3.26), but not necessarily (3.33).

Kinetic Fluxes as Moments

The additional moments can be directly computed as

$$\begin{aligned} q_x &= (\text{-54} \quad \text{-10} \quad 10 \quad 54 \quad \text{-30} \quad \text{-2} \quad 2 \quad 30 \quad \text{-30} \quad \text{-2} \quad 2 \quad 30 \quad \text{-54} \quad \text{-10} \quad 10 \quad 54) \\ q_y &= (54 \quad 30 \quad 30 \quad 54 \quad 10 \quad 2 \quad 2 \quad 10 \quad \text{-10} \quad \text{-2} \quad \text{-2} \quad \text{-10} \quad \text{-54} \quad \text{-30} \quad \text{-30} \quad \text{-54}) \\ \sigma_{xy} &= (\text{-9} \quad \text{-3} \quad 3 \quad 9 \quad \text{-3} \quad \text{-1} \quad 1 \quad 3 \quad 3 \quad 1 \quad \text{-1} \quad \text{-3} \quad 9 \quad 3 \quad \text{-3} \quad \text{-9}) \end{aligned}$$

To fullfill the conditions (3.26), we project these moment vectors to the non-equilibrium phase space by applying $P = (id - ME_0)$. The resulting vectors are then normalized and form the lines of E_1.

Correspondingly we chose $G = E_1^T$, and construct the equations according to (3.27) and (3.28).

A.4.3 Direct Asymptotic Expansion

The conditions for 2nd order in Thm. 3.6.1 are met in the case of the 16 discrete velocities model.
Nontheless, in this section we show how to directly do the asymptotic expansion of (3.50/3.51) in ε.

Let us abbreviate (3.50) and (3.51) as

$$\partial_t \begin{pmatrix} \rho \\ \mu \end{pmatrix} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \rho \\ \mu \end{pmatrix} + \frac{1}{\varepsilon} \begin{pmatrix} 0 & 0 \\ 0 & E \end{pmatrix} \begin{pmatrix} \rho \\ \mu \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{A.38}$$

with $A = E_0\mathbf{c}\cdot\nabla M$, $B = E_0\mathbf{c}\cdot\nabla G$, $C = E_1\mathbf{c}\cdot\nabla M$, $D = E_1\mathbf{c}\cdot\nabla G$,
$E = E_1 K G$.

Inserting the expansion $\mu = \varepsilon\mu_1 + \varepsilon^2\mu_2$ into (A.38) yields

$$\begin{aligned} \partial_t\rho + A\rho + B\left(\varepsilon\mu_1 + \varepsilon^2\mu_2\right) &= 0 \\ \partial_t\left(\varepsilon\mu_1 + \varepsilon^2\mu_2\right) + C\rho + D\left(\varepsilon\mu_1 + \varepsilon^2\mu_2\right) + \frac{1}{\varepsilon}E\left(\varepsilon\mu_1 + \varepsilon^2\mu_2\right) &= 0 \end{aligned} \tag{A.39}$$

and short calculations reveal that

$$\begin{aligned} \mu_1 &= -E^{-1}C\rho, \\ \mu_2 &= -E^{-1}D\mu_1 - E^{-1}\partial_t\mu_1 = E^{-1}DE^{-1}C\rho + E^{-1}E^{-1}C\partial_t\rho \\ &\overset{\text{Euler}}{=} E^{-1}DE^{-1}C\rho - E^{-1}E^{-1}CA\rho. \end{aligned} \tag{A.40}$$

Plugging this into (A.39) yields

$$\partial_t\rho + A\rho = \varepsilon BE^{-1}C\rho + \varepsilon^2 B\left(-E^{-1}DE^{-1}C\rho + E^{-1}E^{-1}CA\rho\right), \tag{A.41}$$

or, by using the definitions of A,...,E:

$$\begin{aligned} \partial_t\rho + E_0\mathbf{c}\cdot\nabla M\rho = \quad & \varepsilon E_0\mathbf{c}\cdot\nabla G\left(E_1KG\right)^{-1}E_1\mathbf{c}\cdot\nabla M\rho \\ & - \varepsilon^2 E_0\mathbf{c}\cdot\nabla G\left(E_1KG\right)^{-1}E_1\mathbf{c}\cdot\nabla G\left(E_1KG\right)^{-1}E_1\mathbf{c}\cdot\nabla M\rho \\ & + \varepsilon^2 E_0\mathbf{c}\cdot\nabla G\left(E_1KG\right)^{-1}\left(E_1KG\right)^{-1}E_1\mathbf{c}\cdot\nabla ME_0\mathbf{c}\cdot\nabla M\rho. \end{aligned} \tag{A.42}$$

