Analysis of Lattice-Boltzmann Methods

Asymptotic and Numeric Investigation of a Singularly Perturbed System

Martin Kilian Rheinländer

Zur Verleihung des akademischen Grades
Doktor der Naturwissenschaften
(*Doctor rerum naturalium, Dr. rer. nat.*)
vom Fachbereich Mathematik und Statistik
der Universität Konstanz genehmigte
Dissertation.

Datum der Präsentation: 08.05.07
Datum des Rigorosums: 25.07.07

Referent: **Prof. Dr. Michael Junk**, Universität Konstanz
Koreferent: **Prof. Dr. Axel Klar**, TU Kaiserslautern

Sanctifica mihi omne primogenitum.
Primitias frugum terrae tuae deferes in domum Domini Dei tui.

Vulgata, liber exodi (13,2 & 23,19)

Preface – Vorwort

Der Anstoß zu meinem Dissertationsvorhaben entstammt dem BMBF-Projekt "Adaptive Gittersteuerung für Lattice-Boltzmann Verfahren zur Simulation von Füllprozessen im Gießereibereich" am *Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM)* in Kaiserslautern. Im Laufe der Projektarbeit ergaben sich grundlegende Fragen, von denen einige Gegenstand der vorliegenden Dissertationsschrift geworden sind, welche einen Großteil meiner Arbeit der Jahre 2003 bis 2006 zusammenfaßt.

Mein besonderer Dank ergeht an Herrn Prof. Dr. Michael Junk für seine wertvolle und wohlwollende Unterstützung. Sein beständiges Interesse sowie die Freiheit, welche er mir gewährt hat, um eigenen Problemstellungen nachzugehen, haben diese Arbeit entscheidend mitermöglicht. Gerne werde ich mich an die vielen Gespräche mit ihm während meiner Assistentenzeit an der Universität Konstanz zurückerinnern: sei es im Allgemeinen, z.B. bei der Planung interessanter Lehrveranstaltungen, oder im Speziellen, wenn der Motor mal wieder ein Tröpfchen Öl benötigte, um nicht zu stottern.
Desweiteren danke ich Herrn Prof. Dr. Axel Klar für die äußerst zügige Übernahme des Koreferats, ohne die der enge Zeitplan des "Endspurts" nicht hätte eingehalten werden können.

Mit dem Vollzug der Promotion endet ein Lebensabschnitt, der für mich in Kaiserslautern begann und über ein kurzes Intermezzo in Saarbrücken nach Konstanz an den Bodensee führte. Den vielen netten Menschen, die mich auf diesem Wege fachlich oder privat begleitet haben, möchte ich ebenfalls meinen Dank aussprechen. Dabei denke ich an etliche Mitarbeiter und Mitdoktoranden vom ITWM (insbesondere aus den Abteilungen SKS und HPC). Außerdem sind vor allem meine Freunde Bernd Büchler, Joshua Karnes und Michael Hawlitzky sowie mein Bruder Peter zu nennen. Die gemeinsamen, teils abenteuerlichen Unternehmungen erscheinen wie Inseln im Meer eines arbeitsreichen Alltags.
Den Pastoren Dietrich Schindler und Jeff Ingram danke ich für die herzliche Aufnahme in die *Freie Evangelische Gemeinde (FEG)* Kaiserslautern. Von den vielen lieben Bekannten dort sei insbesondere Familie Markutzik nebst Hulda erwähnt – die zahlreichen gemütlichen Diskussionsabende sind unvergessen.

Aufgrund meiner Lehraufgaben im Fachbereich Mathematik und Statistik der Universität Konstanz war die Arbeit seit Oktober 2004 sehr intensiv. Für die angenehme Arbeitsatmosphäre sei dem Fachbereich und insbesondere den Kollegen der Arbeitsgruppe Numerik gedankt. Unsere konversationsreiche Mittagstischrunde habe ich des öfteren wie das "Luftholen beim Streckentauchen" empfunden.

Nicht zuletzt bin ich meiner Verlobten und meinen Eltern zu tiefem Dank verpflichtet; Vita für ihre Geduld aus der Nähe, Mutter und Vater für die Einfühlsamkeit aus der Ferne und allen dreien für die notwendige moralische Ermutigung.

Konstanz, den 30. Juli 2007 Martin K. Rheinländer

Abstract

Lattice-Boltzmann algorithms represent a quite novel class of numerical schemes, which are used to solve evolutionary partial differential equations (PDE). In contrast to finite difference and finite element schemes, lattice-Boltzmann methods rely on a mesoscopic (kinetic) approach. The essential idea consists in setting up an artificial, grid-based particle dynamics, which is chosen such that appropriate averages provide approximate solutions of a certain PDE, typically in the area of fluid dynamics. As lattice-Boltzmann schemes are closely related to finite velocity Boltzmann equations being singularly perturbed by special scalings, their consistency is not obvious, however.

This work is concerned with the analysis of lattice-Boltzmann methods, where a particular interest lies in some numeric phenomena like initial layers, multiple time scales and boundary layers. As major analytic tool, regular expansions (Hilbert expansion) are employed to establish consistency. Exemplarily, two and three population algorithms are studied in one space dimension, mostly discretizing the advection-diffusion equation. It is shown how these 'model schemes' can be derived from two dimensional schemes in the case of special symmetries.

The analysis of the schemes is preceded by an examination of the singular limit being characteristic of the corresponding scaled finite velocity Boltzmann equations. Convergence proofs are obtained using a Fourier series approach and alternatively a general regular expansion combined with an energy estimate. The appearance of initial layers is investigated by multiscale and irregular expansions. Among others, a hierarchy of equations is found which gives insight into the internal coupling of the initial layer and the regular part of the solution.

Next, the consistency of the model algorithms is considered followed by a discussion of stability. Apart from proving stability for several cases entailing convergence as byproduct, the spectrum of the evolution operator is examined in detail. Based on this, it is shown that the CFL-condition is necessary and sufficient for stability in the case of a two population algorithm discretizing the advection equation. Furthermore, the presentation touches upon the question whether reliable stability statements can be obtained by rather formal arguments.

To gather experience and prepare future work, numeric boundary layers are analyzed in the context of a finite difference discretization for the one-dimensional Poisson equation.

Sommaire

Les méthodes de Boltzmann sur reseau répresentent une classe de schémas numériques assez nouveaux, qui ont été développés pour la résolution des équations aux dérivées partielles (EDP) du type évolutionnaire. Contrairement aux différences finies et aux éléments finis, les méthodes de Boltzmann sur réseau s'inspirent d'une approche mésoscopique (cinétique). L'idée fondamentale consiste à établir une dynamique artificielle pour des particules imaginaires mouvant sur les liens d'un réseau. Les paramètres de cette dynamique peuvent être choisis de telle sorte que des quantités moyennées convergent vers la solution d'une EDP dont l'origine est

notamment la mécanique des fluides. Toutefois, la consistance des schémas de Boltzmann sur réseau n'est pas évidente, puisqu'ils sont étroitement associés aux équations de Boltzmann aux vélocités finies devenant singulièrement perturbées si le nombre de Knudsen est petit.
Cet ouvrage est consacré à l'analyse des méthodes de Boltzmann sur réseau. Un intérêt particulier est accordé aux phénomènes numériques comme les couches initiales, les multi-échelles temporaires et les couches de bord. Des expansions régulières (expansions de Hilbert) sont utilisées comme outil principal pour vérifier la consistance. À titre d'exemple, des algorithmes à deux ou trois populations en dimension une de l'espace sont étudiés. D'abord, ces algorithmes de modèle – surtout ceux résolvant l'équation d'advection-diffusion – sont dérivés comme cas spéciaux des algorithmes bidimensionnels en exploitant certaines symétries.
L'analyse des schémas est précédée par l'examen de la limite singulière étant caractéristique des équations de Boltzmann aux vélocités finies. En employant des séries de Fourier ou alternativement une expansion régulière générale en combinaison avec une majoration de l'énergie, des preuves de convergence sont donnés. L'apparition des couches initiales est étudiée à l'aide d'expansions irrégulières et multi-échelles. Entre autres, on trouve une hiérarchie des équations qui donne des informations sur le couplage interne entre la couche initiale et la partie regulière de la solution.
Après, la consistance des algorithmes modèles est considérée, suivie par une discussion de la stabilité. En plus de démontrer la stabilité dans plusieurs cas (ce qui entraine aussi la convergence), le spectre de l'opérateur d'évolution discrète est examiné en détail. Basé sur ces résultats il est montré que la condition de CFL garantit la stabilité dans le cas d'un algorithme à deux populations discrétisant l'équation d'advection. En outre, la discussion aborde la question de savoir jusqu'à quel point il est possible d'obtenir des connaissances fiables sur le comportement de stabilité par des arguments seulement fondés sur l'analyse formelle.
Enfin, pour élargir l'expérience et pour préparer le travail futur, un exemple d'une couche de bord est analysé dans le contexte d'une discrétisation de différences finies pour l'équation de Poisson unidimensionnelle.

Zusammenfassung

Gitter-Boltzmann Methoden stellen eine verhältnismäßig neue Klasse numerischer Verfahren dar zur Lösung evolutionsartiger partieller Differentialgleichungen. Im Gegensatz zu Standardmethoden aus dem Bereich der finiten Differenzen bzw. finiten Elemente realisieren Gitter-Boltzmann Verfahren einen mesoskopischen (kinetischen) Ansatz. Die Kernidee besteht darin, eine gitterbasierten Pseudo-Teilchendynamik zu formulieren. Es stellt sich dabei heraus, daß gewisse gemittelte Größen die Lösungen bestimmter Differentialgleichungen approximieren, welche vor allem einem strömungsmechanischen Hintergrund entstammen. Allerdings ist die Konsistenz der Gitter-Boltzmann Verfahren keineswegs offensichtlich, nicht zuletzt weil sie in enger Beziehung zu singulär skalierten Boltzmanngleichungen mit endlichen Geschwindigkeitsmodellen stehen.
Diese Arbeit beschäftigt sich mit der Analyse von Gitter-Boltzmann Verfahren.

Besonderes Augenmerk gilt dabei einigen “numerischen Phänomenen” wie dem Auftreten von Anfangsschichten, der Existenz mehrerer Zeitskalen und dem Zustandekommen von Randschichten. Beim Konsistenznachweis dienen reguläre asymptotische Entwicklungen (Hilbert Entwicklungen) als zentrales Hilfsmittel. Beispielhaft werden Gitter-Boltzmann Algorithmen in einer Raumdimension mit zwei und drei Populationen untersucht. Dabei wird zunächst gezeigt, wie sich diese Modellalgorithmen zur Diskretisierung der Advektions-Diffusions Gleichung aus zweidimensionalen Algorithmen unter Ausnutzung spezieller Symmetrieeigenschaften ergeben.
Der Analyse der eigentlichen Schemata vorangestellt ist eine Untersuchung des singulären Grenzwerts bei einer Boltzmanngleichungen mit zwei bzw. drei Geschwindigkeiten. Alternativ lassen sich hier Konvergenzbeweise mittels einer Fourier-Entwicklung bzw. einer allgemeinen regulären Entwicklung kombiniert mit einer Energieabschätzung erzielen. Anfangsschichten werden mittels irregulärer Entwicklungen bzw. Multiskalen-Entwicklungen aufgelöst. Unter anderem stößt man dabei auf eine Hierarchie von Gleichungen, welche Aufschluß über die interne Kopplung der Anfangsschicht mit dem regulären Teil der Lösung geben.
Anschließend wird die Konsistenz der Modellalgorithmen betrachtet, gefolgt von einer Stabilitätsanalyse. Neben etlichen Stabilitätsbeweisen (woraus Konvergenz der jeweiligen Verfahren gefolgert werden kann) wird das Spektrum des diskreten Evolutionsoperators einer genauen Untersuchung unterzogen. Darauf aufbauend läßt sich zeigen, daß sich die CFL-Bedingung sowie Stabilität im Falle eines Zwei-Populationen Algorithmus für die Advektionsgleichung gegenseitig bedingen. Außerdem wird die Möglichkeit erörtert, inwieweit verläßliche Stabilitätsaussagen auch anhand einer formalen Analyse gewonnen werden können.
Um Erfahrung mit numerischen Randschichten für zukünftige Untersuchungen zu sammeln, wird abschließend eine finite Differenzen Diskretisierung für die eindimensionale Poisson Gleichung betrachtet, welche eine Randschicht erzeugt.

Contents

Background and Outline

General framework and context

While computers were rapidly evolving since the sixties of the last century, scientists became aware of the opportunity to employ these powerful machines to simulate physical processes. Often, simulations are based on the numerical and hence approximate solution of certain differential equations, which form the core of a simplifying *mathematical model*, abstracting the concrete process in a precise mathematical language.
In particular, the simulation of fluid flows (CFD[1]) represents a vast field of mathematically and numerically challenging problems whose mastering is of great practical interest for engineering and environmental sciences.

The underlying model equations used in CFD are taken from *continuum mechanics* and *thermodynamics*, see e.g. [18, 11]. By formulating the physical conservation principles for mass, momentum and energy in terms of familiar macroscopic quantities like pressure, flow velocity or temperature, the fundamental equations of *Euler* and *Navier-Stokes* are found.
Since, in general, one cannot directly[2] solve differential equations arising in mathematical models, the equations must undergo some *discretization*. This procedure breaks differential equations down into discrete equations being finally solvable by computers. Nowadays, various kinds of discretization methods are at disposal, among which *finite differences* represent the historically oldest approach. Their intuitive idea consists in considering the solution (e.g. the velocity or pressure field) only in a finite number of time and space points. In contrast, *finite element* and *spectral methods* try to approximate the complicated solution in terms of comparatively simple functions, that are defined in the entire space domain but can be characterized by only a few numbers.

The fundamental equations, that are generally accepted to govern fluid motion, were deduced in the 18^{th} and 19^{th} century when only little was known about the internal structure of matter. Their derivation is based on the *continuum hypothesis*[3], which

[1] Computational Fluid Dynamics

[2] To describe the solutions completely, infinitely (even uncountably) many numbers would be necessary which even a supercomputer cannot keep in its large but limited memory.

[3] The continuum hypothesis is a trick to deal with very large numbers of particles: for example, a large set of coupled oscillators is more easily described by the wave equation than by a huge ODE system. Even in statistical mechanics the phase space density is introduced by this reason, which, however, might be alternatively interpreted as probability density.

considers matter as a continuum, filling up a continuous space and moving in a continuously elapsing time. Even today, the continuum hypothesis has not lost its importance. Numerous differential equations of mathematical physics, whose validity has been well justified by experiments, rely on this basic assumption.
Nevertheless, we now think of matter as composed of tiny, microscopic particles called *molecules* and *atoms*. So nature seems ultimately to be discrete even if these particles exist in overwhelming large numbers. By the end of the 19^{th} century this discovery led to the development of *statistical mechanics* dealing with the behavior of many particle systems. This new physical discipline unsealed another understanding of fluid motion from a quite different perspective. By means of the Boltzmann equation it was shown how conservation laws on the scale of *microscopic* particles give rise to *macroscopic* conservation principles expressed by the Euler and Navier-Stokes equation.

In the early days of computer simulations, some scientists[4] started to study also purely *discrete dynamical systems*. These *cellular automata* (see [64, 65, 66] and [21]) were devised as some sort of counterpart to the discretization of *continuous dynamical systems* described by differential equations. Although the first cellular automata seem like mathematical gadgets, there was a serious objective behind. The goal was to set up an artificial particle dynamics where the interaction of the particles (collisions) obeys simple rules while the system as a whole displays a rather complex behavior revealing also macroscopic regularities.
In 1973 a *lattice-gas cellular automaton* was proposed by Hardy, Pomeau and de Pazzis [22, 23] that was able to produce flow like patterns in two space dimensions. However, it was shown that the results disagreed with the Navier-Stokes equation taken as reference whose accuracy for standard flow regimes is recognized. Only in 1986 Frisch, Hasslacher and Pomeau [20, 19] succeeded by considering a modified lattice-gas automaton which uses a hexagonal instead of a quadratic velocity model.

Now the development began to accelerate. In 1988 McNamara and Zanetti [47] replaced the boolean variables representing pseudo-particle numbers by real-valued quantities interpreted as particle distribution functions. During the following years, the basic lattice-Boltzmann algorithm received the form being still in use today [27], including the introduction of the BGK collision operator and the equilibrium function [51]. So, the essential steps were done which emancipated lattice-Boltzmann methods from their lattice-gas precursors, that were suffering from statistical noise because of the averaging to compute macroscopic quantities. In particular, relaxation type collision operators (BGK, MRT) do no more define explicit collision rules.
Formally, the classical lattice-Boltzmann method has much in common with a special finite difference scheme discretizing the BGK Boltzmann equation of statistical mechanics [24, 26]. Here the circle closes: departing from a general but blurred principle[5] and an artificial, rather unphysical particle dynamics, one had somehow

[4] Among them were mathematicians like von Neumann, Ulam and the computer pioneer Zuse.
[5] A *microscopically* simple dynamics can generate complex and well organized structures on a *macroscopic* scale.

returned to the "real physics" described by the Boltzmann equation.

Much progress took place since the early nineties when the method came up. Today, lattice-Boltzmann methods compete with traditional methods of CFD and they are implemented in computational software tools for industrial purposes[6] with applications to a large range of flow problems[7].
In comparison with other methods, the coding of lattice-Boltzmann schemes is less complex; moreover the code is well suited for parallelization to gain a speed-up of the computational time. These advantages are certainly owed to the *bottom-up* conception, which exploits the simplicity of the dynamics on the microscopic scale. In contrast, traditional methods are designed in *top-down* manner departing from complicated macroscopic models which have to be 'simplified' then by discretization procedures.

In opposition to numeric methods like finite elements, being well established in mathematical research, lattice-Boltzmann methods are mainly developed by physicists and physical engineers. Since the theoretical foundations were consequently embedded in statistical mechanics, lattice-Boltzmann methods became especially appealing for researchers with physical background. This accounts for the origin of certain analytical techniques still widely adopted. Most notably we mention the *Chapman-Enskog expansion* with two time scales (see e.g. [25, 63]), which is employed as major tool to relate lattice-Boltzmann schemes with macroscopic equations like the Navier-Stokes equation. For the classical approach to lattice-Boltzmann methods we refer also to [14, 3, 9, 10, 30].

Although a fundamental understanding of lattice-Boltzmann methods has been reached utilizing these techniques, the explanations often are not satisfying and lack mathematical rigor. To overcome the mysticism occasionally attributed to lattice-Boltzmann methods, it is necessary to interpret them from different sides.
So [46] advocates to give up a too 'narrow viewpoint' (which classifies lattice-Boltzmann methods just as derivatives of lattice-gas automata) in favor of a broader understanding. The paper emphasizes the affinity between lattice-Boltzmann methods and the kinetic theory of Boltzmann-type equations based on discrete velocity models. The latter has become a field of intensive research (see for instance [49]).
Still there is not much literature available dealing with convergence, consistency and stability of lattice-Boltzmann methods. A first approach is found in [16] where the convergence of certain lattice-Boltzmann schemes with respect to a nonlinear diffusion equation and the viscous Burgers equation is shown. However, the considered collision operator admits an H-theorem which cannot be proven for collision operators of relaxation type and polynomial equilibria [68, 69] as used today.
To avoid the relatively cumbersome Chapman-Enskog expansion, other possibilities have been introduced to connect a given lattice-Boltzmann algorithm with its macroscopic target equation (consistency analysis). In particular we mention

[6] For instance: the *ParPac* code at the Fraunhofer ITWM Kaiserslautern, or PowerFLOW commercialized by Exa corporation and based on "digital physics".

[7] Thermal flow, non-Newtonian fluids, flow through porous media, shallow water, acoustics, multiphase and turbulent flows. For references consult [63].

[33, 32], where the standard D2P9 lattice-Boltzmann algorithm was uncovered as a sophisticated finite difference discretization of the Navier-Stokes equation. Meanwhile, another approach has been presented in [34], which turns out to be especially clear and elegant, stimulating further developments of lattice-Boltzmann methods. Similar approaches are also found in [28, 15]. Furthermore it is tried[8] to interpret lattice-Boltzmann algorithms in the framework of finite volume schemes.

There are only few articles concerned with stability of lattice-Boltzmann schemes. Besides some publications examing stability and accuracy (e.g. [52]) by numerical tests, we mention [59] and [43] where a von Neumann stability analysis is essentially performed. Rigorous mathematical analysis is done in [1, 40].

Specific motivation of the thesis

The idea to this dissertation began to form, when I was working on grid coupling[9] and local grid refinement for lattice-Boltzmann methods. Being a novice in this field, I gradually became aware, that lattice-Boltzmann methods were mostly based on plausibility arguments being not proven in the mathematical sense. However, despite of being founded on plausible reasoning, some of my ideas did not work out in a fully satisfactory way. Because of these reasons the wish consolidated to acquire a deeper comprehension of lattice-Boltzmann methods. In particular, certain numerical phenomena like initial layers aroused my interest.
Often, plausibility arguments tend to be superficial thereby risking to neglect crucial details. Hence numerical side effects, that might spoil the accuracy or even excite unstable modes of the scheme, may easily be overlooked or underestimated with respect to their impact.

The main objective of this thesis is to promote a better understanding of lattice-Boltzmann methods, where some attention is given to disturbing numerical effects. Frequently, theoretic investigations are faced with the difficulty that they should stick to problems of practical relevance on the one hand, while requiring a simplified setup to be analytically feasible on the other hand.
Here the focus is on the *principal* comprehension which is to be obtained by a *detailed* study of several selected model problems, illuminating aspects also arising in more general situations. Thanks to their reduced complexity, the model problems permit a clear analysis and thereby offer much insight into the intrinsic workings of lattice-Boltzmann methods. Nevertheless, to bridge the gap mentioned above at least partly, it is shown how the one-dimensional model schemes may be derived as special cases from two-dimensional schemes.

Although lattice-Boltzmann methods are sometimes publicized as being fundamentally different from traditional numerical methods, they should be considered as just another discretization for numerically solving certain partial differential equations.

[8] ICMMES conference 2006, Hampton (VA). See also [62] for the one-dimensional case.

[9] Confer [54]. Observe, that the derivation of moments concerning the D2P9 algorithm for Stokes flow simulations as announced in this paper, is not included in this thesis to avoid further growing of its extent.

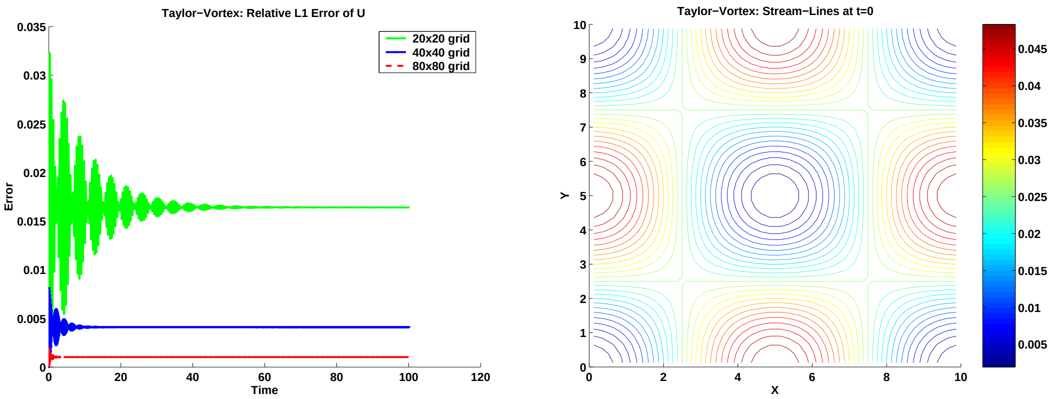


Figure 1: **Phenomenological description of an initial layer.** The curves in the left diagram represent the numerical error of a Stokes flow simulation performed with the D2P9 lattice-Boltzmann algorithm. More precisely, the curves indicate the relative $\mathcal{L}^1$-error in the x-component of the flow velocity plotted versus the time. The simulations were executed on three different grids. Two observations are striking: First, the error seems to be quartered if the grid spacing h is halved, which suggests that the error is of magnitude $\mathrm{O}(h^2)$. Second, the error oscillates at the beginning, where the amplitude (attenuation) is the smaller (stronger), the finer the grid is chosen. These features are typical for an initial layer combining two time scales here. The discrete time scale is manifested by the damping and the oscillations from time step to time step, being hardly visible due to the low resolution of the figure. In contrast, the beat-bellies occur in the fast time scale, which is slower than the discrete time scale but faster than the time scale referring to the labels of the horizontal axis. Considering the time evolution of an arbitrary single population in a fixed node reveals similar oscillations. This indicates that the initial layer affects all populations in roughly the same manner and does not represent an integral phenomenon only appearing in the $\mathcal{L}^1$-norm. The depicted initial layer was observed by simulating a Taylor vortex whose streamlines are given at right. Taylor vortices actually correspond to the eigenmodes of the Stokes operator in a periodic, rectangular domain. Therefore the streamlines remain invariant in space while the stream function decreases exponentially in time. Hence the error should drop likewise. Here, however, this decrease is imperceptible over the simulated time interval because of the relatively small viscosity that was chosen to sustain the initial layer excited by a crude initialization.

However, two particular features must be taken into account:

- The most striking distinction is that the physical quantities of the macroscopic limit equations do not figure as primary variables (populations) in the lattice-Boltzmann algorithms. Besides, there are more primary variables than relevant physical quantities.

- Viewing lattice-Boltzmann algorithms in the light of standard finite difference schemes, they look much like the natural discretization of Boltzmann-type equations so that generally no direct relation with the actual limit equation is suggested.

Roughly speaking, a numerical scheme is called *consistent* with respect to a differential equation if the solution of the differential equation satisfies the discrete equations in every grid node up to a small residual. Due to the first item, the standard notion of consistency does not apply to lattice-Boltzmann schemes. However, extending this notion, one is enabled to examine the consistency of a given lattice-Boltzmann method by means of *regular expansions*.
The resulting analysis shows why the primary variables yield approximate solutions of the limit equation. Additionally, it is found that they encode further information concerning the differentiated solution of the limit equation. This circumstance may justify the enhanced memory requirements which lattice-Boltzmann methods need in comparison with other methods.

If compared with the twoscale Chapman-Enskog expansion, regular asymptotic expansions (truncated power series) seem to be a much more natural approach to analyze lattice-Boltzmann methods. In fact, a regular expansion was already used by Hilbert[10] to study the scaled Boltzmann equation. Furthermore, in numerics, regular expansions date back to early investigations of finite difference methods, to specify the numerical error (see e.g. [50] page 135). Nonetheless, only recently this technique has been applied to lattice-Boltzmann methods [34], which is also adopted here.

Let us finally remark with regard to the second item that under certain scalings, solutions of Boltzmann-type equations display a convergence behavior with respect to other differential equations. This indicates that the convergence of lattice-Boltzmann schemes encapsulates two limit processes:

- The convergence of the scheme under the refinement of the discretization.

- The singular limit inherent in scaled Boltzmann-type equations.

Therefore typical phenomena of scaled Boltzmann-type differential equations can be observed likewise in the context of lattice-Boltzmann algorithms. Among other topics we will particularly focus on this issue which includes the appearance of multiple time scales and initial layers (see figure 1).

[10]That is why it is often referred to as Hilbert expansion in the context of kinetic theory.

Organization and scope of the work

Prerequisites concerning some specific vocabulary, notation and the general strategy of convergence proofs are collected in subsection 1.1.2 (concerning lattice-Boltzmann methods) and section 1.3 (concerning the numeric/asymptotic analysis). Apart from this, the chapters as well as the three sections of the first chapter are nearly self-contained and should be readable independently from each other. The cohesion is mainly generated by the circumstance that different aspects of equal or similar objects are discussed over several chapters.
To keep the notation simple, there is no unique notation for the entire thesis. Instead, the notation may slightly change from chapter to chapter or even from section to section.
Passages appearing in small printing are less important. This concerns digressions providing extra-material, some lemmas and propositions having rather the character of a tool than a result in the specific context as well as some tedious computations. Furthermore the text is complemented by numerous footnotes containing additional explanations, suggestions and cross-references.

To facilitate the orientation, a summary of each chapter is given here:

Chapter 1 conveys some background knowledge about the Boltzmann equation in **subsection 1.1.1**. Above all, the provenance of the Boltzmann equation is motivated and its relation to macroscopic equations of fluid dynamics is established. In contrast to the historic development shortly outlined above, **subsection 1.1.2** introduces to lattice-Boltzmann methods from the perspective of mathematical kinetic theory dealing with Boltzmann-type equations that are much simpler than the original Boltzmann equation but retain essential properties. It is shown that scaled finite-velocity Boltzmann equations (also referred to as lattice-Boltzmann equations) reveal much flexibility to be related with other equations via limit processes. Finally, it is shown how the classical lattice-Boltzmann scheme is derived from finite-velocity Boltzmann equations.
A rather technical paragraph is concerned with the computation of the structure relations pertaining to discrete velocity models. In contrast to one-dimensional velocity models, these relations get more complicated in two and more dimensions; concretely the D2P9 model is considered. Theses relations are essential prerequisites for the analysis of lattice-Boltzmann methods.

The goal of **section 1.2** is to underline the close relationship between lattice-Boltzmann schemes in two space dimensions (being already of practical significance) and the model algorithms to be investigated here.
So, the D1P3 lattice-Boltzmann algorithm is derived from the standard D2P9 algorithm. Combining symmetry properties of the D2P9 algorithm with special initializations, the number of populations as well as the number of spatial dimensions can be reduced such that the D1P3 algorithm is finally extracted. Endowed with a slightly more general equilibrium than found in this derivation, the D1P3 lattice-Boltzmann algorithm serves as our main model scheme.
Notice that the D1P2 algorithm considered in several chapters is a particular case of

the D1P3 scheme. However, it could be derived independently of the D1P3 scheme by considering an appropriate D2P4 lattice-Boltzmann algorithm.

In **section 1.3** we provide some concepts of numerical and asymptotical analysis in as much as they are needed here. Particular attention is paid to the notions of *consistency* and *stability*. The abstract presentation is illustrated by examples.

Before we address lattice-Boltzmann algorithms in chapter 4, 5 and 6, we examine some properties of lattice-Boltzmann equations. Formally, the latter can be investigated more easily since asymptotic expansions with respect to the scaling parameter are applied without being combined with Taylor expansions. Many features of lattice-Boltzmann algorithms are not of specific discrete nature as shown in chapter 2 and 3. Examples are the hierarchy of asymptotic orders, the decoupling of odd and even orders, the occurrence of initial layer etc..

Chapter 2 deals with a D1P2 lattice-Boltzmann equation. In **section 2.1** the difference between the *hyperbolic* and *parabolic* scaling is highlighted. Furthermore, we show that the D1P2 lattice-Boltzmann equation, which is actually a system, can be transformed into a scalar, singularly perturbed equation. For the case, of the parabolic scaling and a linear equilibrium, convergence is demonstrated in **section 2.2** by expanding the solution in terms of a Fourier series. Generally, uniform convergence in time does not extend to the temporal derivative. This indicates the existence of an initial layer, which is expected whenever the solution of the perturbed equation is not initialized compatibly with the limit problem.
Confining ourselves on a single Fourier mode in **section 2.3**, the appearance of the initial layer and its structure is thoroughly analyzed. A twoscale expansion discloses a hierarchy of equations for the initial layer and the regular part of the solution which are mutually coupled.

Chapter 3 presents the analysis of a lattice-Boltzmann equation based on the D1P3 velocity model, which is distinguished from the D1P2 model by an additional rest population. In this case convergence is shown without recourse to an equivalent scalar reformulation. Instead, the proof is divided into a stability part accomplished by an energy estimate and a consistency part performed by a regular asymptotic expansion. It is observed that the regular expansion can only cover very special initial values. Therefore further, irregular expansions are introduced such that arbitrary initial values can be handled. These expansions turn out to describe initial layers. The chapter closes with a short discussion of non-periodic boundary conditions.

Chapter 4 is devoted to the consistency analysis of a D1P3 lattice-Boltzmann algorithm discretizing the advection-diffusion equation. The algorithm is essentially the discrete counterpart of the lattice-Boltzmann equation studied in chapter 3. The regular expansion is formally performed up to arbitrary orders in **section 4.1** providing structural insight. Above all, it is seen how a hierarchy of evolution equations builds up, which determines the asymptotic order functions recursively. Furthermore the asymptotic expansion indicates the convergence order. In **section**

4.2, the consistency of truncated expansions is defined, which are then constructed in **section 4.3**. At last, we throw a quick glance at numerical initial layers displaying more structure in the discrete case.

Whereas the previous investigations were mainly concerned with lattice-Boltzmann equations and algorithms in parabolic scaling, a hyperbolically scaled D1P2 algorithm is discussed in **chapter 5**. The consistency analysis, which proceeds analogously to chapter 3 and 4, shows that the algorithm discretizes the advection equation. However, the major interest of this chapter is focused on the long-term behavior of the scheme which is wrongly predicted by the regular expansion. Actually the evolution of the algorithm contains different times scales which are detected by numerical experiments at the beginning of **section 5.2**. Using a multiscale expansion, the analysis of the algorithm yields a precise description of its behavior, which is validated by examples.

To justify the expansions in chapter 4 and 5 rigorously and to answer thereby the question of convergence, we examine the *stability* of the algorithms in **chapter 6**. After some preparatory remarks about shift matrices to write the algorithms in matrix form, ℓ_∞-stability in time and space is proven in **section 6.2**. The result refers to the parabolically scaled D1P2 algorithm specializing the D1P3 algorithm in chapter 4. A detailed spectral analysis of the evolution matrix continues the section. Among others, eigenvectors and eigenvalues are computed where the latter are compared with the eigenvalues that belong to the evolution operator of the diffusion-advection equation. Furthermore the symmetry properties of the spectrum are discussed.
Section 6.3 considers almost the same D1P2 algorithm as presented in chapter 5 but in *hyperbolic* scaling . Since this algorithm discretizes the advection equation, analyzing the role of the CFL-condition is natural. Using the results in subsection 6.2.2, it is shown that the CFL-condition is necessary and sufficient for stability. However, this statement cannot be extended to the analogous D1P3 scheme as pointed out in 6.5. The section is finished by comparing the actual stability properties with a formal stability discussion based on the multiscale expansion in 5.2.
So far, the considered algorithms refer to periodic boundary conditions. In **section 6.4** we deal with bounce-back type boundary conditions emulating homogeneous Dirichlet and Robin boundary conditions of the macroscopic limit equation. For the diffusive equilibrium, eigenvalues and eigenvectors of the evolution matrix pertaining to the parabolically scaled D1P2 algorithm are computed. However, the consistency analysis of the boundary conditions is left for future work.
Finally, in **section 6.5** we discuss the possibilities to generalize the results of 6.2 and 6.3 to the corresponding D1P3 algorithms. Furthermore we briefly resume the discussion of 6.4 where now an additional advective term is admitted in the equilibrium.

Since kinetic boundary conditions in two and more space dimensions generally produce boundary layers, we exemplarily study this phenomenon in **chapter 7**. Concretely, we consider the one-dimensional Poisson equation with Dirichlet boundary conditions. This is discretized by means of a five-point stencil where different extrapolations are used to deal with the overhanging parts of the stencil near the boundary. After some preliminaries, we prove stability of the resulting schemes in **section 7.4**. It turns out that stability alone is not enough to prove the observed convergence rates. For this, finer estimates are needed resulting from the *damping property* of **section 7.5**. The chapter closes with a presentation of numerical experiments that illustrate the irregular asymptotic expansion which describes the numerical error.

In the **appendix** we compute the eigenvalues of the linear, one-dimensional advection-diffusion operator with constant coefficients for various boundary conditions. From this the eigenvalues of the semigroup of evolution operators associated to the advection-diffusion equation are easily obtained. They are compared with the eigenvalues of lattice-Boltzmann evolution matrices in chapter 6.

Chapter 5 was originally designed as manuscript for a shortcourse which took place at ICMMES[11] 2006 in Hampton (VA). It is also connected with [35] while section 6.3 is based on [55].

[11] International Conference of Mesoscopic Methods for Engineering and Science

Chapter 1

Introduction to lattice Boltzmann methods and their analysis

1.1 Initiation to lattice-Boltzmann methods

Lattice-Boltzmann methods can be approached from three major directions as depicted in figure 1.1:

- cellular automata, particularly lattice-gas automata,
- relaxation schemes and discretization methods (finite differences),
- Boltzmann equation and kinetic theory of gases.

Historically, lattice-Boltzmann methods attracted attention as an important improvement of lattice-gas automata. So they were not originally based on differential equations but rooted in the realm of discrete dynamical systems with boolean variables. From a mathematical point of view, lattice-Boltzmann equations are akin to relaxation schemes which becomes apparent if they are transformed into their equivalent moment system. Finally, a lot of inspiration and vocabulary – including the name lattice-*Boltzmann* – is borrowed from the theory twining around the Boltzmann equation. Unlike some textbooks introducing to lattice-Boltzmann methods (cf. [63], [60]), we shall not retrace the historic development and skip the field of lattice-gas automata completely.

This first section of chapter 1 begins with the physical background touching on some key-words in the field of kinetic theory (for instance: equilibrium, collision invariant, BGK, scaling). Subsection 1.1.1 leans partly on what may be found in textbooks about statistical physics and fluid dynamics; it is mostly drawn from equivalent sections of the books [7, 8]. The last paragraph gives a short survey of the scaled Boltzmann equation referring to the paper [2].
Subsection 1.1.2 imparts the essential ideas leading to "stylized forms of the Boltzmann equation now going by the name *lattice-Boltzmann equation*. These relinquish most mathematical complexities of the true Boltzmann equation without sacrificing physical fidelity in the description of many situations involving complex fluid motion[1]."

[1]Text in quotation marks cited from the synopsis of [60].

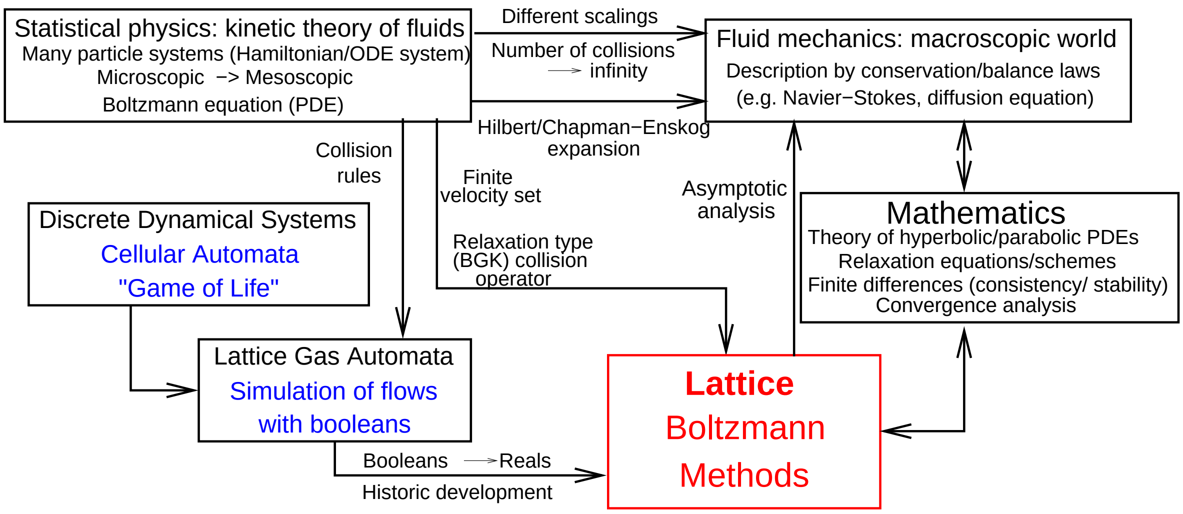


Figure 1.1: Lattice-Boltzmann methods embedded in their mathematical and physical context.

Furthermore, subsection 1.1.2 portrays the kinetic approach and its link to relaxation methods by a simple example. A technical paragraph collects some useful facts that are necessary for the analysis of 2D lattice-Boltzmann methods. Finally, the lattice-Boltzmann algorithm is derived from the lattice-Boltzmann equation.

There might be two principal reasons to deal with the Boltzmann equation and kinetic theory:

- First, the Boltzmann equation describes relevant physics.
- Second, the Boltzmann equation exhibits a rich structure which can be exploited to use it as an approximating tool for solving other equations.

As our objective is to pave a way towards lattice-Boltzmann methods, we take up the second perspective. Therefore we focus on structural aspects and the relation between the Boltzmann equation and macroscopic fluid dynamics. Above all, we do not discuss the intrinsics of the collision operator but recall only certain structurally important properties.

1.1.1 A brief introduction to kinetic theory

Kinetic theory is a branch of statistical physics, which is concerned with *non-equilibrium* phenomena especially in the context of fluids (e.g. gases and liquids). This definition obliges us to explain the term equilibrium more precisely: if a system is left without exterior interaction, it reaches an equilibrium state[2] after a certain while (relaxation time). In this final state the system can be characterized by only

[2]Equilibria are stationary states but the converse is not true. For instance, a rod between two heat sources of different temperature evolves into a stationary state with a constant temperature gradient. If the rod is isolated from the heat sources the temperature relaxes gradually towards an equilibrium marked by a spatially constant temperature.

a few macroscopic quantities like pressure, temperature, concentration etc.. However, in most applications the system is exposed to outer influences resulting into more complicated non-equilibrium states, which also need an adequate description to become computable. For a concrete illustration we imagine a monoatomic gas enclosed inside a box $\mathcal{B} \subset \mathbb{R}^3$.

Microscopic picture

The first approach describes the system on a microscopic level keeping track of every single particle. Each of the N atoms, belonging all to the same species, is modeled as a small *hard sphere* of mass m without any internal degrees of freedom. Hence possible self-rotation[3], vibrational modes or the energetic excitation of the electronic shell are completely disregarded. Similarly to the balls of a billiard game, the atoms interact with each other by *elastic collisions* conserving linear momentum and kinetic energy. Likewise, the atoms undergo *specular*[4] *reflection* if hitting the wall. Mathematically, the gas can be described in the framework of classical mechanics by a system of *Hamilton equations*[5] reducing to the simple case of free motion

$$\dot{\mathbf{x}}_i = \mathbf{v}_i, \qquad \dot{\mathbf{v}}_i = 0 \qquad \text{for } \ i \in \{1, ..., N\},$$

complemented by additional collision[6] and reflection[7] rules. Here $\mathbf{x}_i$ denotes the position and $\mathbf{v}_i$ the velocity of the i^{th} particle as a function of time. Between collisions, the velocities $\mathbf{v}_i$ remain constant and the particles move uniformly.
The state of the whole system corresponds to a point in a $6N$-dimensional state

---

[3]In this model a hard sphere particle has no orientation like a rigid body. Even if it could rotate about some axis, there is no energy exchange admitted by collisions between the rotational and translational degrees of freedom. This would require tangential contact forces (friction). As any interchange between translational and rotational energy is excluded, the rotational energy matters only as a constant, which is conveniently set to zero (free choice of energy scale).

[4]Derived from the Latin word *speculum* = mirror.

[5]For autonomous systems the Hamilton function represents the total energy in terms of spatial coordinates $\mathbf{x}_i$ and generalized momenta $\mathbf{p}_i$. The equations of motion (Hamilton equations) are

$$\dot{\mathbf{x}}_i = \ \nabla_{\mathbf{p}_i} H, \qquad \dot{\mathbf{p}}_i = -\nabla_{\mathbf{x}_i} H, \qquad i \in \{1, ..., N\}.$$

The Hamilton function for the system of N hard spheres is $H(\mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{p}_1, ..., \mathbf{p}_N) = \frac{1}{2m} \sum_{i=1}^{N} p_i^2$, where $\mathbf{p}_i = m\mathbf{v}_i$ and $\mathbf{v}_i$ denotes the velocity of the i^{th} sphere.

[6]If two hard spheres of equal mass (but not necessarily of equal volume) collide with each other, their post-collisional velocities are given in terms of their pre-collisional velocities by

$$\begin{aligned} \mathbf{v}_1^{\text{post}} &= \mathbf{v}_1^{\text{pre}} + \langle \mathbf{v}_2^{\text{pre}} - \mathbf{v}_1^{\text{pre}}, \mathbf{e} \rangle \mathbf{e}\,, \\ \mathbf{v}_2^{\text{post}} &= \mathbf{v}_2^{\text{pre}} - \langle \mathbf{v}_2^{\text{pre}} - \mathbf{v}_1^{\text{pre}}, \mathbf{e} \rangle \mathbf{e}\,. \end{aligned}$$

The unit vector $\mathbf{e}$ points from the center of the first to the center of the second sphere at the moment of collision. Evidently the exchange of momentum (velocity) occurs only in the direction of $\mathbf{e}$.

[7]If $\mathbf{n}$ denotes the unit normal vector of the reflecting surface, the outgoing velocity $\mathbf{v}_{\text{out}}$ of the reflected particle is given in terms of the incoming velocity $\mathbf{v}_{\text{in}}$ by the formula

$$\mathbf{v}_{\text{out}} = \mathbf{v}_{\text{in}} + 2\langle \mathbf{v}_{\text{in}}, \mathbf{n} \rangle \mathbf{n}\,.$$

Here $\mathbf{n}$ is assumed to be directed into the half-space of the particle, i.e. $\langle \mathbf{v}_{\text{in}}, \mathbf{n} \rangle \leq 0$. This formula describes (specular) reflection.

space[8] (also denoted as phase-space or Γ-space) with three spatial coordinates (position) and three momentum coordinates (velocity components) for each particle. In many applications it is not possible to have full information about the system. To deal with this situation of incomplete knowledge, a state-density P is introduced. The integral

$$\int_S P(\mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{v}_1, ..., \mathbf{v}_N)\, \mathrm{d}(\mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{v}_1, ..., \mathbf{v}_N)$$

gives the probability, that the system realizes a state contained in a subset S of the state-space. The evolution of P is governed by *Liouville's equation*[9] which becomes in this case

$$\partial_t P + \sum_{i=1}^{N} \mathbf{v}_i \cdot \nabla_{\mathbf{x}_i} P = 0.$$

Besides initial conditions, this PDE must be supplemented by boundary conditions taking the collisions and reflections into account.

From a theoretical point of view this approach might be satisfying. However, in view of about 10^{24} particles for a system of macroscopic dimension, it is practically

[8]The state space is contained in $\mathcal{B}^N \times (\mathbb{R}^3)^N$. The positional part (for the description of the spatial configuration) is a complicated subset of $\mathcal{B}^N$, which must take into account, that hard spheres can neither overlap each other nor the walls of the box $\mathcal{B}$.

[9]For a general Hamilton function $(\mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{p}_1, ..., \mathbf{p}_N) \mapsto H(\mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{p}_1, ..., \mathbf{p}_N)$ Liouville's equation reads

$$\partial_t P + \sum_{k=1}^{N} \Big(\nabla_{\mathbf{p}_k} H \cdot \nabla_{\mathbf{x}_k} P - \nabla_{\mathbf{x}_k} H \cdot \nabla_{\mathbf{p}_k} P \Big) = 0.$$

The sum is referred to as *Poisson bracket*. Liouville's equation is derived by means of the Hamiltonian flux F^t, which solves Hamilton's equations $\dot{\mathbf{X}}_i = \nabla_{\mathbf{p}_i} H$ and $\dot{\mathbf{P}}_i = -\nabla_{\mathbf{x}_i} H$. The solution corresponds to a curve in the phase-space which is given by $\mathbf{X}_i(t) = F^t_{\mathbf{x}_i}(\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_N, \overline{\mathbf{p}}_1, ..., \overline{\mathbf{p}}_N)$ and $\mathbf{P}_i(t) = F^t_{\mathbf{p}_i}(\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_N, \overline{\mathbf{p}}_1, ..., \overline{\mathbf{p}}_N)$ as function of the time t and the initial values. If Ω denotes an arbitrary measurable subset of the phase space, the probability of finding the state of the system in $F^t(\Omega)$ at time t does not depend on the time. Hence we conclude

$$\int_\Omega P = \int_{F^t(\Omega)} P \qquad \Rightarrow \qquad \frac{\mathrm{d}}{\mathrm{d}t} \int_{F^t(\Omega)} P = 0.$$

As the flux F_t represents a diffeomorphism of the phase space, the right integral over $F^t(\Omega)$ can be written as integral over Ω using the transformation rule.

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{F^t(\Omega)} P(t, \mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{p}_1, ..., \mathbf{p}_N)\, \mathrm{d}(\mathbf{x}_1, ..., \mathbf{x}_N, \mathbf{p}_1, ..., \mathbf{p}_N) \;\; =$$
$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega P\big(t, \mathbf{X}_1(t), ..., \mathbf{X}_N(t), \mathbf{P}_1(t), ..., \mathbf{P}_N(t)\big) \left| \frac{\partial F_t(\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_N, \overline{\mathbf{p}}_1, ..., \overline{\mathbf{p}}_N)}{\partial(\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_N, \overline{\mathbf{p}}_1, ..., \overline{\mathbf{p}}_N)} \right| \mathrm{d}(\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_N, \overline{\mathbf{p}}_1, ..., \overline{\mathbf{p}}_N)$$

Thanks to the *symplecticity* of Hamiltonian fluxes the determinant of its Jacobian is 1. Drawing the time derivative inside the integral yields finally

$$\int_\Omega \Big[\partial_t P + \sum_{i=1}^{N} \dot{\mathbf{X}}_i \cdot \nabla_{\mathbf{x}_i} P + \dot{\mathbf{P}}_i \cdot \nabla_{\mathbf{p}_i} P \Big]\, \mathrm{d}(\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_N, \overline{\mathbf{p}}_1, ..., \overline{\mathbf{p}}_N) = 0.$$

Holding true for arbitrary Ω, this equation must be satisfied pointwise provided the integrand is sufficiently regular. This yields Liouville's equation.

far beyond feasibility to handle the enormous amount of arising data. Moreover, the exact movement of each particle is of little interest, because it does not significantly affect the state of the entire system. Therefore a so-called *mesoscopic* approach has been devised which does not consider individual particles as in the above *microscopic* description.

Motivation of the Boltzmann equation

To obtain the mesoscopic description, the high-dimensional Γ-space is reduced to the six-dimensional phase space (μ-space[10]) of a single particle. Analogously, the function P with $6N+1$ arguments becomes a function $(t, \mathbf{x}, \mathbf{v}) \mapsto f(t, \mathbf{x}, \mathbf{v})$ of only seven arguments, where t stands for the time and $\mathbf{x}, \mathbf{v}$ contain each three components to fix the position and velocity. Similarly to P the distribution function f has a probabilistic meaning depending on how it is gauged:

$$\int_{\mathcal{B}} \int_{\mathbb{R}^3} f(t, \mathbf{x}, \mathbf{v}) \, \mathrm{d}\mathbf{v} \mathrm{d}\mathbf{x} = \begin{cases} 1 & \text{i) probability density} \\ N & \text{ii) particle density} \\ Nm & \text{iii) mass density} \end{cases}$$

For $X \subset \mathcal{B}$ and $V \subset \mathbb{R}^3$ the integral

$$\int_X \int_V f(t, \mathbf{x}, \mathbf{v}) \, \mathrm{d}\mathbf{v} \, \mathrm{d}\mathbf{x}$$

permits the following interpretations:

i) probability that all N particles are in $X \times V$,

ii) expected number of particles in $X \times V$,

iii) expected mass in $X \times V$.

Here we prefer the third interpretation. If $f(t, \mathbf{x}, \mathbf{v})$ is integrated over the entire velocity space it gives the local mass density in space.
The evolution of the distribution function f is prescribed by the *Boltzmann equation*

$$\partial_t f + \mathbf{v} \cdot \nabla_{\mathbf{x}} f = Q(f, f) \tag{1.1}$$

adopting the role of Liouville's equation. The Boltzmann equation becomes quickly comprehensible by a short reasoning of balance: Consider for this a small volume $X \times V$ around a point $(\mathbf{x}, \mathbf{v})$ in the phase-space. Assume that f is scaled so that the integral over the whole phase space corresponds to the number of particles. By what effects will the number of particles change in $X \times V$ over a short time? First, some particles leave X while other enter it from the surrounding. This is taken into account by the advective term $\mathbf{v} \cdot \nabla_{\mathbf{x}} f$. There is a total loss of particles (with velocities in V close to $\mathbf{v}$) if $\mathbf{v}$ is directed parallel to the gradient. Then more particles move into X than stream out. Otherwise, in the case of antiparallel $\mathbf{v}$ and $\nabla_{\mathbf{x}} f$, the reverse process happens and X experiences a total gain of particles.

[10] A mnemonic might be: μ abbreviating the Greek word $\mu\acute{o}\nu o\varsigma$ = single.

Second, the particles collide with each other changing thereby their velocities which causes a gain and loss of particles in V. It is the right hand side $Q(f,f)$ that has to model this effect. The double argument indicates that the collision term depends quadratically on f, which makes the Boltzmann equation nonlinear[11]. Explicitly, the collision term is given by an integral operator reflecting the specific dynamics of the particle interaction (elastic collisions).
In the presence of external forces $\mathbf{K}$ (gravitation) there is a further addend $\mathbf{K} \cdot \nabla_{\mathbf{v}} f$ appearing on the left hand side of the Boltzmann equation (1.1). This term accounts for the acceleration of the particles due to $\mathbf{K}$.

Moments and collision invariants

The transition from the mesoscopic distribution function f to macroscopic quantities is managed by the computation of moments[12]. Moments of $0^{\text{th}}, 1^{\text{st}}$ and 2^{nd} order are of direct physical meaning:

$$\begin{array}{cclcll} 1 & \rightarrow & \rho(t,\mathbf{x}) & = & \int_{\mathbb{R}^3} \quad f(t,\mathbf{x},\mathbf{v})\,\mathrm{d}\mathbf{v} & \text{density of mass} \\ (v_x, v_y, v_z) & \rightarrow & \big(\rho\mathbf{u}\big)(t,\mathbf{x}) & = & \int_{\mathbb{R}^3} \mathbf{v}\, f(t,\mathbf{x},\mathbf{v})\,\mathrm{d}\mathbf{v} & \text{density of momentum} \\ \frac{1}{2}\|\mathbf{v}\|^2 = \frac{1}{2}v^2 & \rightarrow & \kappa(t,\mathbf{x}) & = & \frac{1}{2}\int_{\mathbb{R}^3} v^2 f(t,\mathbf{x},\mathbf{v})\,\mathrm{d}\mathbf{v} & \text{density of energy} \end{array}$$

The conservation of mass (number of the particles), momentum and kinetic energy in the elastic collisions of the particles manifests itself in balance equations for the corresponding fields $\rho, \rho\mathbf{u}$ and ρe. Before these equations can be deduced, an important property of the collision term has to be known.

A function $\mathbf{v} \mapsto \iota(\mathbf{v})$ over the velocity space is called a *collision invariant* if it satisfies

$$\int_{\mathbb{R}^3} Q\big(f,f\big)(t,\mathbf{x},\mathbf{v})\,\iota(\mathbf{v})\;\mathrm{d}\mathbf{v} = 0$$

for all possible distribution functions f. Exploiting the internal structure of the collision term $Q(f,f)$, it is possible to derive a functional equation for ι which only permits a solution of the type

$$\iota(\mathbf{v}) = a + \mathbf{b}\cdot\mathbf{v} + cv^2 \qquad \text{with } a,c \in \mathbb{R},\;\; \mathbf{b} \in \mathbb{R}^3 \text{ and } v = \|\mathbf{v}\|.$$

By integrating the Boltzmann equation (1.1) over the velocity space, the collision term on the right hand side vanishes as $\mathbf{v} \mapsto 1$ is a collision invariant. Then Interchanging the time derivative ∂_t and the spatial nabla operator $\nabla_{\mathbf{x}}$ with the integration, yields the continuity equation for the mass density:

$$\partial_t \rho + \nabla\cdot(\rho\mathbf{u}) = 0. \tag{1.2}$$

[11] It is remarkable that the Boltzmann equation is non-linear in f, whereas Liouville's equation is linear in P. The derivation of the the collision term $Q(f,f)$ requires additional assumptions to obtain a closed equation in f. In particular, it is assumed that the velocities of the particles are statistically independently distributed at different times due to the collisions. This hypothesis is known under the name of *molecular chaos*. It becomes more clear if the Boltzmann equation is derived from the BBGKY hierarchy.

[12] Every polynomial μ with respect to the components of $\mathbf{v}$ generates a moment defined by $m_\mu(t,\mathbf{x}) := \int_{\mathbb{R}^3} \mu(\mathbf{v}) f(t,\mathbf{x},\mathbf{v})\,\mathrm{d}\mathbf{v}$. The order of the moment corresponds to that of the polynomial. The moments of monomials – forming a simple basis in the space of polynomials – are particularly important.

Similarly, one proceeds to obtain the following equations for the time evolution of the momentum and energy densities

$$\partial_t\big(\rho(t,\mathbf{x})\mathbf{u}(t,\mathbf{x})\big) + \int_{\mathbb{R}^3} \big(\mathbf{v}\cdot\nabla_{\mathbf{x}} f(t,\mathbf{x},\mathbf{v})\big)\,\mathbf{v}\,\mathrm{d}\mathbf{v} = 0, \tag{1.3}$$

$$\partial_t\kappa + \tfrac{1}{2}\int_{\mathbb{R}^3} \big(\mathbf{v}\cdot\nabla_{\mathbf{x}} f(t,\mathbf{x},\mathbf{v})\big)\,v^2\,\mathrm{d}\mathbf{v} = 0, \tag{1.4}$$

where the Boltzmann equation has been multiplied by $\mathbf{v}$ and $\frac{1}{2}v^2$ respectively before the integration is performed. The moments of second and third order resulting from the integration of the advection term cannot be completely expressed by $\rho, \rho\mathbf{u}$ and ρe. Therefore we are prevented from writing down a closed system of equations for $\rho, \mathbf{u}$ and e.

Nevertheless, further manipulations of the moment integrals make these terms more explicit and reveal equations that cannot hide a structural relation to the macroscopic equations of fluid dynamics. For this we introduce the *random* or *peculiar velocity*[13] $\mathbf{c} = \mathbf{v} - \mathbf{u}$ where $\mathbf{u}$ is the *streaming* or *bulk velocity*[14]. It is important to note that the mean value of the random velocity vanishes:

$$\int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\mathbf{c}\,\mathrm{d}\mathbf{v} = \int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})(\mathbf{v}-\mathbf{u})\,\mathrm{d}\mathbf{v} = \int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\mathbf{v}\,\mathrm{d}\mathbf{v} - \mathbf{u}\int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\,\mathrm{d}\mathbf{v} = \rho\mathbf{u} - \mathbf{u}\rho = 0.$$

This circumstance is used throughout the following computations.
So we get for the moment of second order in (1.4):

$$\begin{aligned}
\int_{\mathbb{R}^3} f v_i v_j \mathrm{d}\mathbf{v} &= \int_{\mathbb{R}^3} f\ (c_i - u_i)(c_j - u_j)\,\mathrm{d}\mathbf{v} \\
&= \int_{\mathbb{R}^3} f u_i u_j\,\mathrm{d}\mathbf{v} - u_i\int_{\mathbb{R}^3} f c_j\,\mathrm{d}\mathbf{v} - u_j\int_{\mathbb{R}^3} f c_i\,\mathrm{d}\mathbf{v} + \int_{\mathbb{R}^3} f c_i c_j\,\mathrm{d}\mathbf{v} \\
&= \int_{\mathbb{R}^3} f u_i u_j\,\mathrm{d}\mathbf{v} + \int_{\mathbb{R}^3} f c_i c_j\,\mathrm{d}\mathbf{v} \\
&= \rho u_i u_j + p_{ij}.
\end{aligned}$$

The second integral is referred to as (the component ij of) the stress tensor, which is evidently symmetric.
In an analogous manner we proceed with the energy density κ:

$$\begin{aligned}
\kappa = \tfrac{1}{2}\int_{\mathbb{R}^3} f v^2\,\mathrm{d}\mathbf{v} &= \tfrac{1}{2}\int_{\mathbb{R}^3} f\ (\mathbf{c}-\mathbf{u})\cdot(\mathbf{c}-\mathbf{u})\,\mathrm{d}\mathbf{v} \\
&= \tfrac{1}{2}\int_{\mathbb{R}^3} f c^2\,\mathrm{d}\mathbf{v} - \mathbf{u}\cdot\int_{\mathbb{R}^3} f\mathbf{c}\,\mathrm{d}\mathbf{v} + \tfrac{1}{2}u^2\int_{\mathbb{R}^3} f\,\mathrm{d}\mathbf{v} \\
&= \rho e + \tfrac{1}{2}\rho u^2.
\end{aligned} \tag{1.5}$$

Obviously, the density of energy splits[15] into two parts: The first one is attributed to the random motion of the atoms while the second one represents the kinetic energy of the macroscopic convection. e is called the specific internal energy.

[13] Sometimes also called *individual* velocity.
[14] Also called *collective* or *drift* velocity.
[15] This splitting is related to the formula of the variance: $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}X)^2$

Let us now tackle the moment of third order in (1.4) corresponding to the energy flux:

$$
\begin{aligned}
\tfrac{1}{2}\int_{\mathbb{R}^3} f v^2 v_i \,\mathrm{d}\mathbf{v} &= \tfrac{1}{2}\int_{\mathbb{R}^3} f\,(\mathbf{u}+\mathbf{c})\cdot(\mathbf{u}+\mathbf{c})(u_i+c_i)\,\mathrm{d}\mathbf{v}\\
&= \tfrac{1}{2}\int_{\mathbb{R}^3} f\,(u^2+2\mathbf{u}\cdot\mathbf{c}+c^2)(u_i+c_i)\,\mathrm{d}\mathbf{v}\\
&= \tfrac{1}{2}\rho u^2 u_i + \rho e u_i + \mathbf{u}\cdot\int_{\mathbb{R}^3} f\mathbf{c}c_i\,\mathrm{d}\mathbf{v} + \tfrac{1}{2}\int_{\mathbb{R}^3} f c^2 c_i\,\mathrm{d}\mathbf{v}\\
&= \tfrac{1}{2}\rho u^2 u_i + \rho e u_i + \sum_{l=1}^{3} p_{il}u_l + q_i.
\end{aligned}
$$

The last line shows that the energy flux is composed of different terms allowing a clear interpretation: The first and second one are the convective transport of macroscopic kinetic energy and of internal energy. In contrast, the third term reflects the work done by stress forces (acceleration) and the fourth one may be considered as heat flow.
Now, using these results, equations (1.2), (1.3) and (1.4) can be written in the following conservative form, which has the structure: the time derivative of a quantity is equal to the negative divergence of the corresponding flux.

$$\partial_t\rho + \sum_{j=1}^{3}\partial_{x_j}(\rho u_j) = 0 \tag{1.6}$$

$$\partial_t(\rho u_i) + \sum_{j=1}^{3}\partial_{x_j}(\rho u_i u_j + p_{ij}) = 0 \tag{1.7}$$

$$\partial_t(\tfrac{1}{2}\rho u^2 + \rho e) + \sum_{j=1}^{3}\partial_{x_j}\Big(\tfrac{1}{2}\rho u^2 u_j + \rho e u_j + \sum_{j=1}^{3} p_{jl}u_l + q_j\Big) = 0 \tag{1.8}$$

It is possible to derive evolution equations for p_{ij} and q_j from the Boltzmann equation in the same way as it has been done for $\rho, \mathbf{u}$ and e. However, the fluxes of these equations contain moments of third and fourth order. This requires further evolution equations involving even higher moments. Actually, the system cannot be closed in this way. Therefore the above equations are practically useless as long as we lack additional information about the distribution function f, that would allow to compute the unknown moments.

The equilibrium

Although we are primarily interested in non-equilibrium states of fluids, the equilibrium states are of utmost importance. This is due to the fact that the system starts to approach an equilibrium as soon as it can evolve bare of external constraints. As mentioned earlier, equilibria are steady and spatially homogeneous states. The distribution function f_{eq} of an equilibrium is derivable from the Boltzmann equation, noticing that the left hand side $\partial_t f_{\mathrm{eq}} + \mathbf{v}\cdot\nabla_{\mathbf{x}} f_{\mathrm{eq}}$ vanishes due to the stationarity and homogeneity of f_{eq}. In order to be a solution of the Boltzmann equation, this entails immediately that $Q(f_{\mathrm{eq}}, f_{\mathrm{eq}}) = 0$ representing the determining equation for f_{eq}. Once again, it is possible to extract from the collision operator Q by a detailed analysis that f_{eq} must have the general form

$$f_{\mathrm{eq}}(\mathbf{v}) = \exp(a + \mathbf{b}\cdot\mathbf{v} + cv^2) \tag{1.9}$$

with $a \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^3$ and $c < 0$ to be integrable. The proof of (1.9) is based upon the derivation of the collision invariants. (1.9) can be brought into a form reminding of the famous *Maxwell distribution* which is derived in statistical mechanics also by other methods (ensemble theory):

$$f_{\mathrm{eq}}(\mathbf{v}) = M(\mathbf{v}) = \alpha\exp\big(-\beta\|\mathbf{v}-\mathbf{w}\|^2\big).$$

To assign a concrete meaning to the constants $\alpha, \beta > 0$ and $\mathbf{w} \in \mathbb{R}^3$, the macroscopic moments of M must be computed.

For this, the subsequent integrals are of great use:

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\beta z^2} \mathrm{d}z = \sqrt{\frac{\pi}{\beta}}, \qquad \int_{\infty}^{\infty} z^2 \mathrm{e}^{-\beta z^2} \mathrm{d}z = \frac{1}{2}\sqrt{\frac{\pi}{\beta^3}}.$$

As the zeroth moment should yield the density ρ, we find for α

$$\rho = \int_{\mathbb{R}^3} \alpha \mathrm{e}^{-\beta\|\mathbf{v}-\mathbf{w}\|^2} \mathrm{d}\mathbf{v} \;=\; \alpha\sqrt{(\pi/\beta)^3} \qquad \Rightarrow \qquad \alpha \;=\; \frac{\rho}{\sqrt{\pi/\beta}^3}.$$

Using this result we conclude from the the first moment integral

$$\rho\mathbf{u} = \alpha \int_{\mathbb{R}^3} \mathbf{v}\mathrm{e}^{-\beta\|\mathbf{v}-\mathbf{w}\|^2} \mathrm{d}\mathbf{v} = \alpha \int_{\mathbb{R}^3} (\mathbf{z}+\mathbf{w})\mathrm{e}^{-\beta\|\mathbf{z}\|^2} \mathrm{d}\mathbf{z} = \alpha\mathbf{w}\sqrt{(\pi/\beta)^3} \quad \Rightarrow \quad \mathbf{u} = \mathbf{w}$$

In order to obtain the Maxwellian in the usual form we must give a preliminary remark concerning the *temperature*. This quantity is defined via the average energy of a particle due to its random motion. More precisely, if $\tilde{f}(\mathbf{c})$ is the probability for a particle (of mass m) at some point in time and space to have the random velocity $\mathbf{c}$, then the instantaneous local temperature T is defined by the equation $\frac{3}{2}k_B T = \frac{1}{2}m\int_{\mathbb{R}^3} c^2 \tilde{f}(\mathbf{c})\, \mathrm{d}\mathbf{c}$ with k_B denoting the Boltzmann constant. Now this probability is just given by $f(t, \mathbf{x}, \mathbf{c}+\mathbf{u})/\rho(t, \mathbf{x})$. Hence we obtain

$$\begin{aligned}
\tfrac{3}{2}k_B T &= \frac{m}{2\rho}\int_{\mathbb{R}^3} c^2 f(\cdot,\cdot,\mathbf{c}+\mathbf{u})\, \mathrm{d}\mathbf{c} \\
&= \frac{m}{2\rho}\int_{\mathbb{R}^3} \|\mathbf{v}-\mathbf{u}\|^2 f(\cdot,\cdot,\mathbf{v})\, \mathrm{d}\mathbf{v} \\
&= \frac{m}{2\rho}\left[\int_{\mathbb{R}^3} v^2 f(\cdot,\cdot,\mathbf{v})\, \mathrm{d}\mathbf{v} - 2\mathbf{u}\cdot\int_{\mathbb{R}^3} \mathbf{v} f(\cdot,\cdot,\mathbf{v})\, \mathrm{d}\mathbf{v} + \int_{\mathbb{R}^3} u^2 f(\cdot,\cdot,\mathbf{v})\, \mathrm{d}\mathbf{v}\right] \\
&= \frac{m}{\rho}\left[\tfrac{1}{2}\rho u^2 + \rho e - \rho u^2 + \tfrac{1}{2}\rho u^2\right] \;=\; me,
\end{aligned}$$

which relates the temperature with the internal energy. Now, let us attack the moment of the Maxwellian pertaining to v^2:

$$\begin{aligned}
\tfrac{1}{2}\rho u^2 + \rho e &= \tfrac{1}{2}\alpha\int_{\mathbb{R}^3} v^2 \mathrm{e}^{-\beta\|\mathbf{v}-\mathbf{u}\|^2}\, \mathrm{d}\mathbf{v} \\
&= \tfrac{1}{2}\alpha\int_{\mathbb{R}^2} \|\mathbf{z}+\mathbf{u}\|^2 \mathrm{e}^{-\beta\|\mathbf{z}\|^2}\, \mathrm{d}\mathbf{z} \\
&= \tfrac{1}{2}\alpha\int_{\mathbb{R}^2} z^2 \mathrm{e}^{-\beta z^2} \mathrm{d}\mathbf{z} \;+\; \overbrace{\alpha\mathbf{u}\cdot\int_{\mathbb{R}^2} \mathbf{z}\mathrm{e}^{-\beta z^2} \mathrm{d}\mathbf{z}}^{=0} \;+\; \tfrac{1}{2}\alpha u^2\int_{\mathbb{R}^2} \mathbf{z}\mathrm{e}^{-\beta z^2} \mathrm{d}\mathbf{z} \\
&= \tfrac{1}{2}\alpha\int_{\mathbb{R}^3} z^2 \mathrm{e}^{-\beta z^2} \mathrm{d}\mathbf{z} \;+\; \tfrac{1}{2}\rho u^2.
\end{aligned} \tag{1.10}$$

Computing the remaining integral

$$\begin{aligned}
\int_{\mathbb{R}^3} z^2 \mathrm{e}^{-\beta z^2} \mathrm{d}\mathbf{z} &= \int_{\mathbb{R}^3} (z_1^2+z_2^2+z_3^2)\mathrm{e}^{-\beta z_1^2}\mathrm{e}^{-\beta z_2^2}\mathrm{e}^{-\beta z_3^2}\, \mathrm{d}\mathbf{z} = 3\left(\int_{\mathbb{R}} \zeta^2 \mathrm{e}^{-\beta\zeta^2} \mathrm{d}\zeta\right)\left(\int_{\mathbb{R}} \mathrm{e}^{-\beta\zeta^2} \mathrm{d}\zeta\right)^2 \\
&= \tfrac{3}{2}\pi^{3/2}\beta^{-5/2},
\end{aligned}$$

we obtain from (1.10) using the expression for α and the relation between e and T:

$$\rho e = \tfrac{3}{4}\alpha\pi^{3/2}\beta^{-5/2} \qquad \Rightarrow \qquad e = \tfrac{3}{4}\beta^{-1} \qquad \Rightarrow \qquad \beta = \frac{m}{2k_B T}.$$

Thus the Maxwellian becomes

$$M(\mathbf{v}) = \alpha\mathrm{e}^{-\beta\|\mathbf{v}-\mathbf{u}\|^2} = \frac{\rho}{(\pi/\beta)^{3/2}}\mathrm{e}^{-\beta\|\mathbf{v}-\mathbf{u}\|^2} = \rho\left(\frac{m}{2\pi k_B T}\right)^{3/2} \exp\left(-\tfrac{m}{2k_B T}\|\mathbf{v}-\mathbf{u}\|^2\right).$$

It is clearly seen that the equilibrium distribution – in the form of a Maxwellian[16] – is uniquely defined by three independent macroscopic parameters, ρ, $\mathbf{u}$ and T.

A short remark about the long time behavior and the H-theorem

It is a fundamental physical observation that every system, being decoupled from interaction with the outside world, begins to relax into an equilibrium state. This behavior seems to be included in the dynamics described by the Boltzmann equation. Indeed, it has been proven that at least under restrictive conditions the solution of the Boltzmann equation (1.1) converges (weakly) to a Maxwellian while conserving mass, momentum and energy. The space-homogeneous case is discussed in [8] (6.4, page 182), for further results the references therein may be consulted.
It should be remarked that the long time behavior of the distribution function is also backed by the so-called H-theorem[17] which was historically the first clue indicating the important role of equilibrium states.