This compares to the asymptotic expansion of the original kinetic equations, as given in (3.24)

$$\begin{aligned} \partial_t\rho + E_0\mathbf{c}\cdot\nabla M\rho = & -\varepsilon E_0(\mathbf{c}\cdot\nabla)K^\dagger(\mathbf{c}\cdot\nabla)M\rho \\ & - \varepsilon^2 E_0(\mathbf{c}\cdot\nabla)K^\dagger(\mathbf{c}\cdot\nabla)K^\dagger(\mathbf{c}\cdot\nabla)M\rho \\ & + \varepsilon^2 E_0(\mathbf{c}\cdot\nabla)K^\dagger K^\dagger(\mathbf{c}\nabla M)E_0(\mathbf{c}\cdot\nabla)M\rho. \end{aligned} \tag{A.43}$$

Using now the special structure of the 16-discrete velocities model (A.20), we can compute the coefficient matrices for first and second order, and get:

$$\begin{aligned} \partial_t\rho + E_0VM\partial_x\rho = \quad & \varepsilon E_0VG\left(E_1KG\right)^{-1}E_1VM\partial_x^2\rho \\ & - \varepsilon^2 E_0VG\left(E_1KG\right)^{-1}E_1VG\left(E_1KG\right)^{-1}E_1VM\partial_x^3\rho \\ & + \varepsilon^2 E_0VG\left(E_1KG\right)^{-1}\left(E_1KG\right)^{-1}E_1VME_0VM\partial_x^3\rho \end{aligned} \tag{A.44}$$

Computing the products shows equivalence with the Chapman-Enskog expansion of the original kinetic equations. Furthermore one can compute that this equivalence breaks down for third order (super-Burnett).

A.5 Details for the Derivation of Grad's 5-Moment-System

The matrices (4.14) with Hermite functions read

$$\begin{aligned}
M_{\mu\nu} &= \langle \psi_\mu, \phi_\nu \rangle = \begin{pmatrix} \sqrt{6} & 0 \\ 0 & 2\sqrt{6} \end{pmatrix}, \quad M^1_{\mu\nu} = (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi \phi_\nu \rangle = \begin{pmatrix} 0 & 2 \\ 5 & 0 \end{pmatrix}, \\
M^2_{\mu\nu} &= (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^2 \phi_\nu \rangle = \begin{pmatrix} 10 & 0 \\ 0 & 15 \end{pmatrix}, \quad M^3_{\mu\nu} = (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^3 \phi_\nu \rangle = \begin{pmatrix} 0 & 30 \\ \frac{105}{2} & 0 \end{pmatrix}, \\
D^0_{\mu\nu} &= (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \partial_\xi \, \phi_\nu \rangle = \begin{pmatrix} 0 & 2 \\ 3 & 0 \end{pmatrix}, \quad D^1_{\mu\nu} = (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi \, \partial_\xi \, \phi_\nu \rangle = \begin{pmatrix} 6 & 0 \\ 0 & 10 \end{pmatrix}, \\
D^2_{\mu\nu} &= (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^2 \, \partial_\xi \, \phi_\nu \rangle = \begin{pmatrix} 0 & 20 \\ \frac{45}{2} & 0 \end{pmatrix}, \\
V^0_\mu &= (M^{-1})_{\mu\lambda} \langle \psi_\lambda, 1 \rangle = \begin{pmatrix} 0 \\ \sqrt{\frac{3}{8}} \end{pmatrix}, \quad V^1_\nu = (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi \rangle = \begin{pmatrix} \sqrt{\frac{3}{2}} \\ 0 \end{pmatrix}, \\
V^2_\mu &= (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^2 \rangle = \begin{pmatrix} 0 \\ 5\sqrt{\frac{3}{8}} \end{pmatrix}, \quad V^3_\mu = (M^{-1})_{\mu\lambda} \langle \psi_\lambda, \xi^3 \rangle = \begin{pmatrix} 5\sqrt{\frac{3}{2}} \\ 0 \end{pmatrix}, \\
Q_\mu &= \frac{1}{\sqrt{2\pi}} \int \xi^3 e^{-\xi^2/2} \phi_\mu d\xi = \begin{pmatrix} \sqrt{6} \\ 0 \end{pmatrix}.
\end{aligned} \tag{A.45}$$