Final step to derive the Euler equations for inviscid flow

Due to the tendency of the distribution function f to relax into the equilibrium, there are many situations where f is close[18] to a so-called *local Maxwellian*

$$M\big(\rho(t,\mathbf{x}),\mathbf{u}(t,\mathbf{x}),e(t,\mathbf{x});\mathbf{v}\big) = \frac{3\rho(t,\mathbf{x})}{(4\pi e(t,\mathbf{x}))^{3/2}} \exp\big(-\tfrac{3}{4e(t,\mathbf{x})}\|\mathbf{v}-\mathbf{u}(t,\mathbf{x})\|^2\big)$$

with space and time depending macroscopic fields. With the ansatz

$$f(t,\mathbf{x},\mathbf{v}) = M\big(\rho(t,\mathbf{x}),\mathbf{u}(t,\mathbf{x}),e(t,\mathbf{x});\mathbf{v}\big)$$

the quantities p_{ij} and q_i in (1.7) and (1.8) can be computed explicitly in terms of ρ, $\mathbf{u}$ and e yielding thus the necessary closure relations. The following equations are found:

$$p_{ij} = (\tfrac{2}{3}\rho e)\delta_{ij} \qquad \text{and} \qquad q_j = 0.$$

In order to write the fluid-dynamic equations in the familiar form we need to introduce the pressure p. This quantity is defined as the *isotropic* part of the stress tensor, such that the equation $p_{ij} = p\delta_{ij} + \sigma_{ij}$ holds true with a traceless anisotropic part (σ_{ij}). Hence[19]

$$p = \tfrac{1}{3}\sum_{i=1}^{3} p_{ii} = \frac{1}{3}\int_{\mathbb{R}^3} fc^2 \mathrm{d}\mathbf{v}. \tag{1.11}$$

[16] Analogously, the Gaussian normal distribution (being constructed by the same kind of exponential function) is uniquely defined by the normalization condition (0^{th} moment), the mean value (1^{st} moment) and the variance (2^{nd} moment).

[17] Multiplying the Boltzmann equation with $(1+\log(f))$ (assuming $f>0$) and integrating over $\mathbf{v}\in\mathbb{R}^3$ yields

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\textstyle\int_{\mathbb{R}^3} f\log(f)\mathrm{d}\mathbf{v} + \nabla_{\mathbf{x}}\cdot\int_{\mathbb{R}^3} f\log(f)\mathbf{v}\,\mathrm{d}\mathbf{v} = \int_{\mathbb{R}^3}\log(f)Q(f,f)\mathrm{d}\mathbf{v},$$

where we have used the identity $\log(f)\partial_t f = \partial_t\big(f + f\log(f)\big)$ (holding analogously for the nabla operator) to put the derivatives out of the integral. Furthermore, the fact of 1 being a collision invariant has been exploited. Thanks to *Boltzmann's inequality* (which is verified for the BGK collision operator further below) the right hand side is ≤ 0. Integrating spatially over the box $\mathcal{B}$, it can be shown under appropriate boundary conditions (to get rid of the $\nabla_{\mathbf{x}}$-term) that the quantity $H(t) := \int_{\mathcal{B}}\int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\log\big(f(t,\mathbf{x},\mathbf{v})\big)\,\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{x}$ is monotone decreasing or equivalently $\frac{\mathrm{d}}{\mathrm{d}t}H(t)\leq 0$, which is the central statement of the H-theorem. Observe that H corresponds to the *Shannon entropy* in information theory, if f is seen as probability density.

[18] This can be justified by asymptotic expansions.

[19] The formula admits a simple motivation based on the model of hard sphere-like particles. By its internal pressure the gas exerts a force onto the walls of the box confining it. This force is generated by the exchange of momentum, when the particles are reflected at the wall. For a particle of mass m hitting the wall perpendicularly (with velocity $\mathbf{c}$) the momentum transmission

Comparing this with equation (1.5), defining the internal energy e, we infer the relation

$$p = \tfrac{2}{3}\rho e, \tag{1.12}$$

which is known as the *state equation* of the *ideal gas*. Hence, the system (1.6), (1.7), (1.8) attains a closed form

$$\left.\begin{aligned} \partial_t \rho + \nabla\cdot(\rho\mathbf{u}) &= 0 \\ \partial_t(\rho\mathbf{u}) + \nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u}) + \nabla p &= 0 \\ \partial_t(\tfrac{1}{2}\rho u^2 + \rho e) + \nabla\cdot\Big(\tfrac{1}{2}\rho u^2\mathbf{u} + \rho e\mathbf{u} + p\mathbf{u}\Big) &= 0 \end{aligned}\right\} \tag{1.13}$$

if p is eliminated by e employing (1.12). These equations are known as the *Euler equations* in *conservative form*[20] for *inviscid (non-viscous), compressible* flow. It should be observed that a *local Maxwellian* constructed by the solutions of the Euler system is generally not expected to be a solution of Boltzmann's equation. The above system – although deduced from the Boltzmann equation – is not equivalent since it is constituted only by the three lowest moment equations which do not comprise the full dynamics of Boltzmann's equation.

The BGK relaxation model

Due to the complicated nature of the collision operator Q, various simplifications have been proposed, which are known as *collision models*. The most prominent among these alternatives is the *BGK model*[21] retaining two key features of the original collision operator:

I) The existence of the *collision invariants* $1, \mathbf{v}$ and v^2.

II) The tendency of an arbitrary distribution to relax into an equilibrium.

For a distribution function f, the BGK collision operator $J(f)$ is explicitly given by the relaxation term

$$J(f) := \tfrac{1}{\tau}\big(E(f) - f\big)$$

with $\tau > 0$. Here $E(f)$ denotes the *local equilibrium* corresponding to f. This is computed as the local Maxwellian having at any space-time point the same moments as f with respect to the collision invariants $1, \mathbf{v}$ and v^2. Therefore the local Maxwellian depends implicitly on space and time. Let us remark that this collision operator is in a sense more complicated than the original one. Whereas the latter depends quadratically on f, the BGK collision operator is highly non-linear due to moments of f entering the Maxwellian. Yet, the BGK collision term drastically simplifies certain manipulations as we will see below.

This setting automatically ensures condition (I)

$$\int_{\mathbb{R}^3} \iota(\mathbf{v}) f \,\mathrm{d}\mathbf{v} = \int_{\mathbb{R}^3} \iota(\mathbf{v}) E(f)\,\mathrm{d}\mathbf{v} \qquad \Rightarrow \qquad \int_{\mathbb{R}^3} \iota(\mathbf{v}) J(f)\,\mathrm{d}\mathbf{v} = 0,$$

is $2m\mathbf{c}$. The number of impacts is proportional to the density of the particles and their velocity. Statistically, around one sixth of the particles have a velocity pointing orthogonally towards the wall. This explains the integrand in (1.11) including the factor $\frac{1}{3}$ and the quadratic dependence on $\mathbf{c}$.

[20] Strictly speaking the pressure must be integrated into the divergence term. Applying the product rule and substituting e by T, the compressible Euler equations are frequently given in the following form: $\partial_t\rho + \nabla\cdot(\rho v) = 0$, $\partial_t\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \rho^{-1}\nabla p = 0$, $\partial_t T + (\mathbf{u}\cdot\nabla)T + \frac{2}{3}T\nabla\cdot\mathbf{u} = 0$.

[21] Introduced by Bhatnagar, Gross and Krook.

where $\iota(\mathbf{v})$ stands for one of the collision invariants. Requirement (II) is easily checked in the case of a space-homogeneous distribution function $f(t, \mathbf{x}, \mathbf{v}) = \tilde{f}(t, \mathbf{v})$. Then $\rho, \mathbf{u}$ and e are space-independent as well. Moreover, these quantities must be constant in time, too. This is seen by multiplying the BGK equation $\partial_t \tilde{f} = J(\tilde{f})$ with each of the collision invariants and integrating over $\mathbf{v} \in \mathbb{R}^3$ leading to $\partial_t \rho = 0, \partial_t \mathbf{u} = 0$ and $\partial_t e = 0$. Since $\rho, \mathbf{u}$ and e are only numbers (constant functions) $E(f) = M_{\rho,\mathbf{u},e}(\mathbf{v})$ does neither depend on time nor space. Thus the BGK-equation degenerates to an ODE for $\tilde{f}$

$$\partial_t \tilde{f}(t, \mathbf{v}) = \tfrac{1}{\tau}\big(M_{\rho,\mathbf{u},e}(\mathbf{v}) - \tilde{f}(\mathbf{v})\big).$$

with $\mathbf{v}$ as parameter. The solution is

$$\tilde{f}(t, \mathbf{v}) = \tilde{f}(0, t)\mathrm{e}^{-t/\tau} + \big(1 - \mathrm{e}^{-t/\tau} M_{\rho,\mathbf{u},e}(\mathbf{v})\big),$$

which clarifies the meaning of τ. Apparently, this parameter determines how fast $\tilde{f}$ approaches the equilibrium $M_{\rho,\mathbf{u},e}$, justifying the name *relaxation time*.

Likewise, it is an easy task to prove an H-theorem for the BGK collision operator. For this we note that the convexity relation $(y - x)\log(x/y) \leq 0$ holds for all $x, y \in \mathbb{R}^+$. Identifying x with f and $M := E(f)$ with y, we obtain by applying the logarithmic addition theorem

$$0 \geq \int_{\mathbb{R}^3} (M - f)\log(f/M)\mathrm{d}\mathbf{v} = \int_{\mathbb{R}^3} (M - f)\log(f)\mathrm{d}\mathbf{v} - \overbrace{\int_{\mathbb{R}^3} (M - f)\log(M)\mathrm{d}\mathbf{v}}^{=0}.$$

The overbraced integral vanishes since $M - f = \tau J(f)$ corresponds to the collision operator (up to a factor) and $\log(M)$ represents (as function of $\mathbf{v}$) a special linear combination of the collision invariants. Hence we are allowed to infer

$$0 \geq \int_{\mathbb{R}^3} J(f)\log(f)\mathrm{d}\mathbf{v},$$

which is the analogon to the *Boltzmann inequality* being closely related to the H-theorem.

The Boltzmann equation with the BGK collision operator

$$\overbrace{\partial_t f + \mathbf{v} \cdot \nabla_{\mathbf{x}} f}^{=:Df} = \tfrac{1}{\tau}(M - f)$$

admits a formal solution by an expansion. For this we designate the left hand side with Df where D abbreviates the material derivative $\partial_t + \mathbf{v} \cdot \nabla_{\mathbf{x}}$. Shifting terms, we find

$$\begin{aligned} f &= M - \tau D f \\ &= M - \tau D M + \tau^2 D^2 f \\ &= M - \tau D M + \tau^2 D^2 M - \tau^3 D^3 f. \end{aligned}$$

The second equation has been obtained by inserting the first one into itself. Then plugging the first equation into the second results into the third one. This process can be repeated producing a power series in τ. Its coefficients are material derivatives of increasing order applied to the Maxwellian. An exception is the highest order term in τ where the derivative acts still on f itself. For small τ it is imaginable that the truncated expansion (neglecting the last term containing f) approximates the distribution function f.
In particular, we see that the leading order is given by the Maxwellian, i.e. the equilibrium, what we have already anticipated when we have computed p_{ij} and q_j in (1.7) and (1.8). Taking also the next order into account and computing the corresponding moments with respect to $M - \tau DM$ yields more complicated expressions for p_{ij} and q_j. These result finally into a *compressible Navier-Stokes* equation. As a byproduct one finds expressions for the *diffusivity* and *viscosity.* Qualitatively this provides an understanding how these material coefficients vary with temperature and density; quantitatively the macroscopic diffusivity and viscosity can be computed in terms of microscopic parameters with a reasonable ansatz for τ (e.g. mean collision time = mean free path/mean velocity).

Scaling properties

Fluids occur in very different appearances: reaching from water in deep oceanic trenches, exposed to high pressure, up to the thin gas of the outer atmospheric layers. According to the Boltzmann equation (1.1) taking only two effects (free flow and the collision of particles) into account, three major regimes should be distinguished

- *Knudsen gas* : collisions negligible w.r.t. free particle flow
- *Boltzmann gas* : collisions as important as free particle flow
- *Hydrodynamic regime* : collisions dominate free particle flow

It is surprising that the relatively simple mesoscopic model condensating into the Boltzmann equation is able to describe this variety of fluids with enormous differences in physical properties (e.g. density, temperature etc.). An important tool in this context is the scaling of the Boltzmann equation. Before we elucidate this procedure in some more detail, let us recall under what assumptions the Boltzmann equation has been derived.

1) The classical collision operator is based on *hard sphere dynamics.* This model crudely simplifies molecular interaction, because it overlooks further degrees of freedom where energy can be stored. Moreover, there are no contact forces between colliding molecules as between colliding billiard balls. Instead, repulsive and attractive forces between the particles of molecular size are of electric origin (due to the charge distribution within the electronic shells) being relevant over a certain spatial range (force field).

2) The collision term only accounts for *binary* collisions. This assumes quite a low density of the fluid particles, such that ternary collisions and collisions involving even more particles are not significant.

3) The quadratic dependence of the collision term on the distribution function f can only be justified rigorously, if the so-called *Boltzmann-Grad limit* is considered. This couples the size[22] of the particles (described by σ being twice their diameter) to the number N of particles, such that $N\sigma^2$ remains finite if N goes to infinity. So the number of collisions per unit time and unit volume remains constant although the density of particles increases. Since the volume of a single particle is proportional to σ^3 we see that the Boltzmann-Grad limit implies $N\sigma^3 \to 0$ as $N \to \infty$. Thus the volume occupied by the particles vanishes in the limit.

This shows that the Boltzmann equation is especially designed for the regime of a *rarefied gas*, where collisions and free flow have equal importance and where the average distances between particles are much larger than their spatial extension. To capture the case of a Knudsen gas, the collision term can be set to 0. Remarkably, many aspects of the other extreme – the hydrodynamic regime – which is the most complicated one from a microscopic point of view, are correctly described by the Boltzmann equation too.

The natural time scale for the Boltzmann equation is the mean time elapsing between two subsequent collisions of a particle. This time becomes very short in the hydrodynamic regime. In comparison, the change of macroscopic quantities evolves much slower. Therefore the macroscopic observation stretches over a very long time interval if measured in units of the mean collision time. In order to focus on the macroscopic evolution of the fluid, a sort of fast-motion (time-lapse) is needed which is mathematically realized by a time scaling. To be more concrete let us shrink the 'long' time interval $[0, \epsilon^{-\alpha}]$ (with $0 < \epsilon < 1$ and $\alpha > 1$) to the unit interval:

$$\begin{array}{rcl} t \in [0, \epsilon^{-\alpha}] & \longrightarrow & \hat{t} \in [0,1] \quad \hat{t} = \epsilon^{\alpha} t \\ f(t) & \longrightarrow & \hat{f}(\hat{t}) = f(\epsilon^{-\alpha}\hat{t}) \\ \partial_t f(t) & \longrightarrow & \epsilon^{\alpha}\partial_{\hat{t}}\hat{f}(\hat{t}) \end{array}$$

Similarly, it is possible to scale the space variable $\mathbf{x} = \epsilon^{\beta}\hat{\mathbf{x}}$ corresponding to compression ($\beta < 0$) or dilation ($\beta > 0$) of the box containing the fluid. Expressing the Boltzmann equation in the new variables $\hat{t}, \hat{\mathbf{x}}$ and $\hat{f}$, yields[23]

$$\epsilon^{\alpha}\partial_{\hat{t}}\hat{f} + \epsilon^{\beta}\mathbf{v}\nabla_{\hat{\mathbf{x}}}\hat{f} = C(\hat{f}). \tag{1.14}$$

In view of such a parameter-dependent equation it is natural to ask how the solution of (1.14) behaves if ϵ tends to 0 and whether there is some convergence to be observed. Although a positive answer is expected thanks to the physical interpretation of the scaling, the limit should be rather unusual from the mathematical point of view, because (1.14) becomes singular[24] for $\epsilon = 0$.

[22]The size determines the total cross section σ entering the collision integral.

[23]For convenience, hats above the symbols are usually omitted. The following statements hold for a large class of collision operators satisfying certain 'axioms' specified in [2]. To indicate this we write $C(f)$ instead of $Q(f, f)$.

[24]For $\alpha, \beta > 0$ the limit equation is $Q(f, f) = 0$.

Of course, mathematical answers require more precisely formulated questions. Furthermore, the statement of relevant convergence theorems is cumbersome due to a lot of technicalities contained in their proofs. Nevertheless, it can be said that, generally, if there is convergence of the macroscopic moments and the entropy, the limits solve macroscopic fluid dynamic equations. In particular, the limit $\epsilon \to 0$ with $\alpha = \beta = 1$ is related to the *compressible Euler equation* (cf. theorem I of [2]).

It is interesting to note that the compressible Euler equation is not the only macroscopic equation extractable from the Boltzmann equation by such a limit process. Setting $\alpha = 2$ and $\beta = 1$ leads to the *incompressible Navier-Stokes equation.* The increased exponent $\alpha = 2$ that even a longer time scale is necessary to detect the Navier-Stokes equation. Noting that the latter equation models viscous effects, which are ignored by the Euler equation, this becomes physically perspicuous, since viscosity represents the diffusion[25] of momentum. But diffusion processes evolve generally on a slower[26] time scale than pure advection processes (as described by the Euler equation). So a longer time interval is necessary to notice them. It is remarkable that the additional viscous terms of the Navier-Stokes equation are found at the expense of the compressibility.

This, however, does not mean that the Boltzmann equation would not be rich enough to contain also the compressible Navier-Stokes equation, as mentioned already in the previous paragraph. In [2] (theorem II) the solution of an ϵ-depending compressible Navier-Stokes equation[27] is used to construct an approximate solution[28] of the Boltzmann equation. The compressible Navier-Stokes equation is also connected to the Boltzmann equation via the *Chapman-Enskog expansion* as sketched in [7] (page 110) and [8] (11.5, page 333).

Still another ansatz is presented in [2] to formulate theorem III. A background Maxwellian M_0 (constant in time and homogeneous in space) is modulated by a small perturbation in the following way:

$$\hat{f} = M_0(1 + \epsilon^\gamma g_\epsilon)$$

to solve (1.14). Under appropriate conditions, it can be shown that the limits of the macroscopic moments of g_ϵ, if existing, may be weak solutions of the incompressible Navier-Stokes equation or the incompressible Stokes equation or even the

[25]Diffusion of momentum attenuates the kinetic energy of the collective, macroscopic motion and transforms it into the irregular motion of individual particles. Thus it has a frictional effect.

[26]Diffusion is related to stochastic processes. The expected distance from the origin of a random walker starting there at $t = 0$ is proportional to $\sqrt{t}$, while the distance is proportional to t for a 'deterministic' walker with constant velocity. During the time interval $[0, T^2]$ the random walker covers the same (expected) distance as a deterministic walker during $[0, T]$.

[27]Due to the dependence on the perturbation parameter ϵ this could be considered as a *modified equation.* It is also possible to construct other approximate solutions by means of macroscopic equations being independent of the perturbation parameter (Hilbert expansion).

[28]*Approximate solution* means that it satisfies the Boltzmann equation up to a residual that vanishes with ϵ. Theorem II of [2] claims a residual of magnitude $O(\epsilon^2)$. The technique of constructing approximate solutions, that we denote as *prediction functions*, is widely used throughout this work.

incompressible Euler equation depending on the choice of the parameters α, β and γ. For $\alpha = 2, \beta = 1$ and $\gamma = 1$ the incompressible Navier-Stokes equation is found, whereas the incompressible Stokes equation is obtained for $\gamma > 1$, i.e. if the perturbation vanishes faster than $\mathrm{O}(\epsilon)$ for $\epsilon \to 0$. Remarkably, the incompressible Euler equation turns out to be the limit equation if $1 < \beta, \alpha = \beta + 1$ and $\gamma = 1$.

Apart from the Chapman-Enskog expansion, the so-called *Hilbert expansion*[29] (see [7] and [8]) is useful to analyze kinetic equations like the Boltzmann equation. We will take up the idea of Hilbert expansions in the form of *regular asymptotic expansions*. These are our major tool for the analysis of lattice-Boltzmann methods and other numerical schemes.

1.1.2 A primer of lattice-Boltzmann methods

One of the major messages in the previous subsection is the connection between solutions of scaled Boltzmann equations on the one hand and fluid dynamic equations on the other hand. In particular, certain moments of the distribution function f converge to solutions of the Euler or Navier-Stokes equation if the scaling parameter is sent to 0. Hence f can provide approximate solutions for a variety of rather different equations.

From the perspective of numerical methods for PDEs, this point becomes quite interesting on condition that solving the Boltzmann equation requires less effort than solving the fluid dynamic *target equation*. Now, the Boltzmann equation, being an integro-differential equation, is generally very hard to solve because of the six-dimensional phase space and the complicated collision operator. Furthermore the velocity is involved as additional independent variable beside time and space.

However, it has already been mentioned that the original collision term, derived from the physical particle dynamics, may be simplified without loosing its main qualitative properties. It turns out that the Boltzmann equation essentially owes its importance as an approximating tool for solving other equations to a few structural features.

Basic idea: Finite velocity models

A drastic simplification of the Boltzmann equation is achieved by shrinking the velocity space from $\mathbb{R}^3$ to a discrete, actually finite set $\mathcal{S} \subset \mathbb{R}^3$. Formally there is not much to change: instead of $\mathbf{v}$ we will write $\mathbf{s}$ attaining only velocities contained in $\mathcal{S}$. The choice of a velocity set defines a *(velocity) model*. The *discrete velocity*

[29] Although both expansions (with respect to the scaling parameter ϵ) are quite similar, the underlying idea is different: While the Chapman-Enskog ansatz tries to expand the parameter-dependent equation itself (i.e. the occurring differential operators), Hilbert's ansatz assumes a truncated expansion of the solution that is inserted into (1.14). Another feature of the Chapman-Enskog expansion is that it works with several time scales.

Boltzmann equation[30] becomes then

$$\partial_t \mathsf{f} + \mathbf{s} \cdot \nabla \mathsf{f} = J(\mathsf{f}), \qquad \mathbf{s} \in \mathcal{S}, \tag{1.15}$$

which should be compared with (1.1). However, there is only a structural similarity between the collision term $Q(f,f)$ and its counterpart $J(\mathsf{f})$ (collision invariants, existence of equilibrium etc.). Furthermore the nabla operator is written without index because here it can refer to the space variables only.
The discrete distribution function or *population function* f is of the kind $(t, \mathbf{x}, \mathbf{s}) \mapsto \mathsf{f}(t, \mathbf{x}, \mathbf{s})$. Since $\mathbf{s}$ refers to a finite domain, it can be interpreted as a sort of index. This implies that $\mathsf{f}(t, \mathbf{x}, \cdot)$ can be thought of as a tuple. Usually, the discrete velocities in $\mathcal{S}$ are numerated. This allows the notation[31] $\mathsf{f}_k(t, \mathbf{x}) := \mathsf{f}(t, \mathbf{x}, \mathbf{s}_k)$ for $k \in \{1, ..., \#\mathcal{S}\}$. By slight abuse of notation the whole tuple $(\mathsf{f}_k)_k$ is denoted by f too. The components are referred to as *populations* representing the density of fictitious particles moving with one of the velocities in $\mathcal{S}$.

Introducing three diagonal matrices S_1, S_2, S_3, where the k^{th} diagonal entry is given by the first, second and third component of $\mathbf{s}_k$ respectively, (1.15) can also be written in the form

$$\partial_t \mathsf{f} + S_1 \partial_{x_1} \mathsf{f} + S_2 \partial_{x_2} \mathsf{f} + S_3 \partial_{x_3} \mathsf{f} = J(\mathsf{f}).$$

This equation manifests clearly the structure of an *hyperbolic system of first order* with diagonalized constant coefficient matrices. Therefore the $\#\mathcal{S}$ equations for the components are coupled only with each other by the collision term at the right hand side.

From a physical point of view, discrete velocity Boltzmann equations may appear quite weird on the mesoscopic scale. Yet, their algebraic structure and macroscopic behavior, i.e. the evolution of averaged quantities (moments), may reveal much resemblance to the original Boltzmann equation.
This explains why discrete velocity Boltzmann equations are of considerable interest and offer efficient and competing alternatives for simulating macroscopic processes (kinetic approach). The exact macroscopic behavior is determined by the collision operator, which usually reflects physical conservation properties of some generalized mass, momentum etc.. It should be mentioned that the application of discrete velocity Boltzmann equations is not limited to *target equations* of pure fluid dynamic background although they constitute the largest field of applications.

Some notation

Many different velocity models are in use leading to a strongly simplified particle dynamics. A basic requirement with respect to the finite velocity set $\mathcal{S}$ is the

[30] Here the words *discrete* and *finite* are practically used as synonyms; so one may also say finite velocity Boltzmann equation. But mostly we will speak of the *lattice-Boltzmann equation* with regard to our focus. Due to historic reasons, some people reserve this term for the discrete equation (1.34), which we refer to as the *lattice-Boltzmann algorithm* or *scheme*.

[31] In some situations the velocities are directly used as indices. This is particularly advantageous if the Boltzmann equation is considered in one space dimension because then the velocity vectors are signed numbers.

invariance under reflection at the origin ($\mathcal{S} = -\mathcal{S}$). So it will be henceforth assumed tacitly that $-\mathbf{s} \in \mathcal{S}$ whenever $\mathbf{s} \in \mathcal{S}$ is an admissible velocity. This means that particles can move forward and backward[32] like real particles but only in a finite number of directions.
In the context of lattice-Boltzmann methods the most popular velocity models generate cubic grids in $\mathbb{R}^d$ as being subsets of $\{-1, 0, 1\}^d$ where $d \in \{1, 2, 3\}$ denotes the spatial dimension. Examples for $d = 1$ and $d = 2$ are given in table 1.1. However, there are also other models, for instance the hexagonal one with velocity vectors marking the corners of a regular hexagon.
In order to distinguish different models easily, the notation[33] DdP$\#\mathcal{S}$ has been established. The figure behind D indicates the number d of spatial dimensions while P precedes the number $\#\mathcal{S}$ of populations (viz. discrete velocities). Of course, this nomenclature cannot be unambiguous (compare the D2P4 and D2P4X model in table 1.1); nevertheless it is sufficient in most situations by virtue of the context. Due to the invariance condition, $\#\mathcal{S}$ is an even number, unless there is a species of resting particles with zero velocity.

The set of real-valued functions over $\mathcal{S}$ is a $\#\mathcal{S}$-dimensional vector space $\mathcal{F}(\mathcal{S})$ being identifiable with $\mathbb{R}^{\#\mathcal{S}}$. Its elements are denoted by sans-serif characters like f. The standard scalar product in $\mathcal{F}(\mathcal{S}) \cong \mathbb{R}^{\#\mathcal{S}}$ is indicated by brackets $\langle \cdot, \cdot \rangle$. Furthermore, $\mathsf{s}_i : \mathcal{S} \to \mathbb{R},\ 1 \le i \le d$ stands for the function assigning each velocity vector its i^{th} component; moreover $\mathsf{1}$ abbreviates the constant function such that $\mathsf{1}(\mathbf{s}) = 1$ independently of $\mathbf{s} \in \mathcal{S}$. The usually invalid operation $a + \mathsf{f}$ with $a \in \mathbb{R}$ and $\mathsf{f} \in \mathbb{R}^{\#\mathcal{S}}$ is understood in the sense $a + \mathsf{f} := a\mathsf{1} + \mathsf{f}$. Finally, for two elements $\mathsf{f}, \mathsf{g} \in \mathcal{F}(\mathcal{S})$ the product $\mathsf{fg} \in \mathcal{F}(\mathcal{S})$ is defined by component-wise multiplication, i.e. $\big(\mathsf{fg}\big)(\mathbf{s}) := \mathsf{f}(\mathbf{s})\mathsf{g}(\mathbf{s})$. Observe also the identity $\langle \mathsf{f}, \mathsf{gh} \rangle = \langle \mathsf{fg}, \mathsf{h} \rangle$.

The equivalent moment system

In the previous subsection moments were computed as integrals over the velocity space $\mathbb{R}^3$. However, in the case of a discrete velocity space, the integral is superseded by a sum over the velocities. Using the scalar product, the zeroth moment (mass moment) as well as the first and second moments are respectively given by

$$\langle \mathsf{1}, \mathsf{f} \rangle, \quad \langle \mathsf{s}_i, \mathsf{f} \rangle \quad \text{and} \quad \langle \mathsf{s}_i \mathsf{s}_j, \mathsf{f} \rangle \quad \text{with } \ i, j \in \{1, ..., d\}.$$

Similarly, higher moments are obtained.
Since $\mathcal{F}(\mathcal{S})$ is finite dimensional, the space of polynomials in the velocity components is *a fortiori* finite dimensional too (actually it is the same space). This implies that all polynomials can be expressed in terms of $\#\mathcal{S}$ linear independent polynomials forming a basis in $\mathcal{F}(\mathcal{S})$. For the sake of an example let us consider the D2P4 model (see table 1.1). The following equations are found:

$$\mathsf{s}_1^2 + \mathsf{s}_2^2 = 1, \quad \mathsf{s}_1 \mathsf{s}_2 = 0, \quad \mathsf{s}_1^4 = \mathsf{s}_1^2, \quad \mathsf{s}_2^4 = \mathsf{s}_2^2.$$

[32]This condition is necessary to introduce the bounce-back boundary rule to handle boundaries.
[33]In most literature the letter Q is used instead of P: DdQ$\#\mathcal{S}$.

A basis of $\mathcal{F}(\mathcal{S})$ is given by $\{1, \mathsf{s}_1, \mathsf{s}_2, \mathsf{s}_1^2\}$ or alternatively by $\{1, \mathsf{s}_1, \mathsf{s}_2, \mathsf{s}_1^2 - \mathsf{s}_2^2\}$. In the case of other models, different relations are found. If the D2P4X model is chosen, we get, for instance, $\mathsf{s}_1\mathsf{s}_2 \neq 0$.

Consequently, moments with respect to arbitrary polynomials (elements in $\mathcal{F}(\mathcal{S})$) can be computed as linear combinations of moments referring to a basis $\mathsf{p}_1, ..., \mathsf{p}_{\#\mathcal{S}}$ in $\mathcal{F}(\mathcal{S})$. Being a basis means that the row vectors of the matrix

$$M := \begin{pmatrix} \mathsf{p}_1(\mathsf{s}_1) & \dots & \mathsf{p}_1(\mathsf{s}_{\#\mathcal{S}}) \\ \vdots & & \vdots \\ \mathsf{p}_{\#\mathcal{S}}(\mathsf{s}_1) & \dots & \mathsf{p}_{\#\mathcal{S}}(\mathsf{s}_{\#\mathcal{S}}) \end{pmatrix}$$

form a basis in $\mathbb{R}^{\#\mathcal{S}}$. Hence the square matrix M has full rank and is invertible. So we see that the $\#\mathcal{S}$-tuple of moments $m := M\mathsf{f}$ contains the same information as the population function since m is converted into f by $M^{-1}m = \mathsf{f}$.

The circumstance that the population function f can be recovered from a (finite) set of moments is a fundamental difference to the distribution function f of the Boltzmann equation, which originates, of course, from the finite velocity set. In contrast, f cannot be reconstructed uniquely by the knowledge of a finite number of moments. The existence of an infinite number of moments needs additional assumptions on f and even then, the reconstruction remains a rather delicate question (moment problem).

It is possible to transform the finite velocity Boltzmann equation[34]

$$\partial_t \mathsf{f} + \sum_{i=1}^{d} S_i \partial_{x_i} \mathsf{f} = J(\mathsf{f}) \tag{1.16}$$

into an evolution equation for the moment tuple m. Setting $\mathsf{f} = M^{-1}m$ and multiplying the equation by M from the left yields

$$\partial_t m + \sum_{i=1}^{d} M S_i M^{-1} \partial_{x_i} m = M J(M^{-1}m). \tag{1.17}$$

This system is known as the *equivalent moment system*. Generally, it is less decoupled and therefore more complicated than (1.16) because the matrices MS_iM^{-1} are full. On the other hand (1.17) tells much about the macroscopic behavior (physical interpretation) if the moments are well chosen[35].

Lattice-Boltzmann collision term

Lattice-Boltzmann equations represent a special class of discrete velocity Boltzmann equations. Their collision operator is of *relaxation type* tending to relax the

[34]Observe that we could also write $\sum_{i=1}^{d} \mathsf{s}_i \partial_{x_i} \mathsf{f}$ replacing thus the diagonal matrix S_i by $\mathsf{s}_i \in \mathbb{R}^{\#\mathcal{S}}$. Note that the diagonal matrix S_i acts as linear operator $\mathcal{F}(\mathcal{S}) \to \mathcal{F}(\mathcal{S})$ which can be equally represented by a multiplication in $\mathcal{F}(\mathcal{S})$.

[35]cf. the paragraph: The kinetic approach – presentation of an example on page 32

population function f towards a local equilibrium $E(\mathsf{f})$.

$$J(\mathsf{f}) = A\big(E(\mathsf{f}) - \mathsf{f}\big) \qquad \begin{cases} A = \frac{1}{\tau} I & \text{BGK} \\ A = M^{-1}\Lambda M & \text{MRT} \end{cases}$$

The most simple form of J is given by the BGK operator. The collision(relaxation) time τ determines the significance of the collision term; the smaller τ the more dominant it becomes.
It turned out by numerical experiments that better results are obtained by a *multiple relaxation time* (MRT) ansatz (cf. [13]). If the lattice-Boltzmann equation with the BGK collision is transformed into its equivalent moment system,

$$\partial_t m + \sum_{i=1}^{\#\mathcal{S}} M S_i M^{-1} \partial_{x_i} m \;=\; \underbrace{\tau^{-1} I}_{\text{repl. by } \Lambda} M\big(E(M^{-1}m) - M^{-1}m\big)$$

it may be reasonable to assign different relaxation times for certain (groups of) moments. So $\tau^{-1}I$ is replaced by a general diagonal matrix Λ with positive entries. Back-transformation of the equivalent moment system leads to the matrix $A = M^{-1}\Lambda M$.

The equilibrium is the core part of the collision operator, because it defines how the populations are coupled with each other. This influences the macroscopic behavior of the artificial particle dynamics and thus the (macroscopic) target equation that is to be described by the kinetic approach. Generally, the equilibrium is given as a function of certain moments and the discrete velocities. These moments are related to quantities figuring as central variables in the target equation. For concrete examples table 1.1 can be consulted.

Lattice-Boltzmann velocity models

Table 1.1 compiles in the leftmost column some important velocity models in one and two space dimensions. All presented 2D models are compatible with quadratic grids. The second column lists the designation that we use to refer to them.

The D1P2 model has only two velocities $(-1, 1)$. Adding a rest population yields the D1P3 model, which can be used with a weight function parameterized by $\theta > 1$. For $\theta = 1$ the D1P3 model passes over to the D1P2 model. In analogy to the D1P2 model, the D2P4 and D2P4X model can also be equipped with a rest population resulting into D2P5 and D2P5X model respectively. A combination of the D2P4 and D2P4X model leads to the D2P8 model, which becomes the D2P9 model if a rest population is added.
One might be tempted to assume that 2D models are build up by 1D ones. So the D2P4 'velocity star' is undoubtedly obtained by crossing two D1P2 'dumbbells'. Indeed, 1D and 2D models are connected with each other if certain symmetry conditions are satisfied (see next section).