With these matrices, we can compute the matrixfunctions as defined in (4.23)

$$\begin{aligned}
B_{\alpha\gamma}(U, \kappa) &= v\delta_{\alpha\gamma} + \sqrt{\theta} M^1_{\alpha\gamma} - \frac{\sqrt{\theta}}{2} \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\mu M^2_{\alpha\mu} - \kappa_\mu D^1_{\alpha\mu} \right) Q_\gamma \\
&= \begin{pmatrix} v - 3\sqrt{\frac{3}{2}}\sqrt{\theta}\kappa_1 & 2\sqrt{\theta} \\ 5\sqrt{\theta} - \sqrt{\frac{3}{2}}\sqrt{\theta} \left(\sqrt{6} + 4\kappa_2 \right) & v \end{pmatrix}
\end{aligned} \tag{A.46}$$

and

$$\begin{aligned}
C_{\alpha 1} &= \left\{ \kappa_\gamma D^0_{\alpha\gamma} - \frac{1}{2} Q_\gamma \kappa_\gamma \left(-V^0_\alpha - \kappa_\alpha + V^2_\alpha + \kappa_\gamma M^2_{\alpha\gamma} - \kappa_\gamma D^1_{\alpha\gamma} \right) \right\} \frac{\sqrt{\theta}}{\rho} \\
&= \frac{\sqrt{\theta}}{\rho} \begin{pmatrix} -3\sqrt{\frac{3}{2}}\kappa_1^2 + 2\kappa_2 \\ 3\kappa_1 - \sqrt{\frac{3}{2}}\kappa_1 \left(\sqrt{6} + 4\kappa_2 \right) \end{pmatrix},
\end{aligned} \tag{A.47}$$

as well as

$$C_{\alpha 2} = 0_\alpha = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{A.48}$$

and

$$C_{\alpha 3} = \left\{ \frac{V_\alpha^3}{2} + \kappa_\gamma \frac{M_{\alpha\gamma}^3}{2} - \kappa_\gamma \frac{D_{\alpha\gamma}^2}{2} - \frac{3}{2} V_\alpha^1 - \frac{3}{2} \kappa_\gamma M_{\alpha\gamma}^1 + \kappa_\gamma D_{\alpha\gamma}^0 \right.$$
$$\left. - \frac{3}{4} Q_\gamma \kappa_\gamma \left(-V_\alpha^0 - \kappa_\alpha + V_\alpha^2 + \kappa_\gamma M_{\alpha\gamma}^2 - \kappa_\gamma D_{\alpha\gamma}^1 \right) \right\} \frac{1}{\sqrt{\theta}} \tag{A.49}$$
$$= \frac{1}{\sqrt{\theta}} \begin{pmatrix} \sqrt{\frac{3}{2}} - \frac{9}{2}\sqrt{\frac{3}{2}} \kappa_1^2 + 4\kappa_2 \\ \frac{21}{2}\kappa_1 - \frac{3}{2}\sqrt{\frac{3}{2}} \kappa_1 \left(\sqrt{6} + 4\kappa_2 \right) \end{pmatrix}.$$

The production vector reads

$$R_\alpha(U, \kappa) \overset{BGK}{=} -\frac{1}{\tau} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix}. \tag{A.50}$$

A.6 Photonic Crystals

Together with Kersten Schmidt, the author of this thesis was working on finite element computations for photonic crystals. Since this work is not strictly related to the research projects for this thesis, we refer to the published paper [45] and give a short summary of that project. It is worth mentioning that there are further projects with publications (see [9]) going on, building substantially on the results obtained in [45].