The fourth column of the table contains a suitable equilibrium for each model. $\mathcal{E}$ is given as function of macroscopic quantities related to moments of the population

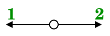
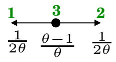
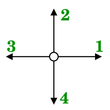
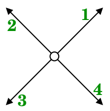
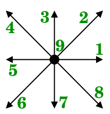
	D1P2	$\nu = \tau$	$\mathcal{E}(u;s) \;=\; \frac{1}{2}u\,(1+\epsilon a s - \epsilon^2\,\tau\, c)$	$\mathsf{q}(s) \;=\; \frac{1}{2}\,q$
	D1P3	$\nu = \frac{\tau}{\theta}$	$\mathcal{E}(u;s) \;=\; \mathsf{w}(s)u\,(1+\epsilon\,\theta a s - \epsilon^2\,\tau\, c)$	$\mathsf{q}(s) \;=\; \mathsf{w}(s)\;q$
	D2P4	$\nu = \frac{\tau}{2}$	$\mathcal{E}(u;\mathbf{s}) \;=\; \frac{1}{4}u\,(1+2\epsilon\,\mathbf{s}\cdot\mathbf{a} - \epsilon^2\,\tau\, c)$	$\mathsf{q}(\mathbf{s}) \;=\; \frac{1}{4}\,q$
	D2P4X	$\nu = \tau$	$\mathcal{E}(u;\mathbf{s}) \;=\; \frac{1}{4}u\,(1+\epsilon\,\mathbf{s}\cdot\mathbf{a} - \epsilon^2\,\tau\, c)$	$\mathsf{q}(\mathbf{s}) \;=\; \frac{1}{4}\,q$
	D2P9	$\nu = \frac{\tau}{3}$	$\mathcal{E}(p,\mathbf{u};\mathbf{s}) = \;3\,\mathsf{w}(\mathbf{s})\mathbf{s}\cdot\mathbf{u}\;+\;\ldots$ $+\,\mathsf{w}(\mathbf{s})\,\epsilon\,\big[3p{+}9(\mathbf{s}{\cdot}\mathbf{a})\,(\mathbf{s}{\cdot}\mathbf{u}){-}3\mathbf{a}{\cdot}\mathbf{u}\big]$	$\mathsf{q}(\mathbf{s}) \;=\; 3\mathsf{w}(\mathbf{s})\,\mathbf{s}\cdot\mathbf{q}$

Table 1.1: Overview of some popular velocity models being also called 'velocity stars'. The equilibrium operator $E(\mathsf{f})$ occuring in the lattice-Boltzmann equation is in the first four cases (from above) explicitly given by $E(\mathsf{f}) := \mathcal{E}\big(\langle 1,\mathsf{f}\rangle;\cdot\big) \in \mathcal{F}(\mathcal{S})$, while in the last case it becomes $E(\mathsf{f}) := \mathcal{E}\Big(\frac{1}{3\epsilon}\langle 1,\mathsf{f}\rangle,\, \binom{\langle \mathsf{s}_x,\mathsf{f}\rangle}{\langle \mathsf{s}_y,\mathsf{f}\rangle};\cdot\Big) \in \mathcal{F}(\mathcal{S})$.

function $(u, p, \mathbf{u})$ and the discrete velocities $(s, \mathbf{s})$. In particular the dependence on the scaling parameter ϵ should be noted.
In conjunction with the lattice-Boltzmann equation in *parabolic scaling*

$$\partial_t \mathsf{f} + \epsilon^{-1}\mathbf{s}\cdot\nabla\mathsf{f} = \epsilon^{-2}\omega\big(E(\mathsf{f}) - \mathsf{f}\big) + \mathsf{q}$$

the following target equations are obtained:

$$\begin{array}{rcll}
\partial_t u \;+\; \partial_x(au - \nu\partial_x u) \;+\; cu & = & q & \text{D1P2, D1P3} \\
\partial_t u \;+\; \nabla\cdot(\mathbf{a}u - \nu\nabla u) \;+\; cu & = & q & \text{D2P4, D2P4X} \\
\partial_t \mathbf{u} \;+\; \mathbf{a}\cdot\nabla\mathbf{u} \;-\; \nu\Delta\mathbf{u} & = & -\nabla p + \mathbf{q}\,, \quad \nabla\cdot\mathbf{u} = 0 & \text{D2P9}
\end{array}$$

The first and second target equation respectively represent a 1D and 2D diffusion advection equation with a linear reactive term cu. The variable u may be interpreted as density or concentration. It is associated with the mass moment $\langle 1, \mathsf{f}\rangle$, that converges to the solution of the corresponding target equation as ϵ tends to 0.

The D2P9 model is complex enough to set up an equilibrium function leading to the Oseen equation[36]. In comparison to the scalar diffusion-advection equation the Oseen equation represents a system comprising the incompressibility condition and two evolution equations for both of the velocity components. For $\mathbf{a} = 0$ the Stokes equation is obtained. In this context the zeroth moment $\langle 1, \mathsf{f}\rangle$ is associated with the

[36] The Oseen equation is obtained as a linearization of the Navier-Stokes equation. So the nonlinear advective term $\mathbf{v}\cdot\nabla\mathbf{v}$ is substituted by $(\mathbf{a}{+}\mathbf{u})\cdot\nabla(\mathbf{a}{+}\mathbf{u}) = \mathbf{a}\cdot\nabla\mathbf{a}{+}\mathbf{a}\cdot\nabla\mathbf{u}{+}\mathbf{u}\cdot\nabla\mathbf{a}{+}\mathbf{u}\cdot\nabla\mathbf{u}$. Here $\mathbf{u}$ is considered as a small disturbance superposing the given velocity field $\mathbf{a}$. Therefore the term quadratic in $\mathbf{u}$ is neglected. Moreover, if $\mathbf{a}$ is constant, then $\mathbf{a}\cdot\nabla\mathbf{u}$ survives alone among the four addends.

pressure p whereas the first moments $\langle \mathsf{s}_x, \mathsf{f}\rangle, \langle \mathsf{s}_y, \mathsf{f}\rangle$ are connected to the respective components of the velocity $\mathbf{u}$.

The correspondence between sources of the target equations $q, \mathbf{q}$ and the pertaining source terms of the lattice-Boltzmann equation is provided by the rightmost column of table 1.1. Finally the third column summarizes the relation between the relaxation time τ and the diffusivity or viscosity ν. In the case of the lattice-Boltzmann algorithm, the relation is modified in so far that τ must be replaced by $\tau - \frac{1}{2}$ which requires $\tau > \frac{1}{2}$ to obtain a positive ν (otherwise the target equation becomes ill-posed).

The kinetic approach – presentation of an example

In the previous subsection it was shown that the Boltzmann equation is related to the compressible Euler equation of fluid dynamics being the prototype of a non-scalar conservation law. In this paragraph we want to illuminate by a simple example how the equilibrium function for a discrete velocity Boltzmann equation is constructed in dependence of the target equation and why the (hyperbolic) scaling comes into play. Furthermore we want to discover the affinity between the kinetic formulation (using a discrete velocity Boltzmann equation) and a relaxation approach.

Let us consider a one-dimensional conservation law that is given by the subsequent evolution equation for some conserved macroscopic quantity ρ:

$$\partial_t \rho + f'(\rho)\partial_x \rho = 0 \quad \text{or alternatively} \quad \partial_t \rho + \partial_x\big(f(\rho)\big) = 0. \tag{1.18}$$

The second equation (the so-called conservative form) fits exactly to the type of equations being accessible by a kinetic approach. Now the following steps are to be done:

- decide upon a discrete velocity model,
- associate the occurring macroscopic quantity ρ with a moment of the population function f, which should approximate ρ,
- construct a suitable equilibrium function depending on the chosen moment.

Here we select the D1P3 model (see table 1.1) endowed with the weight function

$$\mathsf{w}(-1) = \tfrac{1}{2\theta}, \qquad \mathsf{w}(0) = \tfrac{\theta-1}{\theta}, \qquad \mathsf{w}(1) = \tfrac{1}{2\theta} \quad \text{where } \theta > 1.$$

ρ shall be approximated by the mass moment $\varrho := \langle 1, \mathsf{f}\rangle$. The equilibrium operator should be represented by a function of ϱ such that $E(\mathsf{f}) = \mathcal{E}(\varrho) \in \mathcal{F}(\mathcal{S})$. For a heuristic reasoning let us assume that under certain circumstances the population function is close to the equilibrium, i.e. $\mathsf{f} \approx \mathcal{E}(\varrho)$. Plugging $\mathcal{E}(\varrho)$ for f into the lattice-Boltzmann equation

$$\partial_t \mathsf{f} + \mathsf{s}\partial_x \mathsf{f} = \tfrac{1}{\tau}\big(E(\mathsf{f}) - \mathsf{f}\big) \tag{1.19}$$

and taking the mass moment yields

$$\partial_t\big\langle 1,\mathcal{E}(\varrho)\big\rangle + \partial_x\big\langle \mathsf{s},\mathcal{E}(\varrho)\big\rangle \;=\; \tfrac{1}{\tau}\Big\langle 1, E\big(\mathcal{E}(\varrho)\big) - \mathcal{E}(\varrho)\Big\rangle.$$

To connect (1.19) with the conservation law (1.18), we require this equation for ϱ to be equal to (1.18). This provides the following conditions for $\mathcal{E}$:

$$\big\langle 1,\mathcal{E}(\varrho)\big\rangle = \varrho, \qquad \big\langle \mathsf{s},\mathcal{E}(\varrho)\big\rangle = f(\varrho). \tag{1.20}$$

The first condition also assures the vanishing of the right hand side since by definition of the equilibrium operator $E(\mathsf{f}) = \mathcal{E}\big(\langle 1,\mathsf{f}\rangle\big)$ and hence

$$E\big(\mathcal{E}(\varrho)\big) = \mathcal{E}\big(\langle 1,\mathcal{E}(\varrho)\rangle\big) = \mathcal{E}(\varrho)\,.$$

In this situation an appropriate equilibrium function can be guessed[37]:

$$\mathcal{E}(\varrho) = \varrho\mathsf{w} + \theta f(\varrho)\mathsf{sw}.$$

Both conditions in (1.20) are immediately verified for $\mathcal{E}$ using the *structure relations* $\langle 1,\mathsf{w}\rangle = 1$ and $\langle 1,\mathsf{sw}\rangle = \langle \mathsf{s},\mathsf{w}\rangle = 0$.

In order to set up an equivalent moment system we take the monomial basis $\{1,\mathsf{s},\mathsf{s}^2\}$ of $\mathcal{F}(\mathcal{S})$ and introduce two further moments besides ϱ:

$$\langle 1,\mathsf{f}\rangle = \varrho, \qquad \langle \mathsf{s},\mathsf{f}\rangle =: \phi, \qquad \langle \mathsf{s}^2,\mathsf{f}\rangle =: \chi.$$

So we are lead to the following equivalent moment system:

$$\begin{aligned}
&\partial_t\varrho + \partial_x\phi = 0,\\
&\partial_t\phi + \partial_x\chi = \tfrac{1}{\tau}\big(f(\varrho) - \phi\big),\\
&\partial_t\chi + \partial_x\phi = \tfrac{1}{\tau}\big(\tfrac{1}{\theta}\varrho - \chi\big).
\end{aligned}$$

Setting $\phi = f(\varrho)$ in the first equation would just result in the conservation law for ϱ. However, according to the second equation, which can be solved for ϕ, we find $\phi = f(\varrho) - \tau(\partial_t\phi + \partial_x\chi)$. Thus we get $\phi \approx f(\varrho)$ if $\tau(\partial_t\phi + \partial_x\chi)$ is small. This might be enforced by introducing a scaling parameter $\epsilon > 0$ and setting $\phi = f(\varrho) - \epsilon\tau(\partial_t\phi + \partial_x\chi)$. Proceeding similarly with the third equation (in the first equation nothing has to be done) yields the scaled equivalent moment system

$$\begin{aligned}
&\partial_t\varrho + \partial_x\phi = 0,\\
&\partial_t\phi + \partial_x\chi = \tfrac{1}{\epsilon\tau}\big(f(\varrho) - \phi\big),\\
&\partial_t\chi + \partial_x\phi = \tfrac{1}{\epsilon\tau}\big(\tfrac{1}{\theta}\varrho - \chi\big),
\end{aligned}$$

pertaining to

$$\partial_t\mathsf{f} + \mathsf{s}\partial_x\mathsf{f} = \tfrac{1}{\epsilon\tau}\big(E(\mathsf{f}) - \mathsf{f}\big). \tag{1.21}$$

[37] The coefficients of $\mathcal{E}(\varrho)$ (represented in a given basis, e.g. with respect to $\{\mathsf{w},\mathsf{ws}\}$) are determined by solving a linear system.

This is the lattice-Boltzmann equation in the so-called *hyperbolic scaling*. Its solution f as well as the moments ϱ, ϕ and χ depend on the parameter ϵ. Assuming that they remain bounded if ϵ tends to zero, the above reasoning shows that ϱ solves the conservation law (1.18) with a small extra-term vanishing with ϵ. Under further assumptions (e.g. equal initial conditions, uniqueness of the solution etc.) it becomes very plausible that ϱ converges, indeed, for $\epsilon \downarrow 0$ to the solution ρ of the conservation law.

There is another issue that we have not yet touched. Why should one be interested in this limit?
It seems as if one problem (solving (1.18)) has been replaced by another problem (solving (1.21)). Moreover, (1.21) involves three variables – the three population functions contained in f – while (1.18) has only one unknown. Thus (1.21) written in components[38]

$$\begin{aligned}
\partial_t \mathsf{f}_{-1} \;-\; \partial_x \mathsf{f}_{-1} &= \tfrac{1}{\epsilon\tau}\big(\mathcal{E}(\langle 1, \mathsf{f}\rangle) - \mathsf{f}_{-1}\big)\\
\partial_t \mathsf{f}_{0} \qquad\quad &= \tfrac{1}{\epsilon\tau}\big(\mathcal{E}(\langle 1, \mathsf{f}\rangle) - \mathsf{f}_{0}\big)\\
\partial_t \mathsf{f}_{+1} \;+\; \partial_x \mathsf{f}_{+1} &= \tfrac{1}{\epsilon\tau}\big(\mathcal{E}(\langle 1, \mathsf{f}\rangle) - \mathsf{f}_{+1}\big),
\end{aligned}$$

appears even more complicated than the original problem. This, however, is not really true, if the structure of the equations is carefully taken into account. Noticing that the nonlinearity of the conservation law is hidden inside the equilibrium operator, the nonlinear term $f'(\rho)\partial_x\rho$ is modified in so far, that it does no more contain the derivative of the unknown. Furthermore the left hand side is purely linear and of particular simplicity due to the decoupling of the populations. This makes the kinetic formulation of the conservation law by the lattice-Boltzmann equation quite attractive for numerical methods. The price for this structural advantage consists in an inflation of the number of unknowns (here by the factor 3). Nevertheless lattice-Boltzmann methods have gained significant importance for more complicated equations (fluid dynamics) than for the one-dimensional conservation law.

Relaxation methods due to [29] are based on the same idea: to 'buy' a simplification at the cost of an increasing number of variables. Thus a second variable, the flux $\varphi := f(\rho)$, is introduced. In order to obtain an evolution equation for both variables a slight perturbation is added:

$$\left.\begin{aligned}\partial_t\rho + \partial_x\varphi &= 0\\ \varphi &= f(\rho)\end{aligned}\right\} \quad\Rightarrow\quad \left\{\begin{aligned}\partial_t\varrho + \partial_x\phi \quad &= 0\\ \underbrace{\epsilon(\partial_t\phi + \partial_x\varrho)}_{\text{perturbation}} &= f(\varrho) - \phi.\end{aligned}\right.$$

Obviously derivatives and variables are arranged crosswise. To simplify the left hand side further, one may ask for a transformation separating the variables. For this, let us write the system in the following way

$$\underbrace{\begin{pmatrix}\partial_t & \partial_x\\ \partial_x & \partial_t\end{pmatrix}}_{=:D}\binom{\varrho}{\phi} = \tfrac{1}{\epsilon}\,\mathrm{RHS}\left(\binom{\varrho}{\phi}\right) \quad \text{with} \quad \mathrm{RHS}\left(\binom{\varrho}{\phi}\right) = \binom{0}{f(\varrho)-\phi}$$

[38]Here we use the three discrete velocities of $\mathcal{S} = \{-1, 0, 1\}$ as indices.

So we seek a linear variable transformation $\binom{\rho}{\phi} = T\binom{\mathsf{f}_{-1}}{\mathsf{f}_{+1}}$ with an invertible 2×2 matrix T and another invertible 2×2-matrix R such that the transformed system

$$R\,D\,T\,\mathsf{f} = \tfrac{1}{\epsilon}R\ \ \mathrm{RHS}(T\mathsf{f}) \tag{1.22}$$

becomes diagonalized on the left hand side. Now, D is a symmetric matrix implying its diagonalizibility. The eigenvectors of an arbitrary symmetric 2×2 matrix are $\binom{1}{1}$ and $\binom{-1}{1}$. Setting

$$T = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \qquad \text{and} \qquad R = T^{-1} = \tfrac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

the transformed system (1.22) becomes finally

$$\begin{aligned} \partial_t \mathsf{f}_{-1} - \partial_x \mathsf{f}_{-1} &= \tfrac{1}{\epsilon}\left[\tfrac{1}{2}(\mathsf{f}_{-1} + \mathsf{f}_{+1}) - \tfrac{1}{2}f(\mathsf{f}_{-1} + \mathsf{f}_{+1}) - \mathsf{f}_{-1}\right], \\ \partial_t \mathsf{f}_{+1} + \partial_x \mathsf{f}_{+1} &= \tfrac{1}{\epsilon}\left[\tfrac{1}{2}(\mathsf{f}_{-1} + \mathsf{f}_{+1}) + \tfrac{1}{2}f(\mathsf{f}_{-1} + \mathsf{f}_{+1}) - \mathsf{f}_{+1}\right]. \end{aligned}$$

If we define the equilibrium function $\mathcal{E}(\varrho) := \frac{1}{2}\varrho + \frac{1}{2}f(\varrho)\mathsf{s}$ with $\varrho = \mathsf{f}_{-1} + \mathsf{f}_{+1} = \langle 1, \mathsf{f}\rangle$, we just recover the BGK lattice-Boltzmann equation (with $\tau = 1$) referring to the D1P2 model and the weight function $\mathsf{w}(-1) = \mathsf{w}(1) = \frac{1}{2}$.
It seems as though we spared the introduction of an additional third variable with respect to the kinetic approach. However, this impression is deceiving since we could have started directly with the D1P2 model. The additional flexibility of the D1P3 model is optimally used for the kinetic formulation of a system of two conservation laws (e.g. wave equation for sound, 1D waves in compressible media). It should be noted that the D1P2 model ensues from the D1P3 model in the limit of $\theta = 1$.

In summary, the example teaches that the kinetic ansatz and the relaxation ansatz appear as two sides of the same coin. Both give rise to the same (approximate) reformulation of the conservation law. The lattice-Boltzmann equation may be interpreted as a particularly clever way of writing down a relaxation system. But this presupposes also a special structure of the original equation which might not be always available. Therefore the kinetic approach is subsumable into the larger field of relaxation methods.

The kinetic approach is extendible to balance laws generalizing the class of conservation laws. Thus, dissipative terms can be incorporated too. This requires, however, the *parabolic scaling* (compare the relation between the Boltzmann equation on the one hand and the compressible Euler or the incompressible Navier Stokes equation on the other hand, shortly described at the end of the previous subsection).

In general, the more complex the desired target equation the more complex the velocity model must be chosen to gain enough algebraic flexibility. Reference [31] deals with the systematic construction of equilibrium functions for discrete velocity Boltzmann equations.

Although two-dimensional velocity models are not further treated in this thesis, we exemplarily discuss the calculation of the structure relations for the D2P9 model in the next paragraph. Their knowledge is fundamental for the asymptotic analysis of lattice-Boltzmann algorithms, which uses Taylor expansions and regular asymptotic expansions as additional analytic tools.

Structure relations of discrete velocity models

For the design and analysis of lattice-Boltzmann methods the *structure relations* of the underlying velocity model are important to know. These are given as tensorial products of the discrete velocity vectors:

$$R_n := \sum_{\mathsf{s}\in\mathcal{S}} \underbrace{\mathsf{s}\otimes\ldots\otimes\mathsf{s}}_{n \text{ times}} \qquad T_n := \sum_{\mathsf{s}\in\mathcal{S}} \mathsf{w}(\mathsf{s})\ \underbrace{\mathsf{s}\otimes\ldots\otimes\mathsf{s}}_{n \text{ times}}. \tag{1.23}$$

The *weight function* $\mathsf{w}\in\mathcal{F}(\mathcal{S})$ occurring in T_n corresponds to the equilibrium in the 'state of rest', which is a discrete analogon to the normalized Maxwellian with zero bulk velocity. It should be noted that w permits also an interpretation from the perspective of finite differences[39]. In the designing process of lattice-Boltzmann equations w must be appropriately chosen together with the underlying velocity model and the general equilibrium.

Exemplarily, we want to compute the structure relations T_n for the D2P9 model being the most complicated one in table 1.1. For fluid dynamic applications (Navier-Stokes) it is usually combined with the following weight function:

$$\begin{aligned}
\text{rest velocity:} &\quad \mathsf{w}\left(\tbinom{0}{0}\right) = \tfrac{4}{9}\\
\text{axis-parallel velocities:} &\quad \mathsf{w}\left(\tbinom{1}{0}\right) = \mathsf{w}\big(\tbinom{0}{1}\big) = \mathsf{w}\big(\tbinom{-1}{0}\big) = \mathsf{w}\big(\tbinom{0}{-1}\big) = \tfrac{1}{9}\\
\text{diagonal velocities:} &\quad \mathsf{w}\left(\tbinom{1}{1}\right) = \mathsf{w}\big(\tbinom{-1}{1}\big) = \mathsf{w}\big(\tbinom{-1}{-1}\big) = \mathsf{w}\big(\tbinom{1}{-1}\big) = \tfrac{1}{36}
\end{aligned}$$

Observe that by reason of *isotropy* (no preferred directions) $\mathsf{w}(\mathsf{s}) = \mathsf{w}(-\mathsf{s})$. Moreover, w depends only on the speed, i.e. on the norm of the velocity vectors.
The components of the tensor can be computed as scalar products

$$T_{n;\alpha_1\ldots\alpha_n} = \langle \mathsf{w}, \mathsf{s}_{\alpha_1}\ldots\mathsf{s}_{\alpha_n}\rangle, \tag{1.24}$$

which are just the moments of the weight function w. Generally, the commutativity of the product in $\mathcal{F}(\mathcal{S})$ (resulting from the commutativity of multiplication in $\mathbb{R}$) entails that the numbers $T_{n;\alpha_1\ldots\alpha_n}$ remain equal if their indices are permuted. Moreover, the tensor components are invariant with respect to a cyclic change of their indices. In the case of the two-dimensional D2P9 model this means that every index equal to $1(x)$ is set to $2(y)$ and vice versa. The justification of this isotropic property is a little bit less obvious and a proof is given below.

Thanks to these observations, it is sufficient to verify the subsequent relations[40]

$$\begin{gathered}
\langle \mathsf{w}, \mathsf{s}_\alpha\rangle = 0 \quad \langle \mathsf{w}, \mathsf{s}_\alpha\mathsf{s}_\beta\rangle = \tfrac{1}{3}\delta_{\alpha\beta} \quad \langle \mathsf{w}, \mathsf{s}_\alpha\mathsf{s}_\beta\mathsf{s}_\gamma\rangle = 0\\
\langle \mathsf{w}, \mathsf{s}_\alpha\mathsf{s}_\beta\mathsf{s}_\gamma\mathsf{s}_\delta\rangle = \tfrac{1}{9}\left(\delta_{\alpha\beta}\delta_{\gamma\delta} + \delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\delta}\right)
\end{gathered} \tag{1.25}$$

by checking only a few index combinations and by testing the right hand side with respect to the index symmetry relations of T_n, which are obviously satisfied. It is easily seen that

$$T_{1;x}, \quad T_{2;xx}, \quad T_{2;xy}, \quad T_{3;xxx}, \quad T_{3;xxy}, \quad T_{4;xxxx}, \quad T_{4;xxxy}, \quad T_{4;xxyy}$$

agree with the corresponding right hand side. Equation (1.25) gives rise to the assumption that all tensors of odd rank $T_1, T_3, T_5, \ldots$ must vanish. In fact, this is a consequence of the invariance of $\mathcal{S}$ under reflection.

[39] If we consider the velocity set as stencil of a difference operator, it turns out that the weight function is related to the coefficients of a stencil discretizing the Laplacian. An interesting link between the kinetic and numeric perspective is that the fundamental solution of the heat equation (containing the Laplacian) is an exponential function of the same kind as the Maxwellian.

[40] $\delta_{\alpha\beta}$ denotes the Kronecker delta, being 1 for $\alpha = \beta$ and 0 otherwise.

Proposition 1.1. *Let $\mathcal{S}$ be a two-dimensional velocity model satisfying $\mathcal{S} = -\mathcal{S}$ and suppose $\mathsf{w}(\mathbf{s}) = \mathsf{w}(-\mathbf{s})\ \forall \mathbf{s} \in \mathcal{S}$. If $n \in \mathbb{N}$ is odd, the tensors R_n and T_n defined in* (1.23) *vanish.*

Proof: The assertion is shown for R_n. The invariance of the tensor elements under permutation of their indices implies the existence of $i, j \in \mathbb{N}_0$ with $i + j = n$ such that

$$R_{n;\alpha_1,\ldots,\alpha_n} = \langle \mathbf{1}, \mathsf{s}_x^i \mathsf{s}_y^j \rangle = \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{s}_x^i(\mathbf{s})\, \mathsf{s}_y^j(\mathbf{s}).$$

Since n is odd, either i or j is odd. Let us assume without restriction that i is even while j is odd. Setting

$$\mathsf{e}(\mathbf{s}) := \mathsf{s}_x^i(\mathbf{s}) \quad \text{and} \quad \mathsf{o}(\mathbf{s}) := \mathsf{s}_y^j(\mathbf{s})$$

and using the fact that $\mathsf{s}_\alpha(\mathbf{s}) = -\mathsf{s}(-\mathbf{s})$ (the negative component is the component of the negative vector) we infer that $\mathsf{e}(\mathbf{s}) = \mathsf{e}(-\mathbf{s})$ and $\mathsf{o}(\mathbf{s}) = -\mathsf{o}(-\mathbf{s})$ because $(-1)^i = 1$ and $(-1)^j = -1$. Using this and $\mathcal{S} = -\mathcal{S}$ the claim can be verified:

$$\begin{aligned}
R_{n;\alpha_1,\ldots,\alpha_n} = \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{e}(\mathbf{s})\mathsf{o}(\mathbf{s}) &= \frac{1}{2}\Big[\sum_{\mathbf{s}\in\mathcal{S}} \mathsf{e}(\mathbf{s})\mathsf{o}(\mathbf{s}) + \mathsf{e}(\mathbf{s})\mathsf{o}(\mathbf{s})\Big] \\
&= \frac{1}{2}\Big[\sum_{\mathbf{s}\in\mathcal{S}} \mathsf{e}(\mathbf{s})\mathsf{o}(\mathbf{s}) - \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{e}(\mathbf{s})\mathsf{o}(-\mathbf{s})\Big] \\
&= \frac{1}{2}\Big[\sum_{\mathbf{s}\in\mathcal{S}} \mathsf{e}(\mathbf{s})\mathsf{o}(\mathbf{s}) - \sum_{\mathbf{r}\in-\mathcal{S}} \mathsf{e}(-\mathbf{r})\mathsf{o}(\mathbf{r})\Big] \\
&= \frac{1}{2}\Big[\sum_{\mathbf{s}\in\mathcal{S}} \mathsf{e}(\mathbf{s})\mathsf{o}(\mathbf{s}) - \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{e}(\mathbf{s})\mathsf{o}(\mathbf{s})\Big] \\
&= 0.
\end{aligned}$$

Observe that we have renamed the summation variable $\mathbf{s}$ (as being a dummy variable) in the third and fourth line of the above computation. Hence all tensor components of R_n vanish. Using the hypothesis on w the proof proceeds analogously for T_n. ■

To formulate the next proposition we define the rotation matrix

$$D := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

describing a rotation of $\frac{\pi}{2}$ around the origin. From $D\binom{a}{b} = \binom{-b}{a}$ follow the relations

$$\mathsf{s}_x(\mathbf{s}) = \mathsf{s}_y(D\mathbf{s}) \qquad \text{and} \qquad \mathsf{s}_y(\mathbf{s}) = -\mathsf{s}_x(D\mathbf{s}), \tag{1.26}$$

provided $\mathcal{S}$ is invariant under 90°-rotations, i.e. $\mathcal{S} = D\mathcal{S}$. Note that this relation ensures $\mathcal{S} = -\mathcal{S}$ as $D^2 = -I$.

Proposition 1.2. *Suppose the two-dimensional velocity model $\mathcal{S}$ and the weight function satisfy $\mathcal{S} = D\mathcal{S}$ and $\mathsf{w}(D\mathbf{s}) = \mathsf{w}(\mathbf{s})\ \forall \mathbf{s} \in \mathcal{S}$ respectively. Then the moduli of the components pertaining to the tensors defined in* (1.23) *are invariant with respect to a cyclic change of the indices.*

Proof: Let $T_{n,\alpha_1,\ldots,\alpha_n}$ and $T_{n,\beta_1,\ldots,\beta_n}$ two tensor elements such that $\beta_k = y$ if $\alpha_k = x$ or $\beta_k = x$ if $\alpha_k = y$ for $k \in \{1, ..., n\}$. We have to show

$$|T_{n,\alpha_1,\ldots,\alpha_n}| = |T_{n,\beta_1,\ldots,\beta_n}|. \tag{1.27}$$

Thanks to the invariance under permutation of indices, both of the tensor elements permit the subsequent representation with some $i, j \in \mathbb{N}_0, i + j = n$:

$$T_{n,\alpha_1,\ldots,\alpha_n} = \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\, \mathsf{s}_x^i(\mathbf{s})\, \mathsf{s}_y^j(\mathbf{s}) \qquad T_{n,\beta_1,\ldots,\beta_n} = \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\, \mathsf{s}_x^j(\mathbf{s})\, \mathsf{s}_y^i(\mathbf{s}).$$

Starting with $T_{n,\alpha_1,\ldots,\alpha_n}$, applying (1.26) and employing the hypothesis we obtain:

$$\begin{aligned}
\sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\mathsf{s}_x^i(\mathbf{s})\mathsf{s}_y^j(\mathbf{s}) &= \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\mathsf{s}_y^i(D\mathbf{s})(-1)^j \mathsf{s}_x^j(D\mathbf{s}) \\
&= (-1)^j \sum_{\mathbf{r}\in D\mathcal{S}} \mathsf{w}(D^{-1}\mathbf{r})\mathsf{s}_y^i(\mathbf{r})\mathsf{s}_x^j(\mathbf{r}) \\
&= (-1)^j \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\mathsf{s}_y^i(\mathbf{s})\mathsf{s}_x^j(\mathbf{s})
\end{aligned} \tag{1.28}$$

The last line corresponds to $T_{n,\beta_1,\ldots,\beta_n}$ up to the sign. ■

In the case of the D2P9 model proposition 1.2 does not only hold with respect to the moduli but for the tensor elements themselves. The reason for this lies in the circumstance that the model contains only velocities parallel either to the axes or to the diagonals. In particular this entails the relations

$$\begin{array}{ll} \mathsf{s}_x^{2n+1}(\mathbf{s}) = \mathsf{q}^{2n}(\mathbf{s})\mathsf{s}_x(\mathbf{s}) & \mathsf{s}_x^{2n+2}(\mathbf{s}) = \mathsf{q}^{2n}(\mathbf{s})\mathsf{s}_x^2(\mathbf{s}) \\ \mathsf{s}_y^{2n+1}(\mathbf{s}) = \mathsf{q}^{2n}(\mathbf{s})\mathsf{s}_y(\mathbf{s}) & \mathsf{s}_y^{2n+2}(\mathbf{s}) = \mathsf{q}^{2n}(\mathbf{s})\mathsf{s}_y^2(\mathbf{s}) \end{array}$$

with $\mathsf{q}(\mathbf{s}) := \max\big\{|\mathsf{s}_x(\mathbf{s})|, |\mathsf{s}_y(\mathbf{s})|\big\}$. From this the equation

$$\mathsf{s}_x^{2m+1}(\mathbf{s})\,\mathsf{s}_y^{2n+1}(\mathbf{s}) \;+\; \mathsf{s}_x^{2m+1}(D\mathbf{s})\,\mathsf{s}_y^{2n+1}(D\mathbf{s}) \stackrel{!}{=} 0 \tag{1.29}$$

ensues, being equivalent to

$$\Leftrightarrow \quad \mathsf{q}^{2m+2n}(\mathbf{s})\,\mathsf{s}_x(\mathbf{s})\,\mathsf{s}_y(\mathbf{s}) \;+\; \mathsf{q}^{2m+2n}(D\mathbf{s})\,\mathsf{s}_x(D\mathbf{s})\,\mathsf{s}_y(D\mathbf{s}) = 0$$

$$\Leftrightarrow \quad \mathsf{q}^{2m+2n}(\mathbf{s})\,\mathsf{s}_x(\mathbf{s})\,\mathsf{s}_y(\mathbf{s}) \;-\; \mathsf{q}^{2m+2n}(\mathbf{s})\,\mathsf{s}_y(\mathbf{s})\,\mathsf{s}_x(\mathbf{s}) = 0 \qquad \text{(true)}$$

by the above relations. So we can finally conclude that $T_{n,\alpha_1,\ldots,\alpha_n}$ and $T_{n,\beta_1,\ldots,\beta_n}$ are zero if i and j are both odd (note that this enforces n to be even!) because:

$$\begin{aligned} \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\mathsf{s}_x^i(\mathbf{s})\mathsf{s}_y^j(\mathbf{s}) &= \frac{1}{2}\sum_{\mathbf{s}\in\mathcal{S}} \Big[\mathsf{w}(\mathbf{s})\mathsf{s}_x^i(\mathbf{s})\mathsf{s}_y^j(\mathbf{s}) + \mathsf{w}(D\mathbf{s})\mathsf{s}_x^i(D\mathbf{s})\mathsf{s}_y^j(D\mathbf{s})\Big] \\ &= \frac{1}{2}\sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\Big[\mathsf{s}_x^i(\mathbf{s})\mathsf{s}_y^j(\mathbf{s}) + \mathsf{s}_x^i(D\mathbf{s})\mathsf{s}_y^j(D\mathbf{s})\Big] \stackrel{(1.29)}{=} 0. \end{aligned}$$

Thus, three cases have to distinguished:

i) n odd $\Rightarrow$ (1.27) holds without moduli since both sides are 0 (cf. proposition 1.1).

ii) n even but i, j odd $\Rightarrow$ (1.27) holds without moduli since both sides are 0 (special situation satisfied by the D2P9 model).

iii) n even and i, j even $\Rightarrow$ (1.27) holds since $(-1)^j = 1$ in equation (1.28).

Eventually, this shows the isotropic property for the D2P9 model. Of course, it would also be possible to verify it directly using the relations $\mathsf{s} = \mathsf{s}^3 = \mathsf{s}^5 = \ldots$ and $\mathsf{s}^2 = \mathsf{s}^4 = \mathsf{s}^6 = \ldots$ and checking all possible monomials (which are not very much). However, we have preferred to 'fathom' the deeper structural reasons also with regard to other velocity sets.

Let us finally give a basis for $\mathcal{F}(\mathcal{S})$ if $\mathcal{S}$ corresponds to the D2P9 model:

$$\begin{array}{cccl} \mathsf{1}, & \mathsf{s}_x, & \mathsf{s}_y, & \text{(conserved quantities: mass, momentum)} \\ \mathsf{s}_x^2, & \mathsf{s}_y^2, & \mathsf{s}_x\mathsf{s}_y, & \text{(moments related to stress quantities)} \\ \mathsf{s}_x^2\mathsf{s}_y^2, & \mathsf{s}_x\mathsf{s}_y^2, & \mathsf{s}_x^2\mathsf{s}_y, & \text{(no direct physical interpretation)} \end{array}$$

The meaning of the associated moments in a fluid dynamic context is indicated in parenthesis. Usually it is more convenient to work with an orthogonal basis with respect to the scalar product generated by w, i.e. $\langle \mathsf{w}\mathsf{p}_k, \mathsf{p}_l\rangle = 0$ for two different basis functions $\mathsf{p}_k, \mathsf{p}_l \in \mathcal{F}(\mathcal{S})$.

In order to compute the equivalent moment system (or to perform the asymptotic analysis of lattice-Boltzmann schemes using Taylor expansion), it is unavoidable to determine also grid tensors R_n, T_n of rank higher than 4. However, in the case of the D2P9 model the tensor T_6 is no more a multiple of the completely symmetric, isotropic tensor[41], i.e.

$$T_{6;\alpha\beta\gamma\delta\epsilon\eta} \;=\; \langle \mathsf{w}, \mathsf{s}_\alpha\mathsf{s}_\beta\mathsf{s}_\gamma\mathsf{s}_\delta\mathsf{s}_\epsilon\mathsf{s}_\eta\rangle \;\neq\; c\big(\underbrace{\delta_{\alpha\beta}\delta_{\gamma\delta}\delta_{\epsilon\eta} + \delta_{\alpha\beta}\delta_{\gamma\eta}\delta_{\delta\epsilon} + \delta_{\alpha\beta}\delta_{\gamma\epsilon}\delta_{\delta\eta} + \ldots}_{\text{sum extends over all combinations of pairs}}\big).$$

[41] The elements of the completely symmetric, isotropic tensor of rank $2n$ are given as sums, where each addend is a product of n Kronecker deltas. The number of addends is equal to the number A_n of combinations of pairs, that can be formed out of a set of $2n$ elements. These numbers obey the recurrence relation $A_{n+1} = (2n-1)A_n$, because if two elements are added to a set of n elements then there are $2n-1$ different pairs with one of the new elements and for each of these pairs there are A_n combinations of pairs with the remaining elements. So the underbraced sum above counts 15 addends.

This can be seen by computing $T_{6;xxxxxx} = \frac{1}{3}$ and $T_{6;xxxxyy} = \frac{1}{9}$, whereas the right hand side yields $15c$ and $3c$ respectively. Since there is no $c > 0$ which can simultaneously satisfy the conditions $\frac{1}{3} = 15c$ and $\frac{1}{9} = 3c$ the inequality follows. Even if equality held, the right hand side would be quite cumbersome to work with. How to deal with high rank tensors is illustrated by the subsequent example.

Example: A typical multi-sum that has to be computed is the following one:

$$(*) \;=\; \sum_{\alpha,\beta,\gamma,\delta,\epsilon} \partial_\alpha\partial_\beta\partial_\gamma\partial_\delta u_\epsilon \sum_{\mathbf{s}\in\mathcal{S}} \mathsf{w}(\mathbf{s})\,\mathsf{s}_\alpha(\mathbf{s})\,\mathsf{s}_\beta(\mathbf{s})\,\mathsf{s}_\gamma(\mathbf{s})\,\mathsf{s}_\delta(\mathbf{s})\,\mathsf{s}_\epsilon(\mathbf{s})\,\mathsf{s}_\eta(\mathbf{s}).$$

As the sum does not stretch over η we choose for instance $\eta = y$. Observe that the inner sum over $\mathbf{s}$ is different from 0 only if the number of indices equal to x and the number of indices equal to y are both even.

$$\eta = y,\ \epsilon = x \quad \begin{cases} \text{3 other indices equal to } x, & \binom{3}{4} = 4 \text{ possibilities} \quad \sum_{\mathbf{s}\in\mathcal{S}} \cdots = \frac{4}{36} = \frac{1}{9} \\ \text{1 other index equal to } x, & \binom{1}{4} = 4 \text{ possibilities} \quad \sum_{\mathbf{s}\in\mathcal{S}} \cdots = \frac{4}{36} = \frac{1}{9} \end{cases}$$

$$\eta = y,\ \epsilon = y \quad \begin{cases} \text{all other indices equal to } x, & \binom{4}{4} = 1 \text{ possibilities} \quad \sum_{\mathbf{s}\in\mathcal{S}} \cdots = \frac{4}{36} = \frac{1}{9} \\ \text{2 other indices equal to } x, & \binom{2}{4} = 6 \text{ possibilities} \quad \sum_{\mathbf{s}\in\mathcal{S}} \cdots = \frac{4}{36} = \frac{1}{9} \\ \text{no other index equal to } x, & \binom{0}{4} = 1 \text{ possibilities} \quad \sum_{\mathbf{s}\in\mathcal{S}} \cdots = \frac{2}{9} + \frac{4}{36} = \frac{1}{3} \end{cases}$$

Hence the searched sum is for $\eta = y$:

$$(*) \;=\; \tfrac{4}{9}\partial_x^3\partial_y u_x + \tfrac{4}{9}\partial_x\partial_y^3 u_x + \tfrac{1}{9}\partial_x^4 u_y + \tfrac{2}{3}\partial_x^2\partial_y^2 u_y + \tfrac{1}{3}\partial_y^4 u_y.$$

Fluid dynamic equilibrium functions for planar flows

As application of the structure relations we would like to cast a quick glance onto equilibrium functions that are widely employed to simulate viscous, incompressible flows requiring the numerical solution of the Navier-Stokes equation.

$$\mathcal{E}(\rho,\mathbf{u};\mathbf{s}) \;:=\; \rho\,\mathsf{w}(\mathbf{s})\left(1 + 3\mathbf{s}\cdot\mathbf{u} + \tfrac{9}{2}(\mathbf{s}\cdot\mathbf{u})^2 - \tfrac{3}{2}(\mathbf{u}\cdot\mathbf{u})\right) \tag{1.30}$$

$$\mathcal{E}_m(p,\mathbf{u};\mathbf{s}) \;:=\; \mathsf{w}(\mathbf{s})\left(3p + 3\mathbf{s}\cdot\mathbf{u} + \tfrac{9}{2}(\mathbf{s}\cdot\mathbf{u})^2 - \tfrac{3}{2}(\mathbf{u}\cdot\mathbf{u})\right) \tag{1.31}$$

ρ, $\mathbf{u}$ and p are physical quantities standing for the (local) density and velocity of the fluid and for the pressure. The second equilibrium function $\mathcal{E}_m$ has been proposed as a modification of the first one in order to reduce the compressibility error (cf. [25]) inherent to lattice-Boltzmann methods.

In the context of (1.30), the mass density ρ and the momentum density $\rho\mathbf{u}$ are related to the moments of the population function f: $\rho = \langle \mathsf{1}, \mathsf{f}\rangle$ and $\rho u_\alpha = \langle \mathsf{s}_\alpha, \mathsf{f}\rangle$. Expanding the scalar products and introducing our notation, (1.30) becomes

$$\begin{aligned}\mathcal{E}(\rho,\mathbf{u}) = \rho\mathsf{w} \;+\; 3\rho u_x \mathsf{w}\mathsf{s}_x \;+\; 3\rho u_y \mathsf{w}\mathsf{s}_y& \\ +\; \tfrac{9}{2}\rho u_x^2 \mathsf{w}\mathsf{s}_x^2 \;+\; \tfrac{9}{2}\rho u_x^2 \mathsf{w}\mathsf{s}_x^2 \;+\; 9\rho u_x u_y \mathsf{w}\mathsf{s}_x\mathsf{s}_y \;-\; \tfrac{3}{2}\rho(u_x^2 + u_y^2)\mathsf{w}.&\end{aligned}$$

In this form the formulas (1.25) can be applied to compute the 0^{th} and 1^{st} moment of the equilibrium function:

$$\big\langle \mathsf{1}, \mathcal{E}(\rho,\mathbf{u})\big\rangle = \rho, \qquad \big\langle \mathsf{s}_\alpha, \mathcal{E}(\rho,\mathbf{u})\big\rangle = \rho u_\alpha.$$

This implies that the mass and momentum density, being physically conserved quantities, are also conserved by the lattice-Boltzmann collision operator. Correctly, $\mathcal{E}$ should not be given as function of ρ and $\mathbf{u}$ but as function of ρ and $\rho\mathbf{u}$; however that makes it look more complicated.
Inserting $\mathcal{E}$ into the lattice-Boltzmann equation and computing the moments with respect to $\mathsf{1}, \mathsf{s}_x, \mathsf{s}_y$ yields

$$\begin{aligned}\partial_t\langle \mathsf{1},\mathcal{E}\rangle \;+\; \partial_x\langle \mathsf{1},\mathsf{s}_x\mathcal{E}\rangle \;+\; \partial_y\langle \mathsf{1},\mathsf{s}_y\mathcal{E}\rangle &= 0,\\ \partial_t\langle \mathsf{s}_x,\mathcal{E}\rangle + \partial_x\langle \mathsf{s}_x,\mathsf{s}_x\mathcal{E}\rangle + \partial_y\langle \mathsf{s}_x,\mathsf{s}_y\mathcal{E}\rangle &= 0,\\ \partial_t\langle \mathsf{s}_y,\mathcal{E}\rangle + \partial_x\langle \mathsf{s}_y,\mathsf{s}_x\mathcal{E}\rangle + \partial_y\langle \mathsf{s}_y,\mathsf{s}_y\mathcal{E}\rangle &= 0.\end{aligned}$$

Using (1.25) again, the following equations are found:

$$\begin{aligned}
&\partial_t\rho + \partial_x(\rho u_x) + \partial_y(\rho u_y) = 0 && \Leftrightarrow \quad \partial_t\rho + \nabla\cdot(\rho\mathbf{u}) = 0\\
&\left.\begin{aligned}\partial_t(\rho u_x) + \partial_x(\rho u_x^2 + \tfrac{1}{3}\rho) + \partial_y(\rho u_x u_y) &= 0\\ \partial_t(\rho u_y) + \partial_x(\rho u_x u_y) + \partial_y(\rho u_y^2 + \tfrac{1}{3}\rho) &= 0\end{aligned}\right\} && \Leftrightarrow \quad \partial_t(\rho\mathbf{u}) + \nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u}) = -\tfrac{1}{3}\nabla\rho
\end{aligned}$$

These are known as the *barotropic* or *isentropic Euler equations*[42] with the pressure $p(\rho) = \frac{1}{3}\rho$. In contrast to the compressible Euler equations (1.13) the pressure is considered as a function of the density alone. It should be noted that the derivation of the above barotropic Euler equations parallels that of the compressible Euler equations in the previous subsection.

To satisfy the incompressibility condition $\nabla\cdot\mathbf{u} = 0$, is a major difficulty[43] for fluid flow simulations based on lattice-Boltzmann methods. An improvement has been brought by the equilibrium $\mathcal{E}_m$ in (1.31). Proceeding with $\mathcal{E}_m$ as above with $\mathcal{E}$ leads to

$$\begin{aligned}3\partial_t p + \nabla\cdot\mathbf{u} &= 0,\\ \partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} &= -\nabla p.\end{aligned}$$

The resulting system represents a 'hyperbolicization' of the incompressible Euler equation $\partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} = -\nabla p,\ \ \nabla\cdot\mathbf{u} = 0$. If the parabolic scaling is considered, it can be shown that $\mathcal{E}_m$ is related to the incompressible Navier-Stokes equation (see [34]).

Derivation of the lattice-Boltzmann algorithm

In the preceding paragraphs the fundamental ideas of the *kinetic approach* have been explained. Under certain conditions it is possible to formulate a given target equation in terms of a *kinetic* i.e. Boltzmann-type equation. It is expected that this formulation is more amenable to numerical solution methods.
Now we have to tackle the second step of the two-point program: the discretization of such *kinetic equations* leading to *kinetic schemes*. These numerical algorithms then enable the (approximate) solution by means of computer programs.

Actually, there is a great zoo of different techniques to discretize (partial) differential equations. *Finite difference*[44] and *finite element*[45] methods are rather universal. They can be applied without taking the special structure of the kinetic equation

[42] The compressible Euler equation comprises three conservation equations: for the mass density, the momentum density and for a thermodynamic quantity being either the specific internal energy, the temperature or the entropy. The entropy equation is a pure advection equation with respect to the flow field; however it is coupled to the momentum equation due to the dependence of the pressure on the entropy. Under the assumption that the initial distribution of the entropy is a spatially homogeneous field, the entropy remains constant in space and time as long as the other fields are smooth. In this (isentropic) case the pressure becomes only a function of the mass density (barotropic) and the compressible Euler system reduces to two equations for the mass and momentum density forming thus the isentropic or barotropic Euler equation.

[43] The reason for this is that the incompressibility condition represents a degenerated conservation equation that has the character of a constraint. This entails a different role of the pressure as a variable which has to help satisfying this constraint (Lagrange-multiplier). Therefore compressible and incompressible fluid dynamic equations are structurally different, where the first one fits better into the kinetic framework.

[44] This discretization replaces derivatives by difference quotients.

[45] The point of departure for finite elements is a variational formulation based on well-defined function spaces. The idea is to find approximating finite-dimensional function spaces which can be used to formulate a restricted problem. The variational formulation is also useful to introduce the notion of a weak solution.

into account. An alternative is offered by finite volume methods, which are typically applied in the context of balance laws[46].

In opposition, *kinetic particle schemes* represent a very specific possibility of discretization. They are based on the observation that the target equation is reproduced by taking appropriate moments of the lattice-Boltzmann equation, where a suitable equilibrium has been inserted for the population function. The population function is initialized by the equilibrium which is computed from the macroscopic initial values (referring to the target equation). Each population is then advected during a short time step according to the associated discrete velocity. After this, at each (grid) point the moments (especially those being conserved by the collision operator) are calculated and from them the new equilibrium is determined. Then the population function is updated by the new equilibrium function and the process is iterated.
This prescription yields a semi-discrete algorithm in time. A further discretization is necessary to make it fully discrete in space (advection of the populations) and velocity space to compute the moments and the equilibrium function.

The lattice-Boltzmann algorithm works in quite a similar way. In the following we show, how it can be formally derived from the lattice-Boltzmann equation in parabolic scaling. Here the equation is considered for an arbitrary but fixed population[47] $\mathsf{f}_{\mathbf{s}}$:

$$\partial_t \mathsf{f}_{\mathbf{s}} + \epsilon^{-1}\mathbf{s}\cdot\nabla \mathsf{f}_{\mathbf{s}} = \epsilon^{-2}\omega\big(E_{\mathbf{s}}(\mathsf{f}) - \mathsf{f}_{\mathbf{s}}\big) + \mathsf{q}_{\mathbf{s}}. \tag{1.32}$$

In the case of an advective differential operator – as occurring at the left hand side of the lattice-Boltzmann equation – the (projected) characteristic of $\mathsf{f}_{\mathbf{s}}$ is a straight line in the time-space domain passing some point $(t, \mathbf{x})$. Its parameterization is given by

$$\mathbb{R}_0^+ \ni \chi \mapsto (t+\chi,\ \mathbf{x}+\chi\epsilon^{-1}\mathbf{s}) \to \mathbb{R}\times\mathbb{R}^d,$$

where the parameter χ plays the role of a running time while t is fixed. Now let us integrate (1.32) along the characteristic from 0 to ϵ^2:

$$\int_0^{\epsilon^2} \Big[\partial_t \mathsf{f}_{\mathbf{s}}(t+\chi,\mathbf{x}+\tfrac{1}{\epsilon}\chi\mathbf{s})+\tfrac{1}{\epsilon}(\mathbf{s}\cdot\nabla)\mathsf{f}_{\mathbf{s}}(t+\chi,\mathbf{x}+\tfrac{1}{\epsilon}\chi\mathbf{s})\Big]\,\mathrm{d}\chi = \int_0^{\epsilon^2} \mathrm{RHS}_{\mathbf{s}}(t+\chi,\mathbf{x}+\tfrac{1}{\epsilon}\chi\mathbf{s})\,\mathrm{d}\chi.$$

The right hand side of (1.32), comprising the relaxation term and the source, is abbreviated by $\mathrm{RHS}_{\mathbf{s}}$. A sharp glance at the left hand side uncovers the integrand as derivative with respect to the integration variable χ. This circumstance permits an exact evaluation by means of the fundamental theorem of calculus. In contrast the right hand side is only approximately evaluable. Using the primitive left rectangle integration rule, which is justified for ϵ^2 being small, we get

$$\int_0^{\epsilon^2} \frac{\mathrm{d}}{\mathrm{d}\chi}\mathsf{f}_{\mathbf{s}}(t_0+\chi,\mathbf{x}_0+\chi\epsilon^{-1}\mathbf{s})\,\mathrm{d}\chi = \epsilon^2\mathrm{RHS}_{\mathbf{s}}\big|_{\chi=0} + \mathrm{O}(\epsilon^2)$$

[46]Notice that the Boltzmann equation represents a balance law in the phase space.

[47]Component of f referring to $\mathbf{s}\in\mathcal{S}$; it should be noticed that the argument of the equilibrium operator is the full tuple of populations.

and finally

$$\mathsf{f}_{\mathbf{s}}(t+\epsilon^2,\mathbf{x}+\epsilon\mathbf{s})-\mathsf{f}_{\mathbf{s}}(t,\mathbf{x})=\omega\big[E_{\mathbf{s}}\big(\mathsf{f}(t,\mathbf{x})\big)-\mathsf{f}_{\mathbf{s}}(t,\mathbf{x})\big]+\epsilon^2\mathsf{q}_{\mathbf{s}}(t,\mathbf{x})+\mathrm{O}(\epsilon^2). \quad (1.33)$$

The error committed by the quadrature rule is compensated by an additional term. This term should vanish like the squared length of the integration interval, viz. $(\epsilon^2)^2=\epsilon^4$. However, the factor ϵ^{-2} in front of the relaxation term accounts for the magnitude of $\mathrm{O}(\epsilon^2)$. As the population function f is expected to be quite close to the equilibrium $E(\mathsf{f})$, the difference $E(\mathsf{f})-\mathsf{f}$ should be relatively small. Therefore the approximately integrated source term $\epsilon^2\mathsf{q}(t,\mathbf{x})$ is not incorporated into this compensation term.

Neglecting the $\mathrm{O}(\epsilon^2)$-term in (1.33) yields a reasonable difference equation.

$$\mathsf{F}_{\mathbf{s}}(t+h^2,\mathbf{x}+h\mathbf{s})=\mathsf{F}_{\mathbf{s}}(t,\mathbf{x})+\omega\big[E_{\mathbf{s}}\big(\mathsf{F}_{\mathbf{s}}(t,\mathbf{x})\big)-\mathsf{F}_{\mathbf{s}}(t,\mathbf{x})\big]+h^2\mathsf{q}_{\mathbf{s}}(t,\mathbf{x}). \quad (1.34)$$

The parameter ϵ has been baptized h as it plays now the role of a grid spacing. This equation determines for $\mathbf{s}\in\mathcal{S}$ the evolution of a discrete population function F, which is defined on an appropriate time-space mesh $\mathcal{T}(h^2)\times\mathcal{X}(h)$ covering the domain $[0,T]\times\mathcal{X}$ of f. The exact structure of $\mathcal{T}(h^2)\times\mathcal{X}(h)$ is also prescribed by (1.34):

- The time grid $\mathcal{T}(h^2)$ must be uniform with a step-size $\Delta t=h^2$.
- The spatial grid $\mathcal{X}(h)$ must satisfy the following property: $\mathbf{x}\in\mathcal{X}(h)$ requires $\mathbf{x}+h\mathbf{s}\in\mathcal{X}(h)$ for any $\mathbf{s}\in\mathcal{S}$. So the spatial grid is uniform as well, with a grid spacing of $\Delta\mathbf{x}=h\mathbf{s}$ in the direction of $\mathbf{s}$.

Each time step of the lattice-Boltzmann algorithm consists of two substeps being similar to the kinetic particle scheme.

- The evaluation of the of the right hand side of (1.34) which is referred to as the *collision step*. This substep is a nodal operation because it can be executed for every node without accessing the population function at other nodes.
- During the second substep – denoted also as propagation, advection or update step – the computed right hand sides – the so-called *collision products* – are shifted to the neighbor nodes lying in direction of the associated discrete velocity to obtain the new populations.

The essential computational work has to be done in the first substep, while the second substep is mainly a copy process.
For grid nodes situated closest to spatial boundary $\partial\mathcal{X}$, suitable boundary conditions must be provided. It is possible to introduce bounce-back like boundary conditions for those populations residing at the boundary nodes that cannot be updated by (1.34). However, one has to bear in mind, that the standard bounce-back rules loses some accuracy, unless the boundary node is situated in a very specific position with respect to $\partial\mathcal{X}$ (half grid-spacing distance).

It is important to note that the singular parameter ϵ occurring in the lattice-Boltzmann equation (e.g. in front of the relaxation term) has been completely removed in the discretization above. The reason for this is, that the time step Δt is coupled to ϵ in a specific way which melts two *a priori* independent limit processes into a single, simultaneous limit process. Therefore (1.34) represents *no consistent* discretization of (1.32) but aims directly towards an approximate solution of the target equation.

1.2 Translational invariance and dimensional reduction

It has already been mentioned in the preceding section that 2D and 1D lattice-Boltzmann velocity models are related to each other. In this section we exemplarily show how D1P3 algorithms are extracted from D2P9 algorithms. Generally, two-dimensional lattice-Boltzmann schemes result into one-dimensional lattice-Boltzmann schemes if certain symmetry conditions are fulfilled. Therefore the latter ones provide appropriate starting points for theoretical investigations of the first ones – not just due to structural similarities but, what weighs more, as a kind of simplifying reformulation.

The first paragraph discusses shear flows as translationally invariant solutions of the incompressible Navier-Stokes equation (INSE). For this class of flows it is possible to boil down the nonlinear, two-dimensional INSE to become the linear diffusion equation in one space dimension. The following paragraphs are devoted to the D2P9 model where we try to mimick this process of simplification. Besides the reasons mentioned above it is interesting to check, whether the D2P9 lattice-Boltzmann algorithm exhibits similar symmetries as those equations, that can be approximated by it (Stokes, Oseen, Navier-Stokes).

Definition 1.1. *(Translational invariance of sets and functions)*

i) A subset $\Omega \subset \mathbb{R}^n$ *is called* translationally invariant *with respect to the* displacement *vector* $\mathbf{d} \in \mathbb{R}^n \setminus \{0\}$ *iff:* $\forall \alpha \in \mathbb{R}: \quad \mathbf{x} \in \Omega \quad \Rightarrow \quad \mathbf{x} + \alpha\mathbf{d} \in \Omega.$

ii) Let $\Omega \subset \mathbb{R}^n$ *be* translationally invariant *with respect to* $\mathbf{d}$. *A function* $f : \Omega \to \mathbb{R}^n$ *is called* translationally invariant *or simply* constant *in the direction of* $\mathbf{d}$ *iff:* $\forall \alpha \in \mathbb{R}: \quad f(\mathbf{x}) = f(\mathbf{x} + \alpha\mathbf{d}).$

Obviously, translationally invariant subsets correspond to stripes parallel to $\mathbf{d}$. These sets are unbounded and do not permit a *finite, regular discretization* with a constant grid spacing. This however would be very desirable with regard to computer simulations. Therefore we introduce two planar manifolds. In the following, the constants H, L denote some positive numbers.

$$\text{planar torus:} \quad \mathscr{T} := \mathbb{R}/L \times \mathbb{R}/H \qquad \text{planar cylinder:} \quad \mathscr{Z} := [0, L] \times \mathbb{R}/H$$

The torus corresponds to the rectangle $[0, L] \times [0, H]$ where opposite sides are identified. This manifold is translationally invariant in any direction. For flow simulations it is realized by a rectangle with periodic boundary conditions in x- and y-direction. If only one pair of opposite sides is identified, the planar cylinder is obtained. Here the sides parallel to the x-axis are chosen such that the cylinder becomes translationally invariant in y-direction. The cylinder models a pipe where periodic boundary conditions are assumed at the inlet and outlet to avoid more complicated inflow and outflow conditions.

Shear Flows

Shear flows are distinguished by the fact that the streaming velocity is everywhere parallel to a fixed flow direction. Moreover the incompressibility enforces the velocity field to be translationally invariant[48] in this direction. It turns out that shear flows can not only be described as solutions of the incompressible Navier-Stokes equation (INSE) but also as solutions of the simpler diffusion equation. This circumstance facilitates the finding of analytic solutions. Two prominent examples are the *Poiseuille* flow and the *Couette* flow. Both are frequently used as simple benchmarks for numeric Navier-Stokes solvers.

Definition 1.2. *A solution of the INSE is called a* shear flow in y-direction*, if the x-component u of the velocity $\mathbf{v} = \binom{u}{v}$ is identically zero.*

The incompressiblity condition $\nabla \cdot \mathbf{v} = \partial_x u + \partial_y v = 0$ implies immediately that $\partial_y v$ must vanish, because u and *a fortiori* $\partial_x u$ is zero by assumption. Hence, v must be independent of y and the velocity vector $\mathbf{v}$ simplifies in the following way:

$$\mathbf{v}(t,x,y) = \mathbf{v}(t,x) = \begin{pmatrix} 0 \\ v(t,x) \end{pmatrix} \tag{1.35}$$

Proposition 1.3. *The incompressibility condition entails that any solution of the INSE being a shear-flow in y-direction is constant in y-direction.*

We conclude that the velocity field of a shear flow is completely described by a single scalar function of one spatial coordinate. This justifies the commonly used term *quasi-1d flow.* v is often referred to as the *velocity profile* or *flow profile.* Plugging (1.35) into the INSE leads to the evolution equation for v:

$$\begin{aligned} \partial_t \begin{pmatrix} 0 \\ v \end{pmatrix} \; + \; (0 + v\partial_y) \cdot \begin{pmatrix} 0 \\ v \end{pmatrix} \; - \; \nu\Delta \begin{pmatrix} 0 \\ v \end{pmatrix} \; &= \; -\nabla p \\ \begin{pmatrix} 0 \\ \partial_t v \end{pmatrix} \; + \; \begin{pmatrix} 0 \\ v\,\partial_y v \end{pmatrix} \; - \; \nu \begin{pmatrix} 0 \\ \partial_x^2 v + \partial_y^2 v \end{pmatrix} \; &= \; -\begin{pmatrix} \partial_x p \\ \partial_y p \end{pmatrix} \end{aligned}$$

Noting that the nonlinear term disappears, one obtains

$$0 = -\partial_x p \quad \wedge \quad \partial_t v \; - \; \nu\, \partial_x^2\, v \; = \; -\partial_y p.$$

The independence of the velocity with respect to the y-coordinate carries over to the pressure p. This entails $\nabla p = 0$ obliging the pressure to be constant. Being uniquely determined up to an additive constant, the pressure can be set to 0.

Evidently, the flow profile v appears as a solution of the *one-dimensional diffusion equation.* The result is summarized more precisely by the following proposition.

[48] In principle, this requires the flow domain to be translationally invariant as well.

Proposition 1.4. ***(Conservation of translational invariance)***

i) Consider the IVP for the INSE in the torus $\mathscr{T}$:

$$\partial_t \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} - \nu \Delta \mathbf{v} = -\nabla p + \mathbf{g},$$
$$\nabla \cdot \mathbf{v} = 0, \qquad \mathbf{v}(0, \cdot) = \mathbf{v}_0.$$

$$\textit{If} \qquad \mathbf{v}_0(x,y) = \begin{pmatrix} 0 \\ v_0(x) \end{pmatrix} \qquad \textit{and} \qquad \mathbf{g}(t,x,y) = \begin{pmatrix} 0 \\ g(t,x) \end{pmatrix}, \qquad (*)$$

there exists a global solution that is constant in y-direction, provided the diffusion problem

$$\partial_t v - \nu \partial_x^2 v = g \qquad v(0, \cdot) = v_0$$

admits a global solution.

ii) Consider the IBVP problem for the INSE in the cylinder $\mathscr{Z}$:

$$\partial_t \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} - \nu \Delta \mathbf{v} = -\nabla p + \mathbf{g}$$
$$\nabla \cdot \mathbf{v} = 0 \qquad \mathbf{v}(0, \cdot) = \mathbf{v}_0$$
$$\mathbf{v}(t,x,0) = \mathbf{w}_0(t,x) \qquad \mathbf{v}(t,x,H) = \mathbf{w}_H(t,x)$$

If $(*)$ is satisfied and moreover for some $a \in \mathbb{R}$

$$\mathbf{w}_0(t,x) = \begin{pmatrix} a \\ w_0(t) \end{pmatrix}, \qquad \mathbf{w}_H(t,x) = \begin{pmatrix} a \\ w_H(t) \end{pmatrix},$$

there exists a global solution that is constant in y-direction provided the advection-diffusion problem

$$\partial_t v + a \partial_y v - \nu \partial_y^2 v = g \qquad v(0, \cdot) = v_0$$
$$v(t,0) = w_0(t) \qquad v(t,H) = w_H(t)$$

has a global solution.

Proof: The solution of the Navier-Stokes problem is obtained by setting $\mathbf{v}(t,x,y) = \mathbf{v}(t,x) = \binom{0}{v(t,x)}$ in first case and $\mathbf{v}(t,x,y) = \mathbf{v}(t,x) = \binom{a}{v(t,x)}$ in the second one. ■

Remark 1.1. If the given data, i.e. the initial velocity field and the exterior force, are constant in y-direction, the solution for any IVP of the INSE on a translationally invariant domain inherits this property. Therefore the INSE preserves translational invariance.

Remark 1.2. Parallel shear flows are solutions of the Stokes and the Navier-Stokes equation, since the nonlinear, convective term of the Navier-Stokes equation vanishes in the case of the special velocity field.

Reduction of the D2P9 lattice-Boltzmann algorithm

Let us consider the lattice-Boltzmann scheme with the D2P9 velocity model on a rectangular grid discretizing the torus. $I, J \in \mathbb{N}$ denote the number of grid nodes in x- and y-direction respectively. The populations F_k with $k \in \{1, ..., 9\}$ are indexed as indicated in table 1.1. Moreover the F_k's are taken as functions of the time step index n and two spatial indices i, j specifying the grid node the F_k's belong to:

$$\mathsf{F}_k : \mathbb{N}_0 \times \mathbb{Z}/I\mathbb{Z} \times \mathbb{Z}/J\mathbb{Z} \to \mathbb{R}.$$

The equilibrium $\mathcal{E}$ is assumed to depend on the mass moment R (interpreted as pseudo-density) and the moments U, V of first order (interpreted as velocity components) like (1.30) and (1.31). For convenience we set

$$\left.\begin{array}{c} \mathsf{E}_k(n,i,j) := \mathcal{E}_k\big(R(n,i,j), U(n,i,j), V(n,i,j)\big) \qquad \text{with} \\ R := \sum_k \mathsf{F}_k, \quad U := \sum_k \mathsf{F}_k s_{kx}, \quad V := \sum_k \mathsf{F}_k s_{ky}. \end{array}\right\} \tag{1.36}$$

The populations evolve in time according to equation (1.34) that we write in the subsequent form:

$$\left.\begin{array}{l} \mathsf{F}_k(n+1,i,j) = (1-\omega)\,\mathsf{F}_k(n, i - s_{kx}, j - s_{ky}) \;+\; \omega\,\mathsf{E}_k(n, i - s_{kx}, j - s_{ky}) \\ \Leftrightarrow \;\; \mathsf{F}_k(n+1, i + s_{xk}, j + s_{ky}) \;=\; (1-\omega)\,\mathsf{F}_k(n,i,j) \;+\; \omega\,\mathsf{E}_k(n,i,j) \end{array}\right\} \tag{1.37}$$

The choice of the spatial index sets $\mathbb{Z}/I\mathbb{Z}$, $\mathbb{Z}/J\mathbb{Z}$ encodes directly the periodic boundary conditions along all sides of the rectangular domain.

The following property of the lattice-Boltzmann algorithm is as simple as fundamental.

Proposition 1.5. ***(Conservation of translational invariance)***
If the populations are initialized such that they do not vary with j, then they remain independent of j for all iterations (translational invariance in y-direction).

Proof: The proof is done by recurrence. For $n = 0$ the assertion is true by assumption. Let us now conclude that it remains true from the n^{th} to the $(n+1)^{\text{th}}$ iteration:
By hypothesis of induction the $\mathsf{F}_k(n,\cdot,\cdot)$'s on the right hand side are constant with respect to j. Thus the $\mathsf{E}_k(n,\cdot,\cdot)$'s do not vary with j either, since the equilibrium is assumed to depend not explicitly on j. So it depends on j only via the moments R, U, V being however linear combinations of the $\mathsf{F}_k(n,\cdot,\cdot)$'s. Hence the right hand side is independent of j and so must be the left hand side. This is equivalent to the claim that the populations of the $(n+1)^{\text{th}}$ iteration are independent of j. ∎

Remark 1.3. Since the populations F_k are allowed to attain arbitrary real values, the set of all admissible states pertaining to the lattice-Boltzmann scheme forms a vector space being isomorphic to $\mathbb{R}^{9IJ}$. Evidently, the subset of states which are independent of j represents a subspace of dimension $9I$. The proposition is equivalent to the statement, that this subspace is *invariant* under the *evolution operator* of the lattice-Boltzmann scheme.

Let us suppose that the assumption of proposition 1.5 is fulfilled. Consequently, the index j can be omitted in (1.37). Illustratively spoken, this means that the rectangular two-dimensional grid is replaced by a one-dimensional grid corresponding to a cross-section along the x-axis. The accompanying *one-dimensional evolution equations* become:

$$\left.\begin{array}{rcccl} \mathsf{F}_{1,2,8}(n+1,i) &=& (1-\omega)\,\mathsf{F}_{1,2,8}(n,i-1) &+& \omega\,\mathsf{E}_{1,2,8}(n,i-1), \\ \mathsf{F}_{3,7,9}(n+1,i) &=& (1-\omega)\,\mathsf{F}_{3,7,9}(n,i) &+& \omega\,\mathsf{E}_{3,7,9}(n,i), \\ \mathsf{F}_{4,5,6}(n+1,i) &=& (1-\omega)\,\mathsf{F}_{4,5,6}(n,i+1) &+& \omega\,\mathsf{E}_{4,5,6}(n,i+1). \end{array}\right\} \qquad (1.38)$$

The nine populations are grouped into three triples:

the right-moving populations F_1, F_2, F_8 with $s_{kx} = 1$,
the resting populations F_3, F_7, F_9 with $s_{kx} = 0$,
the left-travelling populations F_4, F_5, F_6 with $s_{kx} = -1$.

Obviously, the populations of each triple share the same evolution equation. Therefore one might be tempted to reduce the number of populations and equations by taking simply one equation out of every triple. However, this idea does not work out, since the evaluation of the equilibrium requires the computation of the moments R, U and V. These, however, can not be computed by the knowledge of three populations alone and therefore the system is *not closed*.

$$\begin{array}{rcl} R &=& \mathsf{F}_3+\mathsf{F}_7+\mathsf{F}_9 \;+\; \mathsf{F}_1+\mathsf{F}_2+\mathsf{F}_8 \;+\; \mathsf{F}_4+\mathsf{F}_5+\mathsf{F}_6 \\ U &=& \mathsf{F}_1+\mathsf{F}_2+\mathsf{F}_8 \;-\; \big(\mathsf{F}_4+\mathsf{F}_5+\mathsf{F}_6\big) \\ V &=& \mathsf{F}_2-\mathsf{F}_8 \;+\; \mathsf{F}_3-\mathsf{F}_7 \;+\; \mathsf{F}_4-\mathsf{F}_6 \end{array}$$

Actually, the full knowledge of the 9 populations is not necessary. The moments R, U and V are available as soon as the following quantities are known:

$$\begin{array}{rr} \mathsf{A}_+ := \mathsf{F}_1+\mathsf{F}_2+\mathsf{F}_8 & \mathsf{D}_+ := \mathsf{F}_2-\mathsf{F}_8 \\ \mathsf{A}_0 := \mathsf{F}_3+\mathsf{F}_7+\mathsf{F}_9 & \mathsf{D}_0 := \mathsf{F}_3-\mathsf{F}_7 \\ \mathsf{A}_- := \mathsf{F}_4+\mathsf{F}_5+\mathsf{F}_6 & \mathsf{D}_- := \mathsf{F}_4-\mathsf{F}_6 \end{array} \Rightarrow \left\{\begin{array}{rcl} R &=& \mathsf{A}_+ + \mathsf{A}_0 + \mathsf{A}_- \\ U &=& \mathsf{A}_+ \qquad\;\; - \mathsf{A}_- \\ V &=& \mathsf{D}_+ + \mathsf{D}_0 + \mathsf{D}_- \end{array}\right. \qquad (1.39)$$

Each of these 6 linear combinations is only composed of those populations that belong to the same triple. In order to obtain the evolution system for the new quantities, the equations (1.38) have to be combined in the same manner. Since R, U, and V can be expressed only in terms of the new quantities, the system is *closed*. Therefore we are able to formulate the next proposition.