Short Summary: Computation of the band structure of two-dimensional Photonic Crystals with hp-Finite Elements

Photonic crystals are periodic materials with discontinuous dielectrical properties. Their band structure and corresponding eigenmodes can be efficiently computed with the finite element method (FEM). For second order elliptic boundary value problems with piecewise analytic coefficients it is known that the solution converges extremly fast, i.e. exponentially, when using p-FEM for smooth and hp-FEM for polygonal interfaces and boundaries. In our project, we discretised the variational eigenvalue problems for photonic crystals with smooth and polygonal interfaces in scalar variables. We used quasi-periodic boundary conditions by means of p- and hp-FEM for the transverse electric (TE) and transverse magnetic (TM) modes. Our computations showed exponential convergence of the numerical eigenvalues in settings of smooth and polygonal lines of discontinuity for the dielectric material properties.

References

[1] K. Aoki, P. Degond, S. Takata, and H. Yoshida. *Diffusion Models for Knudsen Compressors*. Phys. Fluids, 19(11)(117103), 2007.

[2] H. Babovsky. *Die Boltzmann-Gleichung*. Leitfäden der angewandten Mathematik und Mechanik. Teubner, 1998.

[3] H. Babovsky. *A Numerical Model for the Boltzmann Equation with Applications to Micro Flows*. Comput. Math. Appl., 58:791–804, 2009.

[4] D. Benedetto, E. Caglioti, and M. Pulvirenti. *A one Dimensional Boltzmann Equation with Inelastic Collisions*. Milan Journal of Mathematics, 67(1):169–179, 1997.

[5] P. L. Bhatnagar, E. P. Gross, and M. Krook. *A Model for Collision Processes in Gases. I. Small Amplitude Processes in Charged and Neutral One-Component Systems*. Phys. Rev., 94(3):511–525, 1954.

[6] G. A. Bird. *Molecular Gas Dynamics and the Direct Simulation of Gas Glows*, volume 12 of *Oxford Engineering Science Series*. Oxford University Press, New York, 1994.

[7] A. V. Bobylev. *Instabilities in the Chapman-Enskog Expansion and Hyperbolic Burnett Equations*. J. Stat. Phys., 124(2-4):371–399, 2006.

[8] A.V. Bobylev. *The Chapman-Enskog and Grad Methods for Solving the Boltzmann Equation*. Sov. Phys. Dokl., 27:29–31, 1982.

[9] H. Brandsmeier, K. Schmidt, and Ch. Schwab. *A Multiscale hp-FEM for 2D Photonic Crystal Bands*. J. Comp. Phys., 230:349–374, 2011.

[10] J. E. Broadwell. *Study of Rarefied Shear Flow by the Discrete Velocity Method*. J. Fluid Mech., 19:401–414, 1964.

[11] A. Caiazzo, M. Junk, and M. Rheinländer. *Comparison of Expansion Methods*. Comput. Math. Appl., 58:883–897, 2009.

[12] C. Cercignani. *The Boltzmann Equation and its Applications*, volume 67 of *Applied Mathematical Sciences*. Springer, New York, 1988.

[13] C. Cercignani and Sir R. Penrose. *Ludwig Boltzmann: The Man Who Trusted Atoms*. Oxford University Press, Oxford, 1998.

[14] S. Chapman and T. G. Cowling. *The Mathematical Theory of Non-Uniform Gases*. Cambridge University Press, Cambridge, 1970.

[15] N. Crouseilles, P. Degond, and M. Lemou. *A Hybrid Kinetic-Fluid Model for Solving the Vlasov-BGK Equation.* J. Comput. Phys., 203(2):572–601, 2005.

[16] C. M. Dafermos. *Hyperbolic Conservation Laws in Continuum Physics.* Springer, Berlin, 2nd edition, 2005.

[17] B. Düring, P Markowich, J.-F. Pietschmann, and M.-Th. Wolfram. *Boltzmann and Fokker-Planck Equations Modelling Opinion Formation in the Presence of Strong Leaders.* Proc. R. Soc. A, 465(2112):3687–3708, 2009.

[18] U. S. Fjordholm and S. Mishra. *Accurate Numerical Discretizations of Non-Conservative Hyperbolic Systems.* Research Report 2010-25, Seminar for Applied Mathematics, ETH Zürich, 2010.

[19] R. W. Freund and R. H. W. Hoppe. *Stoer/Bulirsch: Numerische Mathematik 1.* Springer, 10th edition, 2007.

[20] E. Godlewsky and P.A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, volume 118 of Applied Mathematical Sciences. Springer, Berlin, 1996.

[21] S. K. Godunov. *A Difference Scheme for Numerical Solution of Discontinuous Solution of Hydrodynamic Equations.* Math. Sbornik, 47:271 – 306, translated by US Joint Publ. Res. Service, JPRS 7226, 1969, original in Russian, 1959.

[22] H. Grad. *On the Kinetic Theory of Rarefied Gases.* Comm. Pure Appl. Math., 2:331–407, 1949.

[23] H. Grad. *Principles of the Kinetic Theory of Gases*, volume 12 of Handbuch der Physik. Springer, Berlin, 1958.

[24] G. M. Graf. Allgemeine Mechanik, Lecture Notes, 2002.

[25] G. M. Graf. Theorie der Wärme, Lecture Notes, 2005.

[26] X.-J. Gu and D. Emerson. *A Computational Strategy for the Regularized 13 Moment Equations with Enhanced Wall-Boundary Conditions.* J. Comput. Phys., 225:263–283, 2007.

[27] A. Harten, P. D. Lax, and B. Van Leer. *On Upstream Differencing and Godunov-Type Schemes for Hyperbolic Conservation Laws.* SIAM Review, 25(1):35–61, 1983.

[28] R. M. Iverson. *The Physics of Debris Flows.* Rev. Geophys., 35(3):245–296, 1997.

[29] S. Jin and M. Slemrod. *Regularization of the Burnett Equations via Relaxation.* J. Stat. Phys., 103 (5-6):1009–1033, 2001.

[30] P. Kauf, M. Torrilhon, and M. Junk. *Scale-Induced Closure for Approximations of Kinetic Equations.* J. Stat. Phys., 141(5):848–888, 2010.

[31] V.I. Kolobov, R. R. Arslanbekov, V. V. Aristov, A. A. Frolova, and S. A. Zabelok. *Unified Solver For Rarefied and Continuum Flows with Adaptive Mesh and Algorithm Refinement.* JCP, 223(2):589–608, 2007.

[32] K. Königsberger. *Analysis 1.* Springer, Heidelberg, 5th edition, 2001.

[33] O. E. Lanford III. *Time Evolution of Large Classical Systems*, volume 38 of Lecture Notes in Physics. Springer, 1975.

[34] M. Lemou and L. Mieussens. *A new Asymptotic Preserving Scheme Based on Micro-Macro Formulation for Linear Kinetic Equations in the Diffusion Limit.* 31(1):334–368, 2008.

[35] R. J. LeVeque. *Finite-Volume Methods for Hyperbolic Problems.* Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.

[36] C. D. Levermore. *Moment Closure Hierarchies for Kinetic Theories.* J. Stat. Phys., 83(5-6):1021–1065, 1996.

[37] G. N. Markelov, A. V. Kashkovsky, and Ivanov M. S. *Aerodynamics of Space Station 'Mir' During Aeroassisted Controlled Descent.* AIP Conference Proceedings of Rarefied Gas Dynamics, 585:745–752, 2001.

[38] L. Mieussens. *Discrete-Velocity Models and Numerical Schemes for the Boltzmann-BGK Equation in Plane and Axisymmetric Geometries.* J. Comp. Phys., 162(2):429–466, 2000.

[39] S. Mishra. Private communication.

[40] I. Müller and T. Ruggeri. *Rational Extended Thermodynamics*, volume 37 of Springer Tracts in Natural Philosophy. Springer, New York, 1998.

[41] Sir I. Newton. *Philosophia Naturalis Principia Mathematica.* books.google.com, London, 1686.

[42] K. Nipp and D. Stoffer. *Lineare Algebra.* vdf Hochschulverlag, 5$^{\text{th}}$ edition, 2002.

[43] M. Planck. *Vorlesungen über Thermodynamik.* Walter de Gruyter and Co., Berlin, Leipzig, 7$^{\text{th}}$ edition, 1922.

[44] G. Puppo et al. To be submitted.

[45] K. Schmidt and P. Kauf. *Computation of the Band Structure of Two-Dimensional Photonic Crystals with hp Finite Elements.* Comp. Meth. App. Mech. Engr., 198:1249 – 1259, 2009.

[46] Kersten Schmidt and Sébastien Tordeux. *Asymptotic Modelling of Conductive Thin Sheets.* Z. Angew. Math. Phys., 61(4):603–626, 2010.