Proposition 1.6. *Consider the D2P9 lattice-Boltzmann scheme on a rectangular grid discretizing the torus. If the initial conditions are constant in y-direction, then the lattice-Boltzmann algorithm can be reduced to a one-dimensional scheme using 6 instead of 9 populations without affecting the numeric values of the three moments R, U and V.*

The elimination of further populations is only possible under extra assumptions. We distinguish between two possibilities:

i) (pure) shear mode: $R = 0 \;\wedge\; U = 0$ ii) (pure) acoustic mode: $V = 0$

Proposition 1.7. *Consider the D2P9 lattice-Boltzmann scheme on a rectangular grid discretizing the torus.*

i) ***Shear mode:*** *Suppose that the initial conditions are constant in y-direction and that moreover the moments R and U vanish for all iterations. Then the D2P9 scheme can be reduced to a D1P3 scheme without affecting the numeric value of V.*

ii) ***Acoustic mode:*** *Suppose that the initial conditions are constant in y-direction and that moreover the moment V vanishes for all iterations. Then the D2P9 scheme can be reduced to a D1P3 scheme without affecting the numeric values of R, U.*

Proof:

ad i) Using the assumption $R = 0, U = 0$ and setting

$$\begin{aligned}
\mathcal{E}_+^D(V) &:= \mathcal{E}_2(0,0,V) - \mathcal{E}_8(0,0,V),\\
\mathcal{E}_0^D(V) &:= \mathcal{E}_3(0,0,V) - \mathcal{E}_7(0,0,V),\\
\mathcal{E}_-^D(V) &:= \mathcal{E}_4(0,0,V) - \mathcal{E}_6(0,0,V),
\end{aligned}$$

the following system is directly derived from (1.38):

$$\begin{aligned}
\mathsf{D}_+(n+1,i) &= (1-\omega)\,\mathsf{D}_+(n,i-1) + \omega\,\mathcal{E}_+^D\big(V(n,i-1)\big),\\
\mathsf{D}_0(n+1,i) &= (1-\omega)\,\mathsf{D}_0(n,i) + \omega\,\mathcal{E}_0^D\big(V(n,i)\big),\\
\mathsf{D}_-(n+1,i) &= (1-\omega)\,\mathsf{D}_-(n,i+1) + \omega\,\mathcal{E}_-^D\big(V(n,i+1)\big).
\end{aligned}$$

Since $V = \mathsf{D}_+ + \mathsf{D}_0 + \mathsf{D}_-$, the system is closed and corresponds to the evolution equation of the D1P3 algorithm.

ad ii) Proceeding analogously as above

$$\begin{aligned}
\mathcal{E}_+^A(R,U) &= \mathcal{E}_1(R,U,0) + \mathcal{E}_2(R,U,0) + \mathcal{E}_8(R,U,0),\\
\mathcal{E}_0^A(R,U) &= \mathcal{E}_3(R,U,0) + \mathcal{E}_7(R,U,0) + \mathcal{E}_9(R,U,0),\\
\mathcal{E}_-^A(R,U) &= \mathcal{E}_4(R,U,0) + \mathcal{E}_5(R,U,0) + \mathcal{E}_6(R,U,0),
\end{aligned}$$

we stumble again on the evolution equation of the D1P3 lattice-Boltzmann scheme:

$$\begin{aligned}
\mathsf{A}_+(n+1,i) &:= (1-\omega)\,\mathsf{A}_+(n,i-1) + \omega\,\mathcal{E}_+^A\big(R(n,i-1),U(n,i-1)\big),\\
\mathsf{A}_0(n+1,i) &:= (1-\omega)\,\mathsf{A}_0(n,i) + \omega\,\mathcal{E}_0^A\big(R(n,i),U(n,i)\big),\\
\mathsf{A}_-(n+1,i) &:= (1-\omega)\,\mathsf{A}_-(n,i+1) + \omega\,\mathcal{E}_-^A\big(R(n,i+1),U(n,i+1)\big).
\end{aligned}$$

Due to $R = \mathsf{A}_+ + \mathsf{A}_0 + \mathsf{A}_-$ and $U = \mathsf{A}_+ - \mathsf{A}_-$ the system is closed. ∎

In the sequel we have to verify that the assumption of the proposition can be fulfilled. So it must be shown that the D2P9 algorithm can be initialized so that R and U or V remain zero over all iterations.

Remark 1.4. (Boundary conditions) In order to study the translation of boundary conditions from the full D2P9 scheme to the reduced D1P3 scheme, we consider a rectangular grid discretizing the cylinder. The populations are then functions of the following type:

$$\mathsf{F}_k : \mathbb{N}_0 \times \{1, ..., I\} \times \mathbb{Z}/J\mathbb{Z} \to \mathbb{R}$$

$\{1, ..., I\}$ indicates that in x-direction no periodic boundary conditions are applied. Hence the populations $\mathsf{F}_1, \mathsf{F}_2, \mathsf{F}_8$ at the left boundary nodes ($i = 1$) and the populations $\mathsf{F}_4, \mathsf{F}_5, \mathsf{F}_6$ at the right boundary nodes ($i = I$) cannot be updated by the standard propagation (1.37). A workaround is offered by the *bounce-back rule*, which approximates homogeneous Dirichlet (no-slip) boundary conditions for the target equation along the walls ($x = 0$ and $x = L$) of the cylinder.
In this case the populations $\mathsf{F}_1, \mathsf{F}_2, \mathsf{F}_8$ and $\mathsf{F}_4, \mathsf{F}_5, \mathsf{F}_6$ are updated by *retroflection*[49] of the collision products, that would otherwise leave the domain. More precisely:

$$\mathsf{F}_k(n+1, i, j) = (1-\omega)\, \mathsf{F}_{\bar{k}}(n, i, j) + \omega\, \mathsf{E}_{\bar{k}}(n, i, j). \tag{1.40}$$

Here $\bar{k}$ denotes the index of the population whose velocity vector is the negative of the velocity vector associated with the population k, i.e. $\binom{s_{\bar{k}x}}{s_{\bar{k}y}} = -\binom{s_{kx}}{s_{ky}}$. Note that the retroflection involves collision products and populations pertaining to the same node. Therefore the translational invariance is not affected and proposition 1.5 is also valid in the case of the cylinder.
Similarly, the one-dimensional equations (1.38) are supplemented by

$$\mathsf{F}_k(n+1, i) = (1-\omega)\, \mathsf{F}_{\bar{k}}(n, i) \;+\; \omega\, \mathsf{E}_{\bar{k}}(n, i), \tag{1.41}$$

which has to be applied whereever (1.38) fails due to the presence of boundaries.
Under the assumptions of proposition 1.7 let us now derive the boundary conditions for the D1P3 scheme from the above bounce-back rule for the D1P2 scheme. Writing down the bounce-back rule for F_2, F_8

$$\begin{aligned}
\mathsf{F}_2(n+1, i) &= (1-\omega)\, \mathsf{F}_6(n, i) \;+\; \omega\, \mathcal{E}_6(0, 0, V(n, i)) \\
\mathsf{F}_8(n+1, i) &= (1-\omega)\, \mathsf{F}_4(n, i) \;+\; \omega\, \mathcal{E}_4(0, 0, V(n, i))
\end{aligned}$$

we find

$$\begin{aligned}
\mathsf{F}_2(n+1, i) - \mathsf{F}_8(n+1, i) &= (1-\omega)\, \Big[\mathsf{F}_6(n, i) - \mathsf{F}_4(n, i)\Big] \\
&\quad + \omega\, \Big[\mathcal{E}_6(0, 0, V(n, i)) - \mathcal{E}_4(0, 0, V(n, i))\Big] \\
\Rightarrow \quad \mathsf{D}_+(n+1, i) &= -\Big[(1-\omega)\, \mathsf{D}_-(n, i) \;+\; \omega\, \mathcal{E}^D_-\big(V(n, i)\big)\Big]
\end{aligned}$$

and similarly $\mathsf{D}_-(n+1, i) \;=\; -\Big[(1-\omega)\, \mathsf{D}_+(n, i) \;+\; \omega\, \mathcal{E}^D_+\big(V(n, i)\big)\Big]$.
As expected, we obtain for D_+, D_- bounce-back type rules differing from the original bounce-back rule only by a flipping of sign. Proceeding analogously, the same boundary conditions are also found for A_- and A_+.

[49] In opposition to the *reflection*, the *retroflection* flips the whole velocity vector and not only its normal component. So, a particle hitting a wall with the velocity $\mathbf{v}_{in}$, is retroflected if it leaves the wall with $\mathbf{v}_{out} = -\mathbf{v}_{in}$, whereas it is reflected if $\mathbf{v}_{out} = \mathbf{v}_{in} - 2\mathbf{n} \cdot \mathbf{v}_{in}$. Here $\mathbf{n}$ denotes the unit normal vector at the position of impact, that points into the wall.

Remark 1.5. Clearly, all statements remain true *mutatis mutandis*, if we assume translational invariance in x-direction. Moreover, if the lattice-Boltzmann algorithm is driven by some external source then the translational invariance of the initial state is only conserved if the force is also translationally invariant with respect to the same direction.

The Stokes equilibrium

In order to become more concrete concerning the reduction of the D2P9 algorithm we have to specify the equilibrium. Let us start with the *Stokes* equilibrium being a linear function of the moments R, U and V introduced in (1.36).

$$\mathcal{E}_k^S(R,U,V) := \mathsf{w}_k \left(R \;+\; 3\,U\,s_{xk} \;+\; 3\,V\,s_{yk} \right). \tag{1.42}$$

Observe that this equilibrium corresponds to the linear part of (1.31). If (1.42) is combined with the parabolic scaling, U and V converge to the x- and y-component of the velocity while the pseudodensity R approximates the pressure.

Lemma 1.1. ***(Shear mode)*** *Consider the D2P9 scheme with the Stokes equilibrium on an one-dimensional grid with either periodic or bounce-back boundary conditions. Assume that the following relations hold true at the initialization $n = 0$. Then they hold true for all iterations:*

i) $\mathsf{F}_1(n,i) \;+\; \mathsf{F}_2(n,i) \;+\; \mathsf{F}_8(n,i) \;=\; 0$

ii) $\mathsf{F}_3(n,i) \;+\; \mathsf{F}_7(n,i) \;+\; \mathsf{F}_9(n,i) \;=\; 0$

iii) $\mathsf{F}_4(n,i) \;+\; \mathsf{F}_5(n,i) \;+\; \mathsf{F}_6(n,i) \;=\; 0$

Proof: The verification is done by recurrence. According to the assumption the relations are true for $n = 0$. So, their correctness has to be checked for the $(n+1)^{\text{th}}$ step provided their validity at the n^{th} step is given.
From i) - iii) we obtain immediately

$$R(n,i) = 0 \qquad \text{and} \qquad U(n,i) = 0.$$

Therefore we conclude:

$$\begin{aligned}
&\mathsf{E}_1(n,i) + \mathsf{E}_2(n,i) + \mathsf{E}_8(n,i)\\
&\qquad = \mathcal{E}_1^S\big(0,0,V(n,i)\big) + \mathcal{E}_2^S\big(0,0,V(n,i)\big) + \mathcal{E}_8^S\big(0,0,V(n,i)\big)\\
&\qquad\qquad = 0 \;+\; \tfrac{1}{12}V(n,i) \;-\; \tfrac{1}{12}V(n,i) \;=\; 0\,.
\end{aligned}$$

Since this relation together with i) holds for all spatial indices i, we find by using the evolution equation (1.38):

$$\begin{aligned}
&\mathsf{F}_1(n+1,i) \;+\; \mathsf{F}_2(n+1,i) \;+\; \mathsf{F}_8(n+1,i) \;=\\
&\qquad (1-\omega)\left[\mathsf{F}_1(n+1,i-1) \;+\; \mathsf{F}_2(n+1,i-1) \;+\; \mathsf{F}_8(n+1,i-1)\right]\\
&\qquad\; + \;\omega\;\; \left[\mathsf{E}_1(n+1,i-1) \;+\; \mathsf{E}_2(n+1,i-1) \;+\; \mathsf{E}_8(n+1,i-1)\right]\\
&\qquad\qquad = \;(1-\omega)\cdot 0 + \omega\cdot 0 \;=\; 0.
\end{aligned}$$

The crucial point is, that the collision products updating the populations occurring in i) come from the same neighbor node. Therefore the recurrence assumption can be applied. Similarly ii) and iii) are proved. In the case of retroflection, equation (1.41) has to be used instead of equation (1.38). ■

Let $v_0 = v_0(x)$ be an initial profile of a shear flow in y-direction and let the coordinates of a regular one-dimensional grid be given by $\{x_i\}_{i\in\mathbb{Z}/I\mathbb{Z}}$ and $\{x_i\}_{i\in\{1,\dots,I\}}$ for the torus and for the cylinder respectively.

Proposition 1.8. *The following initializations of the D2P9 scheme satisfy the assumption of the preceding lemma.*

$$i)\quad \mathsf{F}_k(0,i,j) \;=\; \mathsf{F}_k(0,i) \;=\; \mathcal{E}_k^S\big(0,0,v_0(x_i)\big) \tag{1.43}$$

$$ii)\quad \mathsf{F}_k(0,i,j) \;=\; \mathsf{F}_k(0,i) \;=\; \mathcal{E}_k^S\big(0,0,v_0(x_i)\big) \;+\; \gamma\,\partial_x v_0(x_i)\,\mathsf{w}_k s_{kx}\, s_{ky} \tag{1.44}$$

Remark 1.6. The additional correction term in (1.44) provokes a smoother initial behavior. γ is a coefficient being of no importance in this context.

Now, we are just in the situation, where proposition 1.7 i) is applicable. Utilizing the same notation as in its proof, the Stokes equilibrium reduces to

$$\mathcal{E}_+^{DS}(V) = \tfrac{1}{6}V\;, \qquad \mathcal{E}_0^{DS}(V) = \tfrac{2}{3}V\;, \qquad \mathcal{E}_-^{DS}(V) = \tfrac{1}{6}V\;. \tag{1.45}$$

It should be noticed that we use the D1P3 velocity set $\{-1,\ 0,\ +1\} \equiv \{-,\ 0,\ +\}$ in order to index quantities referring to the D1P3 velocities. The initializations (1.43) and (1.44) take the following form for the D1P3 scheme:

$$i)\quad \mathsf{D}_\sigma(0,i) \;=\; \mathcal{E}_\sigma^{DS}\big(v_0(x_i)\big) \tag{1.46}$$

$$ii)\quad \mathsf{D}_\sigma(0,i) \;=\; \mathcal{E}_\sigma^{DS}\big(v_0(x_i)\big) \;+\; \tfrac{1}{18}\gamma\,\sigma\,\partial_x v_0(x_i) \tag{1.47}$$

D_0 is interpreted as the rest population; $\mathsf{D}_+, \mathsf{D}_-$ represent populations moving to the right and left respectively. Their mass moment (sum of D_-,D_0 and D_+) is exactly V, which is the first order moment of the D2P9 scheme approximating the velocity component in y-direction.

As in the D2P9 case the rest population is associated with the strongest weight $\mathsf{w}_0 = \frac{2}{3}$, the other two populations are equally weighted, i.e. $\mathsf{w}_+ = \mathsf{w}_- = \frac{1}{6}$, such that the isotropy of the scheme is not destroyed and $\sum_\sigma \mathsf{w}_\sigma = 1$. The next theorem recapitulates the whole result we have proved so far.

Theorem 1.1. ***(Reduction theorem for shear modes)*** *Consider the D2P9 scheme with the Stokes equilibrium (1.42) on a rectangular grid discretizing the torus or the cylinder. Furthermore consider the D1P3 scheme with the equilibrium defined in (1.45) on the one-dimensional grid representing the cross-section of the rectangular grid in x-direction. Suppose the D2P9 scheme is initialized by (1.43) or (1.44) and the D1P3 scheme is initialized by (1.46) or (1.47) respectively. Then both schemes produce identical numeric results for the approximate flow profile V.*

Remark 1.7. (Forcing) In the case of external forces, lemma 1.1 remains only true if the source term, which has to be added to the right hand sides of the evolution equation (1.38) and the bounce-back rule (1.41), is of the form $G_y(n,i)\, s_{ky}$. Physically, this reflects the fact that a shear flow in y-direction can only be driven by a force which points in this direction and varies spatially only in the perpendicular direction.

Remark 1.8. If we consider the D2P9 scheme on an one-dimensional grid with I nodes, then the global state space is a vector space of dimension $9I$. The equations i), ii) and iii) of lemma 1.1 define a $6I$ dimensional subspace, which is asserted to be *invariant* under the evolution operator. Note, that this subspace is contained in the larger subspace of dimension $7I$ defined by $R = 0$ and $U = 0$. However, this latter subspace seems to be not invariant under the evolution. An explanation for this can be obtained by the asymptotic analysis yielding an interpretation for quantities like $\mathsf{F}_1 + \mathsf{F}_2 + \mathsf{F}_8$. It turns out that they represent components of the stress tensor, which have to vanish in order to be consistent with the shear flow and and to keep R and U equal to zero.

Remark 1.9. It is possible to prove a stronger version of lemma 1.1 claiming additionally, that F_1, F_5 and F_9 do not change if they are initialized by zero. Note furthermore, that it is not possible to reconstruct the populations of the D2P9 scheme in an unique way from the those of the D1P3 scheme. In that sense the schemes are not equivalent.

Let us now turn to the acoustic modes:

Lemma 1.2. ***(Acoustic mode)*** *Consider the D2P9 scheme with the Stokes equilibrium on an one-dimensional grid with either periodic or bounce-back boundary conditions. Assume that the following relations hold true at the initialization $n = 0$. Then they hold true for all iterations:*

$$
\begin{array}{lllll}
\textit{i)} & \mathsf{F}_2(n,i) & - & \mathsf{F}_8(n,i) & = 0 \\
\textit{ii)} & \mathsf{F}_3(n,i) & - & \mathsf{F}_7(n,i) & = 0 \\
\textit{iii)} & \mathsf{F}_4(n,i) & - & \mathsf{F}_6(n,i) & = 0
\end{array}
$$

Proof: The verification is analogous to the proof of lemma 1.1.
i)-iii) yields immediately

$$V = (\mathsf{F}_2 - \mathsf{F}_8) + (\mathsf{F}_3 - \mathsf{F}_7) + (\mathsf{F}_4 - \mathsf{F}_6) = 0 + 0 + 0 = 0$$

From this we infer:

$$
\begin{aligned}
\mathsf{E}_2(n,i) - \mathsf{E}_8(n,i) &= \mathcal{E}_2^S\big(R(n,i), U(n,i), 0\big) - \mathcal{E}_8^S\big(R(n,i), U(n,i), 0\big) \\
&= \tfrac{1}{36}\big(R(n,i) + 3U(n,i) + 0\big) \; - \; \tfrac{1}{36}\big(R(n,i) + 3U(n,i) - 0\big) \\
&= 0
\end{aligned}
$$

In the same way it is shown that: $\mathsf{E}_3(n,i) - \mathsf{E}_7(n,i) = 0 \quad \mathsf{E}_4(n,i) - \mathsf{E}_6(n,i) = 0$. Since this relation together with i) holds for all spatial indices i, we find by using the evolution equation (1.38):

$$
\begin{aligned}
\mathsf{F}_2(n+1,i) - \mathsf{F}_8(n+1,i) = \;& (1-\omega)\big[\mathsf{F}_2(n,i-1) - \mathsf{F}_8(n,i-1)\big] \\
& + \omega\,\big[\mathsf{E}_2(n,i-1) - \mathsf{E}_8(n,i-1)\big] \\
= \;& (1-\omega)\cdot 0 + \omega \cdot 0 = 0
\end{aligned}
$$

The configuration of the cylinder implies that F_2 is updated by retroflection iff F_8 is updated by retroflection. Therefore we obtain in this case at the left boundary ($i = 1$):

$$\begin{aligned}
\mathsf{F}_2(n+1,1) - \mathsf{F}_8(n+1,1) = &\ (1-\omega)\big[\mathsf{F}_{\bar{2}}(n,1) - \mathsf{F}_{\bar{8}}(n,1)\big] \\
&+ \omega\ \big[\mathsf{E}_{\bar{2}}(n,1) - \mathsf{E}_{\bar{8}}(n,1)\big] \\
= &\ (1-\omega)\big[\mathsf{F}_6(n,2) - \mathsf{F}_4(n,2)\big] \\
&+ \omega\ \big[\mathsf{E}_6(n,2) - \mathsf{E}_4(n,2)\big] \\
= &\ (1-\omega)\cdot 0 + \omega \cdot 0 = 0
\end{aligned}$$

Similarly, ii) and iii) are shown. ■

Let $\rho_0 = \rho_0(x)$ be the initial distribution of the pseudodensity and $\mathbf{u}_0(x,y) = \binom{u_0(x)}{0}$ the initial velocity. If the D2P9 algorithm is initialized by

$$\mathsf{F}_k(0,i,j) \ = \ \mathsf{F}_k(0,i) \ = \ \mathcal{E}_k^S\big(\rho_0(x_i), u_0(x_i), 0\big), \tag{1.48}$$

then the hypothesis of the preceding lemma is fulfilled and proposition 1.7 ii) can be employed. In the case of the acoustic modes the Stokes equilibrium reduces to

$$\mathcal{E}_\sigma^{AS}(R,U) := \mathsf{w}_\sigma\,\big(R + 3\,\sigma\,U\big), \qquad \sigma \in \{-,\,0,\,+\}. \tag{1.49}$$

Consequently, the D1P3 scheme is initialized by

$$\mathsf{A}_\sigma(0,i) \ = \ \mathcal{E}_\sigma^{AS}\big(\rho_0(x_i), u_0(x_i), 0\big). \tag{1.50}$$

The following theorem is the acoustic mode version of the main result, theorem 1.1.

Theorem 1.2. ***(Reduction theorem for acoustic modes)*** *Consider the D2P9 scheme with the Stokes equilibrium (1.42) on a rectangular grid discretizing the torus or the cylinder. Furthermore consider the D1P3 scheme with the equilibrium defined by (1.49) on the one-dimensional grid corresponding to the cross-section of the rectangular grid along the x-axis. If the D2P9 scheme is initialized by (1.48) and the D1P3 scheme by (1.50) then the numeric results for the moments R, U computed according to* (1.36) *and* (1.39) *are identical.*

Remark 1.10. (Forcing) As far as external forces are concerned, lemma 1.2 remains only true if the force term is given by $G_x(n,i)\, s_{kx}$. This corresponds fully to our expectation, since a non-vanishing y-component of the force would automatically excite a non-zero y-velocity violating thus equations i), ii) and iii) of lemma 1.2.

Remark 1.11. Equations i), ii) and iii) also characterize a $6I$ dimensional subspace of the global state space. The situation is analogous to that described in remark 1.8.

The Navier-Stokes Oseen equilibrium

The *Navier-Stokes* equilibrium for the D2P9 model is defined by

$$\begin{aligned}
\mathcal{E}_k^N(\rho,u,v) := &\ \mathsf{w}_k\Big[\rho + 3\tbinom{s_{kx}}{s_{ky}}\cdot\tbinom{u}{v}\Big] \\
&+\ \tfrac{3}{2}\,\gamma\,\mathsf{w}_k\,\Big[\ 3\,\Big(\tbinom{s_{kx}}{s_{ky}}\cdot\tbinom{u}{v}\Big)\ \Big(\tbinom{s_{kx}}{s_{ky}}\cdot\tbinom{u}{v}\Big)\ -\ \tbinom{u}{v}\cdot\tbinom{u}{v}\ \Big].
\end{aligned} \tag{1.51}$$

Linearizing the terms quadratic in $\binom{u}{v}$ around some constant velocity field $\binom{a_x}{a_y}$ leads to the *Oseen* equilibrium

$$\begin{aligned} \mathcal{E}_k^O(\rho,u,v) := {} & \mathsf{w}_k\Big[\rho + 3\tbinom{s_{kx}}{s_{ky}}\cdot\tbinom{u}{v}\Big] \\ & + \; 3\,\gamma\,\mathsf{w}_k\Big[\,3\,\Big(\tbinom{s_{kx}}{s_{ky}}\cdot\tbinom{a_x}{a_y}\Big)\;\Big(\tbinom{s_{kx}}{s_{ky}}\cdot\tbinom{u}{v}\Big)\;-\;\tbinom{a_x}{a_y}\cdot\tbinom{u}{v}\,\Big]. \end{aligned} \tag{1.52}$$

Is it possible to reduce the D2P9 scheme also if it is used with the Navier-Stokes or the Oseen equilibrium?
Remember the proof of lemma 1.1. The point, where the equilibrium came explicitly into play, was the verification of the following relations:

$$\begin{aligned} \mathcal{E}_1(0,0,V) \;+\; \mathcal{E}_2(0,0,V) \;+\; \mathcal{E}_8(0,0,V) \;&=\; 0 \\ \mathcal{E}_3(0,0,V) \;+\; \mathcal{E}_7(0,0,V) \;+\; \mathcal{E}_9(0,0,V) \;&=\; 0 \\ \mathcal{E}_4(0,0,V) \;+\; \mathcal{E}_5(0,0,V) \;+\; \mathcal{E}_6(0,0,V) \;&=\; 0 \end{aligned} \tag{1.53}$$

Let us check, whether the Navier-Stokes equilibrium satisfies these conditions.

$$\begin{aligned} \mathcal{E}_1^N(R,U,V) \;&=\; \tfrac{1}{9}\big(R+3U\big) + \tfrac{\gamma}{6}\big(2U^2 - V^2\big) \\ \mathcal{E}_2^N(R,U,V) \;&=\; \tfrac{1}{36}\big(R+3U+3V\big) + \tfrac{\gamma}{24}\big(2U^2+6UV+2V^2\big) \\ \mathcal{E}_8^N(R,U,V) \;&=\; \tfrac{1}{36}\big(R+3U-3V\big) + \tfrac{\gamma}{24}\big(2U^2-6UV+2V^2\big) \end{aligned}$$

$$\begin{aligned} \Rightarrow\;\; & \mathcal{E}_1^N(R,U,V) \;+\; \mathcal{E}_2^N(R,U,V) \;+\; \mathcal{E}_8^N(R,U,V) \;=\; \tfrac{1}{6}R \;+\; \tfrac{1}{2}U \;+\; \tfrac{\gamma}{2}U^2 \\ \Rightarrow\;\; & \mathcal{E}_1^N(0,0,V) \;+\; \mathcal{E}_2^N(0,0,V) \;+\; \mathcal{E}_8^N(0,0,V) \;=\; 0 \end{aligned}$$

So, the Navier-Stokes equilibrium satisfies the first equation of (1.53). Repeating the computation for the Oseen equilibrium

$$\begin{aligned} \mathcal{E}_1^O(R,U,V) \;&=\; \tfrac{1}{9}\big(R+3U\big) \;+\; \tfrac{\gamma}{3}\big(2a_xU - a_yV\big) \\ \mathcal{E}_2^O(R,U,V) \;&=\; \tfrac{1}{36}\big(R+3U+3V\big) \;+\; \tfrac{\gamma}{12}\big(2a_xU+3a_xV+3a_yU+2a_yV\big) \\ \mathcal{E}_8^O(R,U,V) \;&=\; \tfrac{1}{36}\big(R+3U-3V\big) \;+\; \tfrac{\gamma}{12}\big(2a_xU-3a_xV-3a_yU+2a_yV\big) \end{aligned}$$

$$\begin{aligned} \Rightarrow\;\; & \mathcal{E}_1^O(R,U,V) \;+\; \mathcal{E}_2^O(R,U,V) \;+\; \mathcal{E}_8^O(R,U,V) \;=\; \tfrac{1}{6}R \;+\; \tfrac{1}{2}U \;+\; \sigma a_x\, U \\ \Rightarrow\;\; & \mathcal{E}_1^O(0,0,V) \;+\; \mathcal{E}_2^O(0,0,V) \;+\; \mathcal{E}_8^O(0,0,V) \;=\; 0 \end{aligned}$$

yields the same result. The other two equations of (1.53) are verified in the same manner. Therefore lemma 1.1 is also valid for the Navier-Stokes and Oseen equilibrium. The reduced equilibria come out as

$$\mathcal{E}_\sigma^{DN}(V) \;:=\; \mathsf{w}_\sigma\, V\,, \tag{1.54}$$

$$\mathcal{E}_\sigma^{DO}(V) \;:=\; \mathsf{w}_\sigma V \;+\; 3\gamma\,\mathsf{w}_\sigma\,\sigma\,a_x\,V\,. \tag{1.55}$$

In complete analogy, theorem 1.1 therefore also holds for the case of the Navier-Stokes and Oseen equilibrium.

Theorem 1.3. ***(Reduction theorem for shear modes)*** *Consider the D2P9 scheme with the Navier-Stokes equilibrium (1.51) or the Oseen equilibrium (1.52) on a rectangular grid discretizing the torus or the cylinder. Furthermore consider the D1P3 scheme on the one-dimensional grid representing the cross-section of the rectangular grid in x-direction. Suppose that the D2P9 scheme is initialized by the equilibrium (1.51) or (1.52) respectively, where $\rho_0 = 0$, $u_0 = 0$ and the velocity profile are inserted. Likewise the D1P3 scheme is initialized by the equilibrium (1.54) or (1.55) respectively using the same velocity profile v_0. Then both schemes produce identical numeric results for V which is calculated according to* (1.36) *and* (1.39)*.*

Remark 1.12. It should be noticed, that the reduced Navier-Stokes equilibrium (1.54) is equal to the reduced Stokes equilibrium (1.45). Therefore simulations of shear flows with the D2P9 scheme yield the same numerical result for V independently of the chosen equilibrium (Stokes/Navier-Stokes). (It is not asserted that the populations are equal!) This parallels the fact, that shear flows are simultaneously solutions of the Stokes and the Navier-Stokes equation (cf. remark 1.2).

The whole procedure can now be repeated for the acoustic modes. Recall that the proof of lemma 1.2 can only be carried out if the equilibrium satisfies the following equations:

$$\begin{aligned}
\mathcal{E}_2(R,U,0) &- \mathcal{E}_8(R,U,0) = 0,\\
\mathcal{E}_3(R,U,0) &- \mathcal{E}_7(R,U,0) = 0,\\
\mathcal{E}_4(R,U,0) &- \mathcal{E}_6(R,U,0) = 0.
\end{aligned} \tag{1.56}$$

However these equations are just fulfilled by the Navier-Stokes equilibrium. Hence for the Oseen equilibrium this kind of reduction cannot be performed.

1.3 An abstract framework of numerical analysis

In this section we sketch the background underlying our analysis of lattice-Boltzmann methods. The presented approach is a standard technique splitting the analysis into separate studies of consistency and stability. These concepts are widely used in numerous variations. Here we want to elaborate a guideline of the main idea which may be clouded in lengthy proofs.

The general setup of a numerical scheme

Consider a mapping $M : X \to Y$, not necessarily linear, where X and Y denote two normed vector spaces. A numerical scheme, like a finite difference scheme, for solving the equation

$$M(x) = 0 \in Y \tag{1.57}$$

generally introduces the following objects:

- the domain $\mathcal{H} \subset (0, 1]$ of the discretization parameter h with an accumulation point at 0,

- two families of normed vector spaces $\mathcal{X} = \{(X_h, \|\cdot\|_{X_h}) \; : \; h \in \mathcal{H}\}$ and $\mathcal{Y} = \{(Y_h, \|\cdot\|_{Y_h}) \; : \; h \in \mathcal{H}\}$ discretizing[50] X and Y respectively,
- a family of operators $\mathcal{M} = \{M_h : X_h \to Y_h \mid h \in \mathcal{H}\}$ discretizing M.

The mapping M_h is referred to as the *discretization* or *algorithm* on X_h. Furthermore an element $x_h \in X_h$ is called a *numeric solution* of (1.57) if

$$M_h(x_h) = 0 \in Y_h. \tag{1.58}$$

Assuming that (1.58) admits a unique solution for every $h \in \mathcal{H}$, the numerical scheme defines a *fiber*[51] (alternatively denoted as section or path) through $\mathcal{X}$. Usually there is not a direct inclusion relation between X_{h_1}, X_{h_2} and Y_{h_1}, Y_{h_2} for $h_1 \neq h_2$. In order to compare elements between different spaces of the same family one can define linear restriction[52] and continuation (extension) operators using interpolation for instance. Moreover, a restriction operator $R_h : X \to X_h$ and an continuation operator $C_h : X_h \to X$ must be available for comparison with the solution of (1.57).

Fundamental notions

Definition 1.3. ***(Convergence)*** *Denoting by x, x_h the unique solution of* (1.57) *and* (1.58) *respectively, the numerical scheme is called* convergent *if*

$$\lim_{h\to 0} \|x_h - R_h x\|_{X_h} = 0 \qquad \text{or} \qquad \lim_{h\to 0} \|C_h x_h - x\|_X = 0.$$

If the convergence is specified more precisely by

$$\|x_h - R_h x\|_{X_h} \in \mathrm{O}(h^\alpha) \qquad \text{or} \qquad \|C_h x_h - x\|_X \in \mathrm{O}(h^\alpha), \tag{1.59}$$

the exponent α is referred to as convergence rate[53]*.*

Let us underline that this definition makes only sense if the norm $\|\cdot\|_{X_h}$ and the continuation operator C_h are reasonably defined. In order to obtain the same convergence rate from both equations in (1.59), $R_h C_h$ should correspond to the identity on X_h and the operator norm of R_h should be uniformly bounded with respect to h. Under these circumstances the right equation in (1.59) implies the left one:

$$\mathrm{O}(h^\alpha) \ni \underbrace{\|R_h\|_{\mathcal{L}(X,X_h)}}_{\mathrm{O}(1)} \|C_h x_h - x\|_X \geq \|R_h C_h x_h - R_h x\|_{X_h} = \|x_h - R_h x\|_{X_h}$$

[50] Do not confuse X_h and Y_h with the discretization of the domain on which the elements of X and Y live as functions. So X_h does not represent any spatial or temporal grid but a function-space over a grid.

[51] A fiber is a map assigning to each $h \in \mathcal{H}$ an element $x_h \in X_h$. By means of restriction and extension operators we can define continuous fibers. $(x_h)_h$ is continuous at $h_0 \in \mathcal{H}$ with respect to $(R_{h_1,h_2} : X_{h_1} \to X_{h_2})_{h_1,h_2}$ if there exists for any $\epsilon > 0$ a $\delta > 0$ such that $\|R_{h_0,h} x_{h_0} - x_h\|_{X_h} < \epsilon$ for all h with $|h - h_0| < \delta$.

[52] In the context of conforming finite elements the term projection is preferred since in this case $X_h \subset X$.

[53] Usually the convergence rate refers to the maximal exponent α that can be taken, such that $\|x_h - R_h x\|_{X_h} \in \mathrm{O}(h^\alpha)$ but $\|x_h - R_h x\|_{X_h} \notin \mathrm{o}(h^{\alpha+\beta})$ for every $\beta > 0$. Observe that such an exponent need not exist necessarily. For instance, $\exp(-1/h)$ vanishes more rapidly than any power.

The fibers $(x_h)_{h\in\mathcal{H}}$ and $(R_hx)_{h\in\mathcal{H}}$ can be interpreted as elements of the product space[54]

$$\mathbf{X} := \prod_{h\in\mathcal{H}} X_h.$$

Definition 1.4. ***(Asymptotic similarity)*** *Two fibers* $\mathbf{u} = (u_h)_{h\in\mathcal{H}} \in \mathbf{X}$ *and* $\mathbf{v} = (v_h)_{h\in\mathcal{H}} \in \mathbf{X}$ *are called* asymptotically similar *if*

$$\lim_{h\to 0} \|u_h - v_h\|_{X_h} = 0.$$

The asymptotic similarity is (at least) of order α *if* $\|u_h - v_h\|_{X_h} \in \mathrm{O}(h^\alpha)$ *for* $h \downarrow 0$.

Thus convergence means that the fibers $(x_h)_h$ and $(R_hx)_h$ are asymptotically similar.

If the mapping M is continuous, convergence should imply $M(C_hx_h) \to 0 \in Y$ as $h \downarrow 0$. In many situations M is only defined on a subspace $\tilde{X} \subset X$ and it is complicated to construct C_h such that $C_hX_h \subset \tilde{X}$ instead of $C_hX_h \subset X$. So M may not be applicable to C_hx_h. Therefore it is more convenient to adopt the converse perspective. As R_hx is close to x_h for small h, $M_h(R_hx)$ should be close to $0 \in Y_h$.

Definition 1.5. ***(Consistency with exact solution)*** *Let* x *be the solution of* (1.57)*. A numerical scheme is* consistent *if*

$$\lim_{h\to 0} \|M_h(R_hx) - \underbrace{M_h(x_h)}_{=0}\|_{X_h} = \lim_{h\to 0} \|M_h(R_hx)\|_{X_h} = 0,$$

meaning that $\big(M_h(R_hx)\big)_h$ *is asymptotically similar to the zero-fiber in* $(Y_h)_h$.

In the case of lattice-Boltzmann methods the solution of the target equation corresponding to (1.57) cannot be inserted directly into the discrete equations, where the population functions occur as primary variables. Therefore we need a more general notion of consistency.

Definition 1.6. ***(Consistency of a fiber)*** *A fiber* $(\hat{x}_h)_{h\in\mathcal{H}}$ *is called* consistent *of order* α *if*

$$\|M_h(\hat{x}_h)\|_{X_h} \in \mathrm{O}(h^\alpha).$$

Alternatively, we say that the residual *of* $\hat{x}_h$*, which is just the quantity* $M_h(\hat{x}_h)$*, is of magnitude* $\mathrm{O}(h^\alpha)$.

The order of consistency depends essentially on how M_h is defined. The equation $\tilde{M}_h(x_h) = 0$ with $\tilde{M}_h := h^\beta M_h$ is equivalent to (1.58) but the order of consistency becomes different.
$(\hat{x}_h)_h$ can be considered as an *approximate* numeric solution. We refer to it also as *comparison* or *prediction function.* Besides consistency, another property is of paramount importance.

[54]There are elements in $\mathbf{X}$ which have no finite norm with respect to $|||\cdot||| := \sup_h \|\cdot\|_{X_h}$. In order to obtain a normed vector space one has to consider the subset of normable elements $\tilde{\mathbf{X}} := \{\mathbf{x} \in \mathbf{X} \;:\; \sup_h \|x_h\|_{X_h} < \infty\}$. Observe, however, that $\mathbf{X}$ is metrizable by using the seminorms $\|\cdot\|_{X_h}$ to define the metric $d(\mathbf{x},\mathbf{y}) = \sum_h 2^{-n(h)} \frac{\|\mathbf{x}-\mathbf{y}\|_{X_h}}{1+\|\mathbf{x}-\mathbf{y}\|_{X_h}}$ if $\mathcal{H}$ is countable and $n(h) \in \mathbb{N}$ is an index univocally assigned to each $h \in \mathcal{H}$.

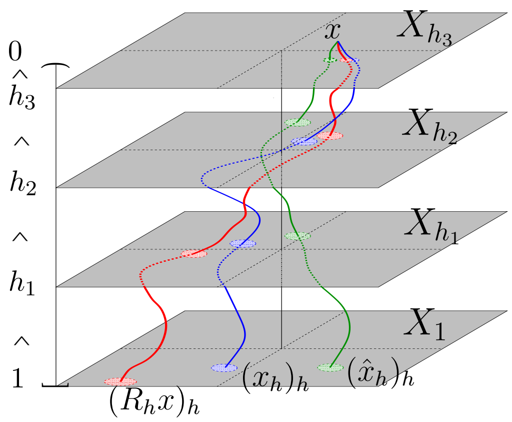


Figure 1.2: Schematic diagram illustrating four spaces of a family $\mathcal{X}$ of vector spaces parametrized by $h \in \mathcal{H} \subset (0,1]$. There are three fibers represented by the curved lines: the projected solution $(R_h x)_h$, the fiber of numeric solutions $(x_h)_h$ and another consistent fiber $(\hat{x}_h)_h$. As these fibers are asymptotically similar, they meet each other in $x \in X$, which could be taken as X_0. It is a deficiency of the illustration that all spaces seem to be isomorphic to one another. This is particularly not correct for examples with numerical background where the dimension grows as h tends to 0.

Definition 1.7. ***(Stability)*** *The numeric scheme is* stable *if M_h is invertible for all $h \in \mathcal{H}$ and if the inverse M_h^{-1} is uniformly Lipschitz continuous with respect to h. So there exists a constant $K > 0$ independent of h such that for $y_1, y_2 \in Y_h$*

$$\|M_h^{-1}(y_2) - M_h^{-1}(y_1)\|_{X_h} < K\|y_2 - y_1\|_{Y_h},$$

which means that $\mathrm{Lip}_{M_h^{-1}} < K$ *for all $h \in \mathcal{H}$.*

The following proposition yields the crucial link between asymptotic similarity (convergence), consistency and stability. Observe that we argue here only in principle ignoring complicating details that may occur in concrete situations.

Proposition 1.9. *Assume that the numeric scheme is* stable*. Then any fiber $(\hat{x}_h)_h$* consistent *of order α is also* asymptotically similar *of equal order to the fiber $(x_h)_h$ of the numeric solutions. In particular, there exists a constant $K > 0$ independent of h such that*

$$\|\hat{x}_h - x_h\|_{X_h} \ < \ K\, h^{\alpha} \qquad \textit{for } h \textit{ sufficiently small.}$$

Proof: Utilizing that $M_h^{-1} \circ M_h$ represents the identity operator on X_h, we obtain:

$$\begin{aligned} \|\hat{x}_h - x_h\|_{X_h} &= \left\| \left(M_h^{-1} \circ M_h\right)(x_h) - \left(M_h^{-1} \circ M_h\right)(\hat{x}_h) \right\|_{X_h} \\ &\leq \mathrm{Lip}_{M_h^{-1}} \left\| M_h(\hat{x}_h) - \underbrace{M_h(x_h)}_{=0} \right\|_{Y_h} = \mathrm{Lip}_{M_h^{-1}} \left\| M(\hat{x}_h) \right\|_{Y_h} \end{aligned} \tag{1.60}$$

The asserted estimate follows directly from the stability and the definition of the O-notation entailing the existence of a constant $L > 0$ with $\left\|M(\hat{x}_h)\right\|_{Y_h} < Lh^{\alpha}$ for h sufficiently small. ∎

(Warning) The computation of the residual order amounts to a consistency analysis that is quite easy to do in many situations. The hope is that the residual order foretells the order of asymptotic similarity or the convergence rate respectively. The latter quantity decides about the accuracy of a scheme and is therefore of practical interest but, unfortunately, more difficult to obtain.
At the outset, it is usually overhasty to infer the convergence rate just from examing the consistency rate, unless precise information about the behavior of M_h^{-1} is available. In general, the Lipschitz constant of the inverse operator may not be uniformly bounded with respect to h. Hence, the consistency rate (order of the residual) and the rate of asymptotic similarity (convergence) can differ from each other. A trivial increase of the consistency order by multiplying (1.58) with h^β – as mentioned above – has an impact on the Lipschitz constant so that the convergence rate remains invariant against such rescalings. In fact, it is quickly seen that the orders gained by the residual are canceled due to the h-dependency of the Lipschitz constant.

As being very general, the estimate in (1.60) is rather crude. Hence it might indicate an order of asymptotic similarity that actually is not realized by the algorithm. For instance, in the case of a linear problem, where $r_h = M_h \hat{x}_h$ denotes the residual, it could happen that

$$\mathrm{O}\Big(\|M_h^{-1} r_h\|_{X_h}\Big) \neq \mathrm{O}\Big(\|M_h^{-1}\|_{\mathcal{L}(Y_h,X_h)}\, \|r_h\|_{Y_h}\Big) \qquad \text{for } h \to 0,$$

because the estimate does not take into account the specific structure of M_h^{-1} and r_h. So the quantity on the left hand side may converge while the quantity on the right hand side is divergent. Related phenomena will be encountered in the forthcoming chapters.

Stability of explicit schemes

The explicit Euler discretization of an initial value problem like

$$z(0) = a, \qquad \partial_t z + Az = 0$$

leads naturally to the following recursion

$$z_h(0) = a_h, \qquad z_h(n+1) = E_h z_h(n). \tag{1.61}$$

$E_h : Z_h \to Z_h$ denotes the discrete evolution operator, that is given by $E_h = I_h - \Delta t_h A_h$ in the case of the Euler scheme. Here Δt_h represents the time step and $A_h : Z_h \to Z_h$ discretizes $A : Z \to Z$ operating on a normed vector space Z, which contains the functions of the spatial argument. Obviously, (1.61) does not comply immediately with the form of equation (1.58), which is typical for stationary problems. Since lattice-Boltzmann methods are subsumed in the class of explicit iterations, we want to see how the notion of stability concretizes in this context.

For this we cast the recursion (1.61) into the form of (1.58). As the discretization refers to a compact time interval $[0,T]$, only a finite number $N_h \in \mathbb{N}$ of time steps

$\Delta t_h > 0$ is to be considered. Both N_h and Δt_h depend on h and satisfy $N_h \Delta t_h \approx T$. So the iterative equations for $n \in \{0, 1, ..., N_h\}$ can be summarized into a single matrix equation.

$$\underbrace{\begin{pmatrix} 1 & & & & \\ -E_h & 1 & & & \\ & -E_h & 1 & & \\ & & \ddots & \ddots & \\ & & & -E_h & 1 \end{pmatrix} \begin{pmatrix} z_h(0) \\ z_h(1) \\ z_h(2) \\ \vdots \\ z_h(N_h) \end{pmatrix} = \begin{pmatrix} a_h \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{M_h(x_h)=b_h}$$

In contrast to (1.58) we now consider a non-zero right hand side. Obviously, the variable x_h must be identified with an array storing the values of all iterations. This suggests the setting

$$X_h = \underbrace{Z_h \times \ldots \times Z_h}_{(N_h+1) \text{ times}},$$

which is equipped with the norm $\|x_h\|_{X_h} = \max_{n=0}^{N_h} \|z_h(n)\|_{Z_h}$. The inverse of the two-band matrix representing M_h is given by a triangular matrix:

$$M_h^{-1} = \begin{pmatrix} 1 & & & & \\ E_h & 1 & & & \\ E_h^2 & E_h & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ E_h^{N_h} & \ldots & E_h^2 & E_h & 1 \end{pmatrix}.$$

So far, these relations hold independently of whether E_h acts linearly or non-linearly on its argument.
Let us now focus on the linear case. May $x_h = \big(x_h(0), x_h(1), \ldots, x_h(N_h)\big)$ denote the numeric solution $M_h x_h = b_h$, and may $\hat{x}_h$ be an approximate numeric solution (prediction function) with $M_h \hat{x}_h = b_h + r_h$ where $r_h = \big(r_h(0), r_h(1), \ldots, r_h(N_h)\big)$ stands for the residual. Repeating the short computation in the proof of proposition 1.9, but exploiting the linearity, yields:

$$\begin{aligned} \|x_h - \hat{x}_h\|_{X_h} = \|M_h^{-1} M_h x_h - M_h^{-1} M_h \hat{x}_h\|_{X_h} &= \big\|M_h^{-1}(\underbrace{M_h x_h}_{b_h} - \underbrace{M_h \hat{x}_h}_{b_h + r_h})\big\|_{X_h} \\ &= \big\|M_h^{-1} r_h\big\|_{X_h} \leq \|M_h^{-1}\|_{\mathcal{L}(X_h)} \|r_h\|_{X_h} \end{aligned} \tag{1.62}$$

The prediction function and the numeric solution are asymptotically similar if the right hand side of (1.62) vanishes for $h \downarrow 0$. In order to estimate $\|M_h^{-1}\|_{\mathcal{L}(X_h)}$ we consider $v_h \in X_h$ with $\|v_h\|_{X_h} = 1$ entailing $\|v_h(k)\|_{Z_h} \leq 1$ for all $k \in \{0, ..., N_h\}$:

$$\begin{aligned} \|M_h^{-1} v_h\|_{X_h} &= \max_{n=0}^{N_h} \Big\| \sum_{k=0}^{n} E_h^k v_h(k) \Big\|_{Z_h} \leq \max_{n=0}^{N_h} \sum_{k=0}^{n} \big\|E_h^k v_h(k)\big\|_{Z_h} \\ &\leq \max_{n=0}^{N_h} \sum_{k=0}^{n} \big\|E_h^k\big\| \underbrace{\big\|v_h(k)\big\|_{Z_h}}_{\leq 1} \leq (N_h + 1) \max_{n=0}^{N_h} \big\|E_h^n\big\|_{\mathcal{L}(Z_h)}. \end{aligned}$$

Thus we conclude

$$\|M_h^{-1}\|_{\mathcal{L}(X_h)} \leq (N_h+1) \max_{n=0}^{N_h} \left\|E_h^n\right\|_{\mathcal{L}(Z_h)}. \tag{1.63}$$

Since E_h^0 corresponds to the identity on Z_h, we see that $\max_{n=0}^{N_h} \left\|E_h^n\right\|_{\mathcal{L}(Z_h)} \geq 1$. This shows that the right hand side of (1.63) becomes arbitrarily large for $h \downarrow 0$ due to the unbounded growth of the factor N_h. The estimate insinuates that, in contrast to definition 1.7, it might not be reasonable to require $\|M_h^{-1}\|_{X_h}$ to be uniformly bounded with respect to h. This motivates the following modified definition of stability.

Definition 1.8. ***(Stability)*** *An explicit iterative scheme is called stable if there is an h-independent constant $K > 0$ such that for all $h \in \mathcal{H}$*

$$\max_{n=0}^{N_h} \left\|E_h^n\right\|_{\mathcal{L}(Z_h)} < K \quad \textit{being equivalent to} \quad \sup_{h\in\mathcal{H}} \max_{n=0}^{N_h} \left\|E_h^n\right\|_{\mathcal{L}(Z_h)} < \infty.$$

Under this condition, any prediction function whose residual vanishes faster than $1/N_h$, i.e. $r_h \in \mathrm{o}(1/N_h) = \mathrm{o}(\Delta t_h)$ for fixed T, is asymptotically similar to the numeric solution. Due to the factor N_h in (1.63), the consistency order is generally greater than the order of asymptotic similarity in opposition to the statement in proposition 1.9.

Asymptotic approach – construction of a prediction function

Equation (1.60) and (1.62) indicate that the convergence analysis consists of two relatively independent steps[55]: estimating Lipschitz constants (stability part) and finding an appropriate prediction function producing a residual of sufficiently high order (consistency part).

The prediction function is searched in an *ansatz space* $\mathbf{A} \subset \mathbf{X}$, whose elements share a special, preferably simple structure or permit a particular representation. A popular ansatz space is given by regular asymptotic expansions

$$\hat{x}_h = \hat{x}_h^{[m]} = R_h x^{(0)} + h R_h x^{(1)} + \ldots + h^m R_h x^{(m)}. \tag{1.64}$$

This choice is inspired by the idea to approximate $h \mapsto x_h$ by polynomials, constituting the most simple class of functions. Thanks to the choice of smooth coefficient functions $x^{(0)}, \ldots, x^{(m)}$ it is possible to evaluate $M_h \hat{x}_h$ by applying differential operators to $x^{(0)}, \ldots, x^{(m)}$ using Taylor expansion. The transition from difference

[55] Although this 'division of labor' works satisfactorily in many situations, we emphasize once more that the standard estimate of the matrix vector product in (1.62) may be crude. Hence $\|M_h^{-1}\hat{x}_h\|_{X_h}$ might reveal a higher order than $\|M_h^{-1}\|_{\mathcal{L}(X_h)}\|\hat{x}_h\|_{X_h}$ because the estimate completely ignores the interaction between the residual r_h and the inverse operator M_h^{-1}. Such situations appear in particular, if $M_h\hat{x}_h = r_h$ is just a very condensed notation for several equations that differ with respect to their order of consistency (e.g. different equations prescribing boundary conditions, initial conditions, the evolution or bulk behavior etc.). This will be one of the major issues in chapter 7.

operators building up M_h to differential operators composing M enables us to connect (1.58) with (1.57).

While the discrete equation (1.58) is introduced to supply approximate solutions for the original equation (1.57) in practical applications, the analysis of (1.58) takes the reverse direction. Departing from the solution x of (1.57), an approximate solution of the scheme (1.58) (prediction function $(\hat{x}_h)_h$) is constructed, that is asymptotically similar to the output of the scheme $(x_h)_h$. In standard situations[56] $x^{(0)}$ turns out to be x such that $x^{(0)} + hx^{(1)} + \ldots + h^m x^{(m)}$ converges to x as $h \to 0$. Transitivity (triangle inequality) entails the asymptotic similarity of $(x_h)_h$ and $(R_h x_h)_h$ which gives the desired convergence.

Analogy with perturbation theory

The asymptotic approach also represents a convenient tool to analyze perturbed problems, which, formally, resemble equation (1.58) where we set $\epsilon \in (0, 1]$ instead of $h \in \mathcal{H}$. The main difference is that the family of operators $(M_\epsilon)_\epsilon$ refers to a single, non-indexed domain and range space, i.e. $M_\epsilon : X \to Y$. Furthermore we associate with M_ϵ mostly a differential operator, while M_h stands for discrete operator in matrix form.

If there exists a limit problem $M_0 x = 0$, that does not differ from $M_\epsilon x = 0$ in a very distinct qualitative manner, we speak of a *regularly perturbed problem.* A prominent example may be the frictionless harmonic oscillator versus a weakly damped one considered over a *bounded* time interval. In this case the movement of the attenuated oscillator differs from those of the ideal one mainly by a slight decrease in the amplitude, which gets almost unnoticeable for smaller and smaller friction constants. Although dissipation destroys periodicity and energy conservation, the damped oscillations converge *uniformly*[57] to the undamped ones as the friction tends to zero.

The lattice-Boltzmann equation pertains to another class of perturbed problems where the operator M_0 is formally not available or leads to (ill-posed) problems whose nature changes abruptly for $\epsilon = 0$. Those problems are referred to as *singularly perturbed.*

Analogously to (1.64), we construct prediction functions in the form of asymptotic expansions $\hat{x}_\epsilon = x^{(0)} + \epsilon x^{(1)} + ... + \epsilon^m x^{(m)}$ to analyze perturbed problems. The asymptotic order functions $x^{(0)}, ..., x^{(m)}$ are determined such that $\hat{x}$ solves the perturbed problem with a small residual.

Stability by energy-type estimates

Let us pass over to the stability analysis, which often proceeds differently from the discrete case in so far as it is less obvious to write down the inverse operator

[56] Observe that the case of lattice-Boltzmann schemes is somewhat more complicated as the the scheme is based on primary variables different from those of the target equation.

[57] This statement does not hold with respect to an unbounded time interval.

M_ϵ^{-1} explicitly. Generally the equation $M_\epsilon x_\epsilon = 0$ decomposes into several equations. This situation is frequently encountered in the guise of boundary or initial conditions.
Assuming that the image space Y of M_ϵ is the Cartesian product of two normed vector spaces U and V the perturbed problem becomes

$$M_\epsilon(x_\epsilon) = 0 \in U_\epsilon \times V_\epsilon \quad \Leftrightarrow \quad \begin{cases} G_\epsilon(x_\epsilon) = 0 \in U_\epsilon \\ F_\epsilon(x_\epsilon) = 0 \in V_\epsilon \end{cases}, \tag{1.65}$$

with $(G_\epsilon : X \to U)_\epsilon$ and $(H_\epsilon : X \to V)_\epsilon$ being two ϵ-dependent operator families. A prediction function $\hat{x}_\epsilon$ is said to have a *residual order* α and β in the first and second equation respectively if there exist constants $K, L > 0$ independent of ϵ, such that

$$\|G_\epsilon(\hat{x}_\epsilon)\|_U \;\leq\; K\,\epsilon^\alpha\,, \qquad \|H_\epsilon(\hat{x}_\epsilon)\|_V \;\leq\; L\,\epsilon^\beta\,.$$

By energy methods it is possible to derive inequalities of the subsequent type, which implies the Lipschitz continuity of the inverse operator[58]. Let us therefore assume that the following estimate holds true for all $z_\epsilon, \hat{z}_\epsilon \in X_\epsilon$ with constants $C_\epsilon, D_\epsilon > 0$ depending possibly on ϵ.

$$\|z_\epsilon - \hat{z}_\epsilon\|_X \;\leq\; C_\epsilon \|G_\epsilon(z_\epsilon) - G_\epsilon(\hat{z}_\epsilon)\|_U \;+\; D_\epsilon \|H_\epsilon(z_\epsilon) - H_\epsilon(\hat{z}_\epsilon)\|_V \tag{1.66}$$

For the solution x_ϵ of (1.65) and an approximate solution $\hat{x}_\epsilon$ satisfying (1.3), we can hence infer:

$$\begin{aligned} \|x_\epsilon - \hat{x}_\epsilon\|_X \;&\leq\; C_\epsilon \| \overbrace{G_\epsilon(x_\epsilon)}^{=0} - G_\epsilon(\hat{x}_\epsilon)\|_U \;+\; D_\epsilon \| \overbrace{H_\epsilon(x_\epsilon)}^{=0} - H_\epsilon(\hat{x}_\epsilon)\|_V \\ &=\; C_\epsilon \|G_\epsilon(\hat{x}_\epsilon)\|_U \;+\; D_\epsilon \|H_\epsilon(\hat{x}_\epsilon)\|_V \\ &\leq\; C_\epsilon\, K\, \epsilon^\alpha \;+\; D_\epsilon\, L\, \epsilon^\beta \end{aligned}$$

If C_ϵ, D_ϵ are found to be independent of ϵ, we obtain asymptotic similarity of order $\min\{\alpha, \beta\}$.

Two examples

The following examples may serve as direct illustration. The first one discusses Euler's well known polygon method in the light of the concepts mentioned so far. In particular we want to show, how the way of writing down the scheme may influence stability and consistency results which – if taken for their own – may give rise to confusion. The second example is not of discrete nature and anticipates the analysis in chapter 3. It shows also the impact of the scaling on Lipschitz bounds for the inverse operator.

[58] Observe, that we do not need the continuity of M_ϵ itself. This would be reflected by an inequality of the form $\|G_\epsilon(z_\epsilon) - G_\epsilon(\hat{z}_\epsilon)\|_U + \|H_\epsilon(z_\epsilon) - H_\epsilon(\hat{z}_\epsilon)\|_V \leq A_\epsilon \|z_\epsilon - \hat{z}_\epsilon\|_X$ with some constant $A_\epsilon > 0$. In the prominent cases, where M_ϵ involves differential operators, such an estimate does usually not hold with respect to the desired norm.

Example 1:
Consider the initial value problem on a bounded time interval:

$$u(0) = u_0 \in \mathbb{R}, \qquad \tfrac{\mathrm{d}}{\mathrm{d}t} u(t) = f\big(u(t)\big) \quad t \in [0,T]. \tag{1.67}$$

To ensure the existence of a unique solution $u : [0,T] \to \mathbb{R}$, the function $f : \mathbb{R} \to \mathbb{R}$ is supposed to be globally Lipschitz-continuous, meaning that $|f(x) - f(y)| \le L|x-y|$ holds for all $x, y \in \mathbb{R}$ with some $L > 0$. After fixing a time step $h = \frac{T}{N}$ for some $N \in \mathbb{N}$, *Euler's polygon method* can be used to solve (1.67) numerically. The approximate solution $\mathbf{v} = [v_0, v_1, ..., v_n] \in \mathbb{R}^{N+1}$ is computed according to the following recursion:

$$v_0 = u_0 \qquad v_k = v_{k-1} + hf(v_k) \quad 1 \le k \le N\ . \tag{1.68}$$

v_k is taken as an approximation of $u(kh)$. Let us now write the scheme in a form corresponding to (1.58). For this we introduce an (invertible) function $F_h : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$, whose unique zero is given by $\mathbf{v}$ as defined in (1.68). Thus equation (1.68) should become equivalent to

$$F_h(\mathbf{v}) = \mathbf{0} \in \mathbb{R}^{N+1}\ .$$

F_h and its inverse are defined via recursions such that all requirements are satisfied by construction:

$$F_h : \mathbf{y} \mapsto \mathbf{r}, \qquad \begin{cases} r_0 = & y_0 - u_0 \\ r_k = & \frac{y_k - y_{k-1}}{h} - f(y_{k-1}), \qquad 1 \le k \le N \end{cases}$$

$$F_h^{-1} : \mathbf{r} \mapsto \mathbf{y}, \qquad \begin{cases} y_0 = & r_0 + u_0 \\ y_k = & y_{k-1} + hf(y_{k-1}) + hr_k, \qquad 1 \le k \le N. \end{cases}$$

The output argument $\mathbf{r} = [r_0, ..., r_N] \in \mathbb{R}^{N+1}$ is considered as the residue of the input argument. Comparing with the general setting, observe that we have set $X_h = Y_h = \mathbb{R}^{N+1}$ which shall be equipped with the maximum norm.

Proposition 1.10. *The Lipschitz constant of F_h^{-1} is uniformly bounded with respect to h, or equivalently, it is an $\mathrm{O}(1)$ quantity for $h \to 0$.*

Proof: In order to show the assertion, suppose that $\mathbf{y}, \mathbf{z}, \mathbf{r}, \mathbf{s} \in \mathbb{R}^{N+1}$ satisfy

$$F_h^{-1}(\mathbf{r}) = \mathbf{y} \qquad \text{and} \qquad F_h^{-1}(\mathbf{s}) = \mathbf{z}\ .$$

We have to find a constant $C > 0$, such that the following estimate holds true for any $N \in \mathbb{N}$.

$$\|\mathbf{y} - \mathbf{z}\|_\infty \ \le\ C\|\mathbf{r} - \mathbf{s}\|_\infty$$

The definition of F_h^{-1} yields directly $|y_0 - z_0| = |r_0 - s_0|$. Let us estimate the other components of $\mathbf{y}, \mathbf{z}$ recursively employing the Lipschitz continuity of f.

$$\begin{aligned} |y_k - z_k| &\le\ |y_{k-1} - z_{k-1}| \ +\ h|f(y_{k-1}) - f(z_{k-1})| \ +\ h|r_k - s_k| \\ &\le\ (1 + hL)|y_{k-1} - z_{k-1}| \ +\ h\|\mathbf{r} - \mathbf{s}\|_\infty \end{aligned}$$

Observe that a recursion of the type $\alpha_k = \gamma\alpha_{k-1} + \beta$ is solved by $\alpha_k = \gamma^k \alpha_0 + \beta \sum_{j=0}^{k-1} \gamma^j$ so that

$$|y_k - z_k| \le\ (1 + hL)^k |y_0 - z_0| \ +\ h\|\mathbf{r} - \mathbf{s}\|_\infty \sum_{k=0}^{N-1} (1 + hL)^k\ .$$

Apply the inequality $1 + x \le \exp(x)$ for $x \ge 0$. Then $(1 + hL)^k \le \big(\exp(hL)\big)^k = \exp(khL) \le \exp(TL)$ and

$$|y_k - z_k| \le\ \exp(TL)|r_0 - s_0| \ +\ T\exp(TL)\|\mathbf{r} - \mathbf{s}\|_\infty.$$

This estimate is valid for $k \in \{1, ..., N\}$. In the case of $k = 0$ we have $|y_0 - z_0| = |r_0 - s_0| \le \|\mathbf{r} - \mathbf{s}\|_\infty$. As $TL \ge 0$ and hence $\exp(TL) \ge 1$, we get:

$$\|\mathbf{y} - \mathbf{z}\|_\infty \ \le\ 2\max\{1, T\}\exp(TL)\ \|\mathbf{r} - \mathbf{s}\|_\infty\ .$$

Thus we obtain $\mathrm{Lip}_{F_h^{-1}} \le 2\max\{1, T\}\exp(TL)$, which is independent of h. ∎

Let us denote by $\hat{\mathbf{v}} := [u(0h), ..., u(Nh)]$ the solution of (1.67) projected onto the time grid, which can also be seen as prediction function. For $u \in \mathcal{C}^2([0,T])$ the standard consistency analysis using Taylor expansion yields the residual order 1, i.e.

$$\|F_h(\hat{\mathbf{v}})\| \leq Kh = \mathrm{O}(h)$$

with some suitably chosen constant $K > 0$. Since the Lipschitz constant of F_h^{-1} is (at least) O(1) we obtain from proposition 1.9, adapted to our context, that the output $\mathbf{v}$ of the Euler method is first order asymptotically similar to $\hat{\mathbf{v}}$, i.e.

$$\|\mathbf{v} - \hat{\mathbf{v}}\|_\infty \;\leq\; 2\max\{1,T\}\exp(TL)\,K\,h\,.$$

According to definition 1.3 the scheme has the convergence order 1 which agrees with the residual order.

Instead of F_h, we could alternatively introduce $\tilde{F}_h$

$$\tilde{F}_h : \mathbf{y} \mapsto \mathbf{r}, \qquad \begin{cases} r_0 = y_0 - u_0 & \\ r_k = y_k - y_{k-1} - hf(y_{k-1}), & 1 \leq k \leq N \end{cases}$$

$$\tilde{F}_h^{-1} : \mathbf{r} \mapsto \mathbf{y}, \qquad \begin{cases} y_0 = r_0 + u_0 & \\ y_k = y_{k-1} + hf(y_{k-1}) + r_k, & 1 \leq k \leq N. \end{cases}$$

Notice that $\tilde{F}_h$ is obtained from F_h, if the equations for $r_1, ..., r_n$ are multiplied by h. As the initial condition is exactly satisfied, the residual of $\hat{\mathbf{v}}$ turns out to be of second order (with respect to $\tilde{F}_h$!). This might feign also second order convergence – but by repeating the estimate of the Lipschitz constant for $\tilde{F}_h^{-1}$, we get an $\mathrm{O}(h^{-1})$ bound, since the factor h in front of the r_k is missing in the recursive definition. Therefore the additional order that would be found in the consistency analysis is canceled in the stability analysis.

The reason for finding the Lipschitz bound of magnitude $\mathrm{O}(h^{-1})$ is the same as the occurrence of the factor $(N_h + 1)$ in equation (1.63). Observe that we did not introduce the evolution operator which could be done for an alternative discussion.

Example 2:
Let us pick up the IVP for the parabolicly scaled D1P3 lattice-Boltzmann equation treated in detail throughout chapter 3.

$$\left.\begin{aligned} \mathsf{f}_\epsilon(0,\cdot) \;-\; \mathsf{f}_\epsilon^{\text{init}} \;&=\; 0 \\ \partial_t \mathsf{f}_\epsilon \;+\; \epsilon^{-1} S \partial_x \mathsf{f}_\epsilon \;+\; \epsilon^{-2} J_\epsilon \mathsf{f}_\epsilon \;&=\; 0 \end{aligned}\right\} \tag{1.69}$$

The initial condition can be formulated by virtue of an operator $G_\epsilon : \mathcal{C}^1_{\text{per}}(\mathcal{X}_T, \mathcal{F}) \to \mathcal{C}^1_{\text{per}}(\mathcal{X}, \mathcal{F})$, that evaluates the argument at $t = 0$ and subtracts the initial value, which is given by a function $\mathsf{f}_\epsilon^{\text{init}} \in \mathcal{C}^1_{\text{per}}(\mathcal{X}, \mathcal{F})$. Defining the operator

$$H_\epsilon : \mathcal{C}^1_{\text{per}}(\mathcal{X}_T, \mathcal{F}) \to \mathcal{C}_{\text{per}}(\mathcal{X}_T, \mathcal{F}), \qquad H_\epsilon := \partial_t + \epsilon^{-1} S \partial_x - \epsilon^{-2} J_\epsilon,$$

the lattice-Boltzmann equation is written as $H_\epsilon \mathsf{f}_\epsilon = 0$. We consider the spaces equipped with the $\mathcal{L}^2(\mathcal{X};\mathcal{F})$ norm and the $\mathcal{C}\big(\mathcal{T};\mathcal{L}^2(\mathcal{X};\mathcal{F})\big)$ norm in the case of time dependent functions. Note that in this example, the involved spaces and norms do not depend on the parameter ϵ.
In section 3.1 the following estimate is derived, that is similar to (1.66) except for the squared norms. Moreover, we need not consider differences $(z_\epsilon - \hat{z}_\epsilon)$, since G_ϵ, H_ϵ are affine linear.

$$\|\mathsf{z}_\epsilon(t,\cdot)\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;\leq\; \mathrm{e}^T \,\|B\|_{\mathcal{L}(\mathcal{F})}\, \|\mathsf{z}(0,\cdot)\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;+\; \mathrm{e}^T\, T\, \|B\|_{\mathcal{L}(\mathcal{F},\mathcal{F})}\, \|\mathsf{q}\|^2_{\mathcal{C}(\mathcal{T};\mathcal{L}^2(\mathcal{X};\mathcal{F}))}$$

The estimate holds for all $\mathsf{z}_\epsilon \in \mathcal{C}^1_{\text{per}}(\mathcal{X}_T, \mathcal{F})$ satisfying $H_\epsilon \mathsf{z}_\epsilon = \mathsf{q}$ with some source function $\mathsf{q} \in \mathcal{C}_{\text{per}}(\mathcal{X}_T, \mathcal{F})$. The initial value problem (1.69) is not modified, if we multiply the second line by ϵ^2. Replacing thus H_ϵ by $\epsilon^2 H_\epsilon = \epsilon^2 \partial_t + \epsilon S \partial_x - J_\epsilon$ the residual order of any ansatz is augmented by 2. But for z_ϵ satisfying $\epsilon^2 H_\epsilon \mathsf{z}_\epsilon = \mathsf{q}$ we obtain the estimate,

$$\|\mathsf{z}_\epsilon(t,\cdot)\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;\leq\; \mathrm{e}^T \,\|B\|_{\mathcal{L}(\mathcal{F})}\, \|\mathsf{z}(0,\cdot)\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;+\; \epsilon^{-4}\mathrm{e}^T\, T\, \|B\|_{\mathcal{L}(\mathcal{F},\mathcal{F})}\, \|\mathsf{q}\|^2_{\mathcal{C}(\mathcal{T};\mathcal{L}^2(\mathcal{X};\mathcal{F}))}$$

that is distinct from the first one by an factor ϵ^{-4} before the second addend. So, the four orders of ϵ, that are gained due to the squared norm, are directly lost by the additional factor ϵ^{-4}. Therefore the order of asymptotic similarity is not ameliorated. Note however, that the estimate might not be sharp.

Comparison of asymptotic similarity and equivalence

In this paragraph we return to the notion of asymptotic similarity to discuss it in the context of asymptotic equivalence – a similar concept frequently found in literature. Let us start with a definition from a slightly different perspective.

Definition 1.9. *Two sequences $(x_n)_n$ and $(y_n)_n$ in a normed vector space X are called* asymptotically similar*, written as $(x_n)_n \sim (y_n)_n$, if:*

$$\lim_{n\to\infty} \|x_n - y_n\| = 0 \qquad \Leftrightarrow \qquad x_n = y_n + r_n \quad \textit{with} \quad \lim_{n\to\infty} r_n = 0 \in X.$$

Asymptotic similarity generalizes the idea of convergence in a natural way. Convergence could be considered as the approximation of a "static object", namely the limit.
In contrast, asymptotic similarity means, that something is approximated, which appears as a "dynamic object", since it does vary by itself while approaching it. It should be stressed, that asymptotic similarity does not presuppose the convergence of $(x_n)_n$ and $(y_n)_n$; they may be arbitrary (divergent) sequences.

Asymptotic similarity represents an *equivalence relation* in the set of X-valued sequences. So, any sequence is trivially asymptotically similar to itself (*reflexivity*) and by definition the relation is *symmetric* as well. Moreover, we see by an application of the triangle inequality, that $(x_n)_n \sim (y_n)_n$ and $(y_n)_n \sim (z_n)_n$ imply $(x_n)_n \sim (z_n)_n$ (*transitivity*). Likewise, two sequences converging two the same limit are always asymptotically similar, since any convergent sequence is asymptotically similar to the constant sequence being equal to its limit.

Closely related to the notion of asymptotic similarity is the idea of *asymptotic equivalence*. Roughly speaking, two convergent sequences are *asymptotically equivalent*, if they approach their common limit with the same speed. This means, that their difference should go faster to zero than the distance to the limit.
Also uniformly divergent[59] sequences are referred to as asymptotically equivalent if they grow faster than their difference does even though the difference might become unbounded too (e.g. factorial and Stirling formula; $x_n := n^2 + n, y_n := n^2$).

Definition 1.10.

i) In a normed vector space X two sequences $(x_n)_n, (y_n)_n \subset X \setminus \{0\}$ are called asymptotically equivalent*, abbreviated by $(x_n)_n \asymp (y_n)_n$, if:*

$$\lim_{n\to\infty} \frac{\|x_n - y_n\|}{\|x_n\|} = 0 \tag{1.70}$$

[59] A sequence $(x_n)_n$ in a normed vector space X is called uniformly divergent, if there exists for any $M > 0$ an index $n(M) \in \mathbb{N}$, such that $\|x_n\| > M$ if $n > n(M)$. They are convergent to infinity in the extended (compactified) real or complex number set.

ii) In particular, two sequences of complex numbers $(w_n)_n, (z_n)_n \subset \mathbb{C} \setminus \{0\}$ *are called* asymptotically equivalent *if*

$$\lim_{n\to\infty} \frac{w_n}{z_n} = 1 \tag{1.71}$$

Although equation (1.70) is not symmetric in x_n and y_n, asymptotic equivalence is also an equivalence relation[60] as suggested by the terminology. For sequences of complex numbers definition i) and ii) are equivalent. i) is a generalization of ii) towards vector spaces, where a division operation is not available.

Note that nullsequences are always asymptotically similar but need not be asymptotically equivalent. Conversely, there are uniformly divergent sequences being asymptotically equivalent but not similar. If the range of two sequences is comprised in a compact interval not containing zero, then both notions imply each other. The precise relation between asymptotic similarity and asymptotic equivalence is clarified by the following proposition.

Proposition 1.11. *In a normed vector space* X *two sequences* $(x_n)_n, (y_n)_n \subset X \setminus \{0\}$ *are* asymptotically equivalent, *if and only if* $\left(\frac{x_n}{\|x_n\|}\right)_n$ *and* $\left(\frac{y_n}{\|y_n\|}\right)_n$ *are* asymptotically similar *and* $\lim\limits_{n\to\infty} \frac{\|x_n\|}{\|y_n\|} = 1$.