[47] H. R. Schwarz. *Numerische Mathematik.* B. G. Teubner, Stuttgart, 4$^{\text{th}}$ edition, 1997.

[48] G. Strang. *On the Construction and Comparison of Difference Schemes.* SIAM Journal on Numerical Analysis, 5(3):506 – 517, 1968.

[49] H. Struchtrup. *Stable Transport Equations for Rarefied Gases at High Orders in the Knudsen Number.* Phys. Fluids, 16(11), 2004.

[50] H. Struchtrup. *Derivation of 13 Moment Equations for Rarefied Gas Flow to Second Order Accuracy for Arbitrary Interaction Potentials.* Multiscale Model. Simul., 3(1):221–243, 2005.

[51] H. Struchtrup. *Macroscopic Transport Equations for Rarefied Gas Flows.* Interaction of Mechanics and Mathematics. Springer, Heidelberg, 2005.

[52] H. Struchtrup and M. Torrilhon. *Regularization of Grad's 13-Moment-Equations: Derivation and Linear Analysis.* Phys. Fluids, 15/9:2668–2680, 2003.

[53] H. Struchtrup and M. Torrilhon. *H-theorem, Regularization, and Boundary Conditions for Linearized 13 Moment Equations.* Phys. Rev. Letters, 99(014502), 2007.

[54] P. Taheri, M. Torrilhon, and H. Struchtrup. *Couette and Poiseuille Microflows: Analytical Solutions for Regularized 13-Moment Equations.* Phys. Fluids, 21(017102), 2009.

[55] R. Temam and A. Miranville. *Mathematical Modeling in Continuum Mechanics*, volume 38 of Interaction of Mechanics and Mathematics. Cambridge University Press, Cambridge, 2005.

[56] M. Torrilhon. *Regularized 13-Moment-Equations.* Proceedings of 5th Intl. Symposium on Rarefied Gas Dynamics, St. Petersburg, Russia, 2006.

[57] M. Torrilhon. *Two-Dimensional Bulk Microflow Simulations Based on Regularized 13-Moment-Equations.* SIAM Multiscale Model.Simul., 5(3):695–728, 2006.

[58] M. Torrilhon. *Hyperbolic Moment Equations in Kinetic Gas Theory Based on Multi-Variate Pearson-IV-Distributions.* Comm. Comput. Phys., 7(4):639–673, 2010.

[59] M. Torrilhon and H. Struchtrup. *Regularized 13-Moment-Equations: Shock Structure Calculations and Comparison to Burnett Models.* J. Fluid Mech., 513:171–198, 2004.

[60] M. Torrilhon and H. Struchtrup. *Boundary Conditions for Regularized 13-Moment-Equations for Micro-Channel-Flows.* J. Comput. Phys., 227 (3):1982–2011, 2008.

[61] C. Wagner, C. Hoffmann, R. Sollacher, J. Wagenhuber, and B. Schürmann. *Second-Order Continuum Traffic Flow Model.* Phys. Rev. E, 54 (5):5073–5085, 1996.

[62] G. Wang, Y. Mei, and S. Giao. *Generalized Inverses: Theory and Computations*, volume 5 of Graduate Series in Mathematics. Science Press, Beijing/New York, 2004.

[63] E. T. Whittaker and G. N. Watson. *A Course of Modern Analysis.* Cambridge University Press, London, 4^{th} edition, 1962.

[64] G. Widmer. *An Efficient Sparse Finite Element Solver for the Radiative Transfer Equation.* Journal of Heat Transfer-Transactions of The Asme, 132, 2010.

Curriculum Vitae

Personal details

Name	Peter Kauf
Date of birth	December 27, 1981
Place of birth	Frauenfeld, Switzerland
Citizenship	Switzerland

Education

10/2006–08/2011	PhD studies in Applied Mathematics at ETH Zürich Zurich, Switzerland
09/2006	Diploma in Mathematics at ETH Zürich
10/2001–09/2006	Studies in Mathematics at ETH Zürich Zurich, Switzerland
02/2001–10/2001	Military Education (Fourier) Bern, Switzerland
02/2001	Matura, Typi A and B at Kantonsschule Frauenfeld
08/1994–02/2001	Gymnasium, Kantonsschule Frauenfeld Frauenfeld, Switzerland