Proof:

"$\Rightarrow$" Applying $\big|\|a\| - \|b\|\big| \le \|a-b\|$ for $a, b \in X$, shows $\lim\limits_{n\to\infty} \|y_n\|/\|x_n\| = 1$:

$$\left|1 - \frac{\|y_n\|}{\|x_n\|}\right| = \left| \left\|\frac{x_n}{\|x_n\|}\right\| - \left\|\frac{y_n}{\|x_n\|}\right\| \right| \le \left\|\frac{x_n}{\|x_n\|} - \frac{y_n}{\|x_n\|}\right\| = \underbrace{\frac{\|x_n - y_n\|}{\|x_n\|}}_{\xrightarrow[n\to\infty]{} 0}$$

[60]The reflexivity is obvious. Suppose equation (1.70) is true. Let us check the symmetry. Using $\lim\limits_{n\to\infty} \|y_n\|/\|x_n\| = \lim\limits_{n\to\infty} \|x_n\|/\|y_n\| = 1$ (see proof of prop. 1.11) we get:

$$0 = 1 \cdot 0 = \lim_{n\to\infty} \frac{\|x_n\|}{\|y_n\|} \lim_{n\to\infty} \frac{\|x_n - y_n\|}{\|x_n\|} = \lim_{n\to\infty} \frac{\|x_n\|}{\|y_n\|} \cdot \frac{\|x_n - y_n\|}{\|x_n\|} = \lim_{n\to\infty} \frac{\|x_n - y_n\|}{\|y_n\|}$$

And this yields the symmetry: $\lim\limits_{n\to\infty} \frac{\|x_n - y_n\|}{\|x_n\|} = 0 \quad \Leftrightarrow \quad \lim\limits_{n\to\infty} \frac{\|y_n - x_n\|}{\|y_n\|} = 0$

Transitivity: Assume $\lim\limits_{n\to\infty} \frac{\|x_n - y_n\|}{\|x_n\|} = 0, \quad \lim\limits_{n\to\infty} \frac{\|y_n - z_n\|}{\|y_n\|} = 0 \quad \overset{?}{\Longrightarrow} \quad \lim\limits_{n\to\infty} \frac{\|x_n - z_n\|}{\|x_n\|} = 0$
Exploiting once more $\lim\limits_{n\to\infty} \|y_n\|/\|x_n\| = 1$, we obtain,

$$0 = 1 \cdot 0 = \lim_{n\to\infty} \frac{\|y_n\|}{\|x_n\|} \lim_{n\to\infty} \frac{\|y_n - z_n\|}{\|y_n\|} = \lim_{n\to\infty} \frac{\|y_n\|}{\|x_n\|} \frac{\|y_n - z_n\|}{\|y_n\|} = \lim_{n\to\infty} \frac{\|y_n - z_n\|}{\|x_n\|}$$

and so we find eventually:

$$\lim_{n\to\infty} \frac{\|x_n - z_n\|}{\|x_n\|} \le \lim_{n\to\infty} \frac{\|x_n - y_n\|}{\|x_n\|} + \lim_{n\to\infty} \frac{\|y_n - z_n\|}{\|x_n\|} = 0 + 0 = 0$$

This is used to get asymptotic similarity of the normalized sequences:

$$\begin{aligned}\left\|\frac{x_n}{\|x_n\|}-\frac{y_n}{\|y_n\|}\right\| &\le \left\|\frac{x_n}{\|x_n\|}-\frac{y_n}{\|x_n\|}\right\|+\left\|\frac{y_n}{\|x_n\|}-\frac{y_n}{\|y_n\|}\right\| \\ &= \frac{\|x_n-y_n\|}{\|x_n\|}+\left|\frac{1}{\|y_n\|}-\frac{1}{\|x_n\|}\right|\|y_n\| \\ &= \underbrace{\frac{\|x_n-y_n\|}{\|x_n\|}}_{\text{by hyp. } \xrightarrow[n\to\infty]{} 0}+\underbrace{\left|1-\frac{\|y_n\|}{\|x_n\|}\right|}_{\xrightarrow[n\to\infty]{} 0}\end{aligned}$$

"$\Leftarrow$" Conversely we find:

$$\begin{aligned}\frac{\|x_n-y_n\|}{\|x_n\|}=\left\|\frac{x_n}{\|x_n\|}-\frac{y_n}{\|x_n\|}\right\| &\le \left\|\frac{x_n}{\|x_n\|}-\frac{y_n}{\|y_n\|}\right\|+\left\|\frac{y_n}{\|y_n\|}-\frac{y_n}{\|x_n\|}\right\| \\ &= \underbrace{\left\|\frac{x_n}{\|x_n\|}-\frac{y_n}{\|y_n\|}\right\|}_{\text{by hyp. } \xrightarrow[n\to\infty]{} 0}+\underbrace{\left|1-\frac{\|y_n\|}{\|x_n\|}\right|}_{\text{by hyp. } \xrightarrow[n\to\infty]{} 0}\end{aligned}$$

■

Appendix: Landau symbols and asymptotic expansions

Asymptotic analysis deals with objects depending on parameters. Mostly these objects are functions that are obtained as solutions of parameter-dependent differential equations. The main task of asymptotic analysis is to investigate the dependence near certain 'critical' values.
Numerical schemes naturally fit into this context such that asymptotic analysis turns out to be a pertinent tool for their analysis. They are formulated as discrete equations characterized by a discretization parameter $h > 0$. The critical value is obtained for $h = 0$. In this case the scheme is not defined; nevertheless it must reveal a converging behavior for $h \to 0$.

Here we recall some basic notation frequently used in asymptotics.

Definition 1.11. ***(Landau symbols)*** *Consider a metric space P, the parameter domain, and let $U \subset X$ and $V \subset Y$ be subsets of two normed vector spaces X and Y. For two functions $f, g : P \times U \to V$ we define*

1) ***(big Oh)*** $f(p,u) \in \mathrm{O}\big(g(p,u)\big)$ *as* $p \to p_0 \in P$

$$:\Leftrightarrow \quad \exists\, K > 0,\ \exists \delta > 0,\ \forall\, p \in B_\delta(p_0) : \|f(p,u)\| < K\|g(p,u)\|$$

2) ***(little oh)*** $f(p,u) \in \mathrm{o}\big(g(p,u)\big)$ *as* $p \to p_0 \in P$

$$:\Leftrightarrow \quad \forall\, \epsilon > 0,\ \exists \delta > 0,\ \forall\, p \in B_\delta(p_0) : \|f(p,u)\| < \epsilon\|g(p,u)\|$$

Observe that the setting of this definition is quite general (e.g. f, g may have only the parameter argument). Furthermore the inequalities in the definition might hold uniformly in the sense that K, δ and ϵ can be chosen independently of u (locally/globally). Additionally, the function g could be independent of u.

Strictly speaking $\mathrm{O}\big(g(p,x)\big)$ and $\mathrm{o}\big(g(p,x)\big)$ designate function sets, where the first set bears the structure of an equivalence class. The Landau symbols are often used in a sloppy manner. So they abusively stand for representants of the corresponding function sets whenever it is of no importance to be more precise.

Definition 1.12. ***(Asymptotic sequences and expansions)*** *Consider a metric space P, the parameter domain, and let $U \subset X$ and $V \subset Y$ be subsets of two normed vector spaces X and Y. A sequence of functions $f_n : P \times U \to V, n \in \{0, 1, 2, ..., N\}$ with $N \in \mathbb{N}_0 \cup \{\infty\}$ is an* asymptotic *sequence for $p \to p_0 \in P$ if*

$$f_{n+1}(p,u) \in \mathrm{o}\big(f_n(p,u)\big) \qquad \textit{for } p \to p_0.$$

The sum of an asymptotic sequence $\sum_n a_n f_n(p,u)$ is called an asymptotic expansion*.*

For more detailed explanations we refer to [48]. Additionally we recommend [53], where the subtle difference between *asymptotic power expansions* and power series is addressed in chapter 9 section 6.

Chapter 2

Scalings and singular limits on the basis of the D1P2 model

Although lattice-Boltzmann equations are not of much interest *per se*, they enjoy important applications in combination with different scalings. In the first part of this chapter we pick up the thread of section 1.1. The intention of the introductory section 2.1 is to stress the following points:

- The significance of the scaling, where the *hyperbolic* scaling is opposed to the *parabolic* scaling.

- The relation between the equilibrium operator and the target equation. This is highlighted using the example of an one-moment-preserving and two-moment-preserving equilibrium operator.

Let us remind that the target equation is formally found by introducing appropriate moments, defining the relevant physical quantities. Expressing the lattice-Boltzmann system in terms of these and further moments, the so-called *equivalent moment system* can be derived which allows to read off the target equation (or target system). However, the choice of the right moments is, in general, neither unique nor obvious. It can only be done by physical intuition or by employing some preknowledge of the system, which may be obtained in a systematic way by asymptotic analysis.

As a special feature, the D1P2 lattice-Boltzmann equation being actually a 2×2 system permits a reformulation into a scalar equation, which is singularly perturbed like the lattice-Boltzmann equation itself. In section 2.2 this circumstance is exploited to investigate the singular limit resulting from the parabolic scaling. The analysis is based on Fourier series leading to a family of singularly perturbed ODEs for the coefficient functions, which reflect the temporal behavior of the original PDE. Particular interest is devoted to the initial layer exhibited by the solutions of these ODEs. A two-scale expansion is motivated and employed in section 2.3 to derive a precise description of the initial layer.

Recall that the (normalized) velocity space $\mathcal{S}$ of the D1P2 model is given by $\{-1, 1\}$. Hence $\mathsf{s} := (-1, 1)^\top \quad \Rightarrow \quad \mathsf{s}^2 = 1$. By reason of isotropy the weights are equally distributed with respect to the discrete velocities, i.e. $\mathsf{w} := (\frac{1}{2}, \frac{1}{2})^\top$.

2.1 Hyperbolic versus parabolic scaling

Hyperbolic scaling for balance and conservation laws

The lattice-Boltzmann equation in *hyperbolic scaling* is generally given by

$$\partial_t \mathsf{f} + \mathsf{s}\partial_x \mathsf{f} = -\epsilon^{-1}\omega\big[\mathsf{f} - E\mathsf{f}\big] + q\mathsf{w} \ , \tag{2.1}$$

where ϵ denotes the scaling parameter and q is a prescribed source. Assuming the D1P2 model as the underlying discrete velocity set, equation (2.1) changes concretely into the system

$$\begin{aligned} \partial_t \mathsf{f}_1 - \partial_x \mathsf{f}_1 &= -\epsilon^{-1}\omega\big[\mathsf{f}_1 - (E\mathsf{f})_1\big] + \tfrac{1}{2}q \ , \qquad (2.2)\\ \partial_t \mathsf{f}_2 + \partial_x \mathsf{f}_2 &= -\epsilon^{-1}\omega\big[\mathsf{f}_2 - (E\mathsf{f})_2\big] + \tfrac{1}{2}q \ . \qquad (2.3)\end{aligned}$$

The component f_1 is the density of the particles moving in left direction, while f_2 represents those particles heading to the right. Note that the equations are coupled via the equilibrium $E\mathsf{f}$.
It turns out, that this system can be connected to the following hyperbolic conservation law[1] with a continuously differentiable *flux function*[2] f

$$\partial_t v + \partial_x\big[f(v)\big] = q \ , \tag{2.4}$$

if the equilibrium operator E is suitably chosen. In order to show this, let us introduce the mass moment u, which is physically the total mass density, and the first moment ϕ, whose meaning becomes more clear before long.

$$u := \langle \mathsf{f}, 1\rangle = \mathsf{f}_1 + \mathsf{f}_2, \qquad \phi := \langle \mathsf{f}, \mathsf{s}\rangle = \mathsf{f}_2 - \mathsf{f}_1 \ . \tag{2.5}$$

The mass moment occurs in the equilibrium operator E that we define[3] by

$$E\mathsf{f} := \langle \mathsf{f}, 1\rangle \mathsf{w} + f\big(\langle \mathsf{f}, 1\rangle\big)\mathsf{s}\mathsf{w} = u\mathsf{w} + f(u)\mathsf{s}\mathsf{w} \ . \tag{2.6}$$

Since the moments of the equilibrium $E\mathsf{f}$ are given by

$$\langle E\mathsf{f}, 1\rangle = u = \langle \mathsf{f}, 1\rangle, \qquad \langle E\mathsf{f}, \mathsf{s}\rangle = f(u)$$

we see that E conserves the mass moment. In contrast, the first moment is not preserved[4]. Let us now pass over to the *equivalent moment system*. This is obtained by performing an algebraic transformation of the lattice-Boltzmann system

[1] If the source q vanishes, we speak of a *proper* conservation law, otherwise the term *balance law* is also used.

[2] Two well-known examples are the linear advection equation with $f(v) = av$ (a constant) and Burgers' equation with $f(v) = \frac{1}{2}v^2$.

[3] Observe that (2.6) is the most general equilibrium in the case of the D1P2 model which depends only on the mass moment u and conserves it also.

[4] Otherwise this would imply $\phi = f(u)$; but this means that the first moment ϕ would be entirely determined by u. Since the variables u, ϕ are equivalent to the populations $\mathsf{f}_1, \mathsf{f}_2$ and since the latter ones are independent of each other, this would be a contradiction. However, the fact that the first moment of the equilibrium yields the flux anticipates the close relation between the first moment ϕ and the flux, see equation (2.9).

by expressing $\mathsf{f}_1, \mathsf{f}_2$ in terms of u, ϕ. Adding equation (2.2) and (2.3), i.e. applying $\langle\cdot, 1\rangle$ to (2.1), yields the evolution equation for u:

$$\begin{aligned} \partial_t\langle\mathsf{f},1\rangle + \partial_x\langle\mathsf{sf},1\rangle &= \epsilon^{-1}\omega\Big[\langle E\mathsf{f},1\rangle - \langle\mathsf{f},1\rangle\Big] + q\langle\mathsf{w},1\rangle \\ \Leftrightarrow \qquad \partial_t u + \partial_x\phi &= \epsilon^{-1}\omega[u-u] + q = q\,. \end{aligned}$$

Likewise, by subtracting (2.2) from (2.3), i.e. applying $\langle\cdot, \mathsf{s}\rangle$ to (2.1), the evolution equation for ϕ is found:

$$\begin{aligned} \partial_t\langle\mathsf{f},\mathsf{s}\rangle + \partial_x\langle\mathsf{sf},\mathsf{s}\rangle &= \epsilon^{-1}\omega\Big[\langle E\mathsf{f},\mathsf{s}\rangle - \langle\mathsf{f},\mathsf{s}\rangle\Big] + q\overbrace{\langle\mathsf{w},\mathsf{s}\rangle}^{=0} \\ \Leftrightarrow \qquad \partial_t\phi + \partial_x u &= \epsilon^{-1}\omega\Big[f(u) - \phi\Big] \end{aligned}$$

So we obtain the subsequent equivalent moment system

$$\begin{aligned} \partial_t u + \partial_x\phi &= q\,, && (2.7) \\ \partial_t\phi + \partial_x u &= -\epsilon^{-1}\omega\big[\phi - f(u)\big]\,. && (2.8) \end{aligned}$$

Observe that this system corresponds just to the *relaxation formulation*[5] pertaining to (2.4), that has been introduced by Jin and Xin in [29]. Solving the second equation (2.8) for ϕ

$$\phi = f(u) - \epsilon\tau\big(\partial_x u + \partial_t\phi\big) \tag{2.9}$$

where we have set $\tau = \omega^{-1}$ and plugging this into the first equation (2.7), produces finally

$$\partial_t u + f'(u)\partial_x u = q + \epsilon\tau\big(\partial_t\partial_x\phi + \partial_x^2 u\big)\,. \tag{2.10}$$

It is possible to eliminate ϕ, such that we obtain a closed equation for u. Differentiating (2.7) with respect to t yields an expression for $\partial_t\partial_x\phi$. So we end up with

$$\begin{aligned} \partial_t u + f'(u)\partial_x u &= q + \epsilon\tau\big(-\partial_t^2 u + \partial_t q + \partial_x^2 u\big)\,, \\ \Leftrightarrow \qquad \epsilon\tau\partial_t^2 u + \partial_t u + f'(u)\partial_x u - \epsilon\tau\partial_x^2 u &= q + \epsilon\tau\partial_t q\,. \end{aligned} \tag{2.11}$$

Thus the mass moment u satisfies the balance law (2.4) if we formally set ϵ to zero. In fact, u is expected to converge to the solution v of (2.4) under the assumptions, that ϕ and its derivatives remain bounded for $\epsilon \downarrow 0$ and that the balance law (2.4) is *stable* under small perturbations.

In general, equation (2.11) determines u in a unique way if supplemented by initial and boundary conditions. Once u is known, ϕ can be computed by means of (2.9). Differently from u, the first moment ϕ is determined as solution of an ordinary differential equation depending on the space variable x as parameter. Equation

[5]The name *relaxation* is motivated by the observation, that the quantity ϕ *relaxes* in the space-homogeneous case towards the flux. This means that it is a solution of the ODE $\dot\phi + \epsilon^{-1}\omega\phi = f(u)$. Starting from the balance law (2.4), as Jin and Xin did it, ϕ appears as an artificially introduced variable. Our derivation of the relaxation system shows, that it is equivalent to a kinetic system, where we understand by a *kinetic system* a PDE system that is formally similar to the Boltzmann equation (like the lattice-Boltzmann equation). In the kinetic framework ϕ is interpreted as a first moment.

(2.9) assigns also a clear physical meaning to ϕ, because it approximates the flux $f(u)$. Let us point out, that equation (2.8) (if multiplied by ϵ) and equation (2.11) represent *singularly perturbed* differential equations, since the (highest) occurring derivatives are preceded by the scaling factor ϵ.

Equilibrium with two conserved moments

In order to illustrate how the equilibrium influences the target equation, let us consider the lattice-Boltzmann equation with another equilibrium:

$$E\mathsf{f} = \langle \mathsf{f}, 1\rangle \mathsf{w} + \langle \mathsf{f}, \mathsf{s}\rangle \mathsf{s}\mathsf{w} = u\mathsf{w} + \phi \mathsf{s}\mathsf{w}\,. \tag{2.12}$$

In contrast to (2.6) this equilibrium conserves both moments since we have

$$\langle E\mathsf{f}, 1\rangle = u = \langle \mathsf{f}, 1\rangle, \qquad \langle E\mathsf{f}, \mathsf{s}\rangle = \phi = \langle \mathsf{f}, \mathsf{s}\rangle\ .$$

Due to this circumstance the ϵ-dependence disappears, when we transform the lattice-Boltzmann equation into the equivalent moment system

$$\partial_t u + \partial_x \phi = q\ , \tag{2.13}$$
$$\partial_t \phi + \partial_x u = 0\ . \tag{2.14}$$

It should be noticed that also this system decouples, resulting in a separate *wave equation* for u and ϕ. In order check this, (2.14) is differentiated with respect to x and solved for $\partial_x\partial_t\phi$. Then it is inserted into (2.13) being also differentiated with respect to x. Similarly, the equation for ϕ is found.

$$\begin{aligned} \partial_t^2 u - \partial_x^2 u &= \partial_t q \\ \partial_t^2 \phi - \partial_x^2 \phi &= -\partial_x q \end{aligned}$$

The D1P2 model is too restricted to admit an adjustable wave speed. In this regard the D1P3 model accords more flexibility which is illuminated briefly by the subsequent excursion.

Equilibrium with two conserved moments for the D1P3 model

Let us shortly digress from the D1P2 model to consider an equilibrium analogous to (2.12) for the D1P3 model already discussed in 1.1.2. Recall that s and w are now given by

$$\mathsf{s} = (-1, 0, 1)^\top, \qquad \mathsf{w} = \left(\tfrac{1}{2\theta}, \tfrac{\theta-1}{\theta}, \tfrac{1}{2\theta}\right)^\top,$$

entailing $\mathsf{s}^2 \neq 1$ and $\langle \mathsf{w}, \mathsf{s}^2\rangle = \frac{1}{\theta}$. The moments u, ϕ are defined as above; additionally a second order moment

$$\psi := \langle \mathsf{f}, \mathsf{s}^2 - \theta^{-1}\rangle$$

is introduced. An equilibrium preserving u and ϕ is obviously given by

$$E\mathsf{f} = \langle \mathsf{f}, 1\rangle \mathsf{w} + \langle \mathsf{f}, \mathsf{s}\rangle \theta \mathsf{s}\mathsf{w} = u\mathsf{w} + \phi \mathsf{s}\mathsf{w}\ .$$

Applying the moment operators to the lattice-Boltzmann equation and observing

$$\begin{aligned} \langle \mathsf{s}\mathsf{f}, \mathsf{s}\rangle &= \langle \mathsf{f}, \mathsf{s}^2\rangle = \langle \mathsf{f}, \mathsf{s}^2 - \theta^{-1}\rangle + \theta^{-1}\langle \mathsf{f}, 1\rangle = \psi + \theta^{-1} u\,, \\ \langle \mathsf{s}\mathsf{f}, \mathsf{s}^2 - \theta^{-1}\rangle &= \langle \mathsf{f}, \mathsf{s}^3 - \mathsf{s}\theta^{-1}\rangle = (1-\theta^{-1})\langle \mathsf{f}, \mathsf{s}\rangle = (1-\theta^{-1})\phi\,, \\ \langle E\mathsf{f}, \mathsf{s}\rangle &= u\,\underbrace{\langle \mathsf{w}, \mathsf{s}^2 - \theta^{-1}\rangle}_{=0} + \theta\phi\,\underbrace{\langle \mathsf{s}\mathsf{w}, \mathsf{s}^2 - \theta^{-1}\rangle}_{(1-\theta^{-1})\langle \mathsf{w}, \mathsf{s}\rangle = 0} = 0\,, \end{aligned}$$

we find the following equivalent moment system:

$$\left.\begin{array}{lcl}\partial_t u+\partial_x\phi &=& q\\ \partial_t\phi+\frac{1}{\theta}\partial_x u+\partial_x\psi &=& 0\\ \partial_t\psi+\frac{1}{\theta}\partial_x\phi &=& -\epsilon^{-1}\omega\psi\end{array}\right\}\quad\Leftrightarrow\quad\left\{\begin{array}{rcl}\partial_t u+\partial_x\phi &=& q\\ \partial_t\phi+\theta^{-1}\partial_x u &=& \epsilon\tau\partial_x\big(\partial_t\psi+(1-\frac{1}{\theta})\partial_x\phi\big)\\ \psi &=& -\epsilon\tau\big(\partial_t\psi+(1-\frac{1}{\theta})\partial_x\phi\big)\end{array}\right.$$

The system to the right arises from the left one by solving the third equation for ψ and substituting the obtained expression into the second equation. A comparison of this system with the equations (2.13), (2.14) leads to the following observations:

- Since the second moment ψ is not conserved by the equilibrium, the right hand side of the evolution equation for ψ (left below) does not vanish. Solving for ψ, the evolution equation for ϕ can be written with a "source term" (right center) in contrast to (2.14). However the right hand side appears with the scaling factor ϵ, such that it becomes of order $\mathrm{O}(\epsilon)$ for well behaved bounded solutions. Therefore the mass moment u and the first moment ϕ of the D1P3 model should converge for $\epsilon\downarrow 0$ to solutions of the wave equations $\partial_t^2 v-\theta^{-1}\partial_x^2 v=\partial_t q$ and $\partial_t^2 v-\theta^{-1}\partial_x^2 v=-\partial_x q$ respectively.
- There is another point to be learned from this example. In 1.1.2 we have already seen how the D1P3 lattice-Boltzmann equation can be connected to a scalar conservation law (using the hyperbolic scaling as well). Amazingly, neither the algorithmic parameter τ (collision time) nor θ (weight parameter) interferes in the limit $\epsilon\downarrow 0$, i.e. in the target equation. In contrast, θ is lent a physical meaning now: The second equation differs from (2.14) by the factor θ^{-1} in front of $\partial_x u$. This shows, that the weight parameter θ determines the propagation speed of the waves. In chapter 3, we will see in the context of the parabolic scaling that the diffusivity is tuned by τ.

Parabolic scaling and viscous balance laws

Let us now consider the lattice-Boltzmann equation in *parabolic scaling*

$$\partial_t\mathsf{f}+\epsilon^{-1}\mathsf{s}\partial_x\mathsf{f}=-\epsilon^{-2}\omega\big[\mathsf{f}-E\mathsf{f}\big]+q\mathsf{w}\,. \tag{2.15}$$

To recognize the impact of the scaling with regard to the target equation, we repeat the transformation into the equivalent moment system. For this we define the mass moment u and the first moment ϕ analogously to the first paragraph:

$$u:=\langle\mathsf{f},1\rangle=\mathsf{f}_1+\mathsf{f}_2,\qquad \phi:=\langle\mathsf{f},\epsilon^{-1}\mathsf{s}\rangle=\tfrac{1}{\epsilon}\langle\mathsf{f},\mathsf{s}\rangle=\tfrac{1}{\epsilon}(\mathsf{f}_2-\mathsf{f}_1)\,.$$

In contrast to (2.5), the first moment is rescaled[6] by the factor ϵ^{-1}, as the quantity $\langle\mathsf{f},\mathsf{s}\rangle$ is deemed to be of magnitude $\mathrm{O}(\epsilon)$. The equilibrium operator E is chosen as in (2.6) with the difference that the scaling factor ϵ occurs in front of the f-term

$$E\mathsf{f}:=\langle\mathsf{f},1\rangle\mathsf{w}+\epsilon f\big(\langle\mathsf{f},1\rangle\big)\mathsf{sw}=u\mathsf{w}+\epsilon f(u)\mathsf{sw}\,.$$

Evidently, this equilibrium operator conserves also the mass moment. So u is expected to play once more the favored role. Indeed, we will see that u approximates the solution of the *viscous* balance law this time:

$$\partial_t v+\partial_x\big[f(v)\big]-\nu\partial_x^2 v=q\,. \tag{2.16}$$

[6]If we associate the populations $\mathsf{f}_1,\mathsf{f}_2$ with particles moving to the left or right respectively, then the hyperbolic and parabolic scaling distinguish each other by the following circumstance: In the case of the hyperbolic scaling the speed of the particles is constant 1 independently of ϵ. In opposition, the parabolic scaling requires, that the speed of the particles grows as ϵ tends to zero (inverse proportionality). If we assume that the first moment should be computed with respect to the (non-normalized) velocity (this seems physically more reasonable), then it is consequent to define the first moment as we have done it. However, a real justification of this definition is only obtained by a reasonable outcome.

We derive the *equivalent moment system* (evolution system for u and ϕ) by the same procedure as above. Applying $\langle\cdot,1\rangle$ to (2.15) results in the evolution equation for u:

$$\begin{aligned} \partial_t\langle \mathsf{f},1\rangle + \epsilon^{-1}\partial_x\langle \mathsf{sf},1\rangle &= \epsilon^{-2}\omega\Big[\langle E\mathsf{f},1\rangle - \langle \mathsf{f},1\rangle\Big] + g\langle \mathsf{w},1\rangle \\ \Leftrightarrow \qquad \partial_t u + \partial_x\phi &= \epsilon^{-2}\omega[u-u] + q = q\,. \end{aligned}$$

Analogically, the application of $\frac{1}{\epsilon}\langle\cdot,\mathsf{s}\rangle$ to (2.15) leads to the evolution equation for ϕ:

$$\begin{aligned} \epsilon^{-1}\partial_t\langle \mathsf{f},\mathsf{s}\rangle + \epsilon^{-2}\partial_x\langle \mathsf{sf},\mathsf{s}\rangle &= \epsilon^{-3}\omega\Big[\langle E\mathsf{f},\mathsf{s}\rangle - \langle \mathsf{f},\mathsf{s}\rangle\Big] + q\langle \mathsf{w},\mathsf{s}\rangle \\ \partial_t\phi + \epsilon^{-2}\partial_x\langle \mathsf{f},\mathsf{s}^2\rangle &= \epsilon^{-3}\omega\Big[\epsilon f(u) - \epsilon\phi\Big] + 0 \;=\; \epsilon^{-2}\omega\Big[f(u)-\phi\Big]\,. \end{aligned}$$

So we arrive at the equivalent moment system

$$\partial_t u + \partial_x\phi = q\,, \tag{2.17}$$
$$\partial_t\phi + \epsilon^{-2}\partial_x u = -\epsilon^{-2}\omega\big[\phi - f(u)\big]\,. \tag{2.18}$$

By solving the above equation for ϕ

$$\phi = f(u) - \tau\partial_x u - \epsilon^2\tau\partial_t\phi \tag{2.19}$$

and setting $\tau := \omega^{-1}$, (2.17) and (2.18) can be combined to

$$\partial_t u + f'(u)\partial_x u - \tau\partial_x^2 u = q + \epsilon^2\tau\partial_t\partial_x\phi\,. \tag{2.20}$$

Identifying τ with the macroscopic parameter ν (viscosity/diffusivity) we see, that the mass moment u satisfies a viscous balance law (2.16) apart from an additional source term. This term is however proportional to ϵ^2 and should produce only a relatively small perturbation of (2.16). In analogy to equation (2.11) we obtain the following closed equation in u

$$\epsilon^2\tau\partial_t^2 u + \partial_t u + f'(u)\partial_x u - \tau\partial_x^2 u = q + \epsilon^2\tau\partial_t q\,. \tag{2.21}$$

	hyperbolic scaling	parabolic scaling
moments	$u=\langle \mathsf{f},1\rangle,\quad \phi=\langle \mathsf{f},\mathsf{s}\rangle$	$u=\langle \mathsf{f},1\rangle,\quad \phi=\frac{1}{\epsilon}\langle \mathsf{f},\mathsf{s}\rangle$
equilibrium	$\mathcal{E}(u)=u\mathsf{w}+f(u)\mathsf{sw}$	$\mathcal{E}(u)=u\mathsf{w}+\epsilon f(u)\mathsf{sw}$
LB equation	$\partial_t\mathsf{f}+\mathsf{s}\partial_x\mathsf{f}=-\epsilon^{-1}\omega[\mathsf{f}-E\mathsf{f}]$	$\partial_t\mathsf{f}+\epsilon^{-1}\mathsf{s}\partial_x\mathsf{f}=-\epsilon^{-2}\omega[\mathsf{f}-E\mathsf{f}]$
target equation	$\partial_t v+\partial_x\big[f(v)\big]=0$	$\partial_t v+\partial_x\big[f(v)\big]-\nu\partial_x^2 v=0$

Table 2.1: Synoptic summary comparing both of the scalings. The mass moment conserving equilibrium gives rise to a scalar target equation. The names of the scalings are borrowed from the classification of the target equations, that come out finally (particularly in this example). In more complicated examples, however, the type of target equation/system might be of hybrid nature. Nevertheless, the parabolic scaling generally produces additional dissipative terms. Therefore it is often referred to as the *diffusive* or *viscous* scaling.

Equivalence of alternative formulations

So far, we have considered the transformation of the evolution equations[7], without losing a word about accompanying initial and boundary conditions. However, evolution equations alone are mathematically not reasonable and must be complemented by further conditions to obtain well-posed problems with unique solutions. Therefore we now give a complete statement of the associated IBVPs[8].
In particular we deal with the question whether the solution of the lattice-Boltzmann system can be reconstructed from the solution of the scalar equation (2.21). We establish the equivalence of the subsequent IBVPs in the sense, that the solution of one IBVP induces solutions to the other ones. For each IBVP we present two different boundary conditions, that are assumed to be homogeneous for simplicity. The IBVPs are considered over a time-space domain $\mathcal{X}_T := [0,T]\times[0,L]$ with $0 < T, L < \infty$.

IBVP for the *lattice-Boltzmann system*:

$$\left.\begin{array}{ll} \text{EQ}: & \begin{array}{l} \partial_t \mathsf{f}_1 - \frac{1}{\epsilon}\partial_x \mathsf{f}_1 = -\frac{1}{\epsilon^2}\omega\big[\mathsf{f}_1 - (E\mathsf{f})_1\big] + \frac{1}{2}q \\ \partial_t \mathsf{f}_2 + \frac{1}{\epsilon}\partial_x \mathsf{f}_2 = -\frac{1}{\epsilon^2}\omega\big[\mathsf{f}_2 - (E\mathsf{f})_2\big] + \frac{1}{2}q \end{array} \\ \text{IC}: & \quad \mathsf{f}_1(0,\cdot) = \mathsf{g}_1 \quad \wedge \quad \mathsf{f}_2(0,\cdot) = \mathsf{g}_2 \\ \text{BC}: & \begin{array}{ll} \text{i)} & \mathsf{f}_2(t,0) = -\mathsf{f}_1(t,0) \ \wedge \ \mathsf{f}_1(t,L) = -\mathsf{f}_2(t,L) \\ \text{ii)} & \mathsf{f}_2(t,0) = \mathsf{f}_1(t,0) \ \wedge \ \mathsf{f}_1(t,L) = \mathsf{f}_2(t,L) \end{array} \end{array}\right\} \tag{2.22}$$

IBVP for the *equivalent moment system*:

$$\left.\begin{array}{ll} \text{EQ}: & \begin{array}{rcl} \partial_t u + \phantom{\epsilon^{-2}}\partial_x \phi & = & q \\ \partial_t \phi + \epsilon^{-2}\partial_x u & = & -\epsilon^{-2}\omega\big[\phi - f(u)\big] \end{array} \\ \text{IC}: & \quad u(0,\cdot) = g \quad \wedge \quad \phi(0,\cdot) = p \\ \text{BC}: & \begin{array}{ll} \text{i)} & u(t,0) = 0 \ \wedge \ u(t,L) = 0 \\ \text{ii)} & \phi(t,0) = 0 \ \wedge \ \phi(t,L) = 0 \end{array} \end{array}\right\} \tag{2.23}$$

IBVP for the *scalar equation*:

$$\left.\begin{array}{lll} \text{EQ}: & & \epsilon^2\tau\partial_t^2 u + \partial_t u + f'(u)\partial_x u - \tau\partial_x^2 u = q + \epsilon^2\tau\partial_t q\,. \\ \text{IC}: & & u(0,\cdot) = g \quad \wedge \quad \partial_t u(0,\cdot) = h \\ \text{BC}: & \text{i)} & u(t,0) = 0 \ \wedge \ u(t,L) = 0 \\ & \text{ii)} & f\big(u(t,0)\big) - \tau\partial_x u(t,0) = 0 \ \wedge \ f\big(u(t,L)\big) - \tau\partial_x u\big(t,L\big) = 0 \end{array}\right\} \tag{2.24}$$

It is clear that (2.22) and (2.23) must be equivalent, since the variables $(\mathsf{f}_1, \mathsf{f}_2)$ and (u, ϕ) are related to each other by an invertible linear transformation. In

[7] Reduction of the D1P2 lattice-Boltzmann system into a scalar equation via the equivalent moment system.

[8] Shortcut for initial boundary value problem. Further abbreviations: EQ = (evolution) equation, IC = initial condition and BC = boundary condition.

particular, the bounce-back-type boundary condition with sign flipping yields at the left boundary

$$\mathsf{f}_2(t,0) = -\mathsf{f}_1(t,0) \quad \Leftrightarrow \quad \mathsf{f}_1(t,0) + \mathsf{f}_1(t,0) = 0 \quad \Leftrightarrow \quad u(t,0) = 0\,,$$

what corresponds to homogeneous Dirichlet boundary condition for the mass moment u. Similarly, the second boundary condition results in homogeneous Dirichlet boundary conditions for ϕ:

$$\mathsf{f}_2(t,0) = \mathsf{f}_1(t,0) \ \Leftrightarrow\ \mathsf{f}_1(t,0) - \mathsf{f}_2(t,0) = 0 \ \Leftrightarrow\ \epsilon\phi(t,0) = 0 \ \Leftrightarrow\ \phi(t,0) = 0\,.$$

The equivalence of (2.23) and (2.24) turns out to be more delicate, as one is obliged to differentiate (loss of information), when deriving the scalar equation from the equivalent moment system. So, the reverse way, from (2.24) to (2.23), must comprise an integration giving rise to an undetermined integration constant. Before we settle this problem in the next proposition, let us check how the initial and boundary conditions of (2.23) translate into those of (2.24).
From the first evolution equation of (2.23) we obtain the following relation between the initial data p and h:

$$h(x) = \partial_t u(0,x) = q(0,x) - \partial_x\phi(0,x) = q(0,x) - \partial_x p(x)\,.$$

As far as the BC ii) in (2.23) is concerned, $\phi(t,0) = 0$ implies $\partial_t\phi(t,0) = 0$ and this gives BC ii) in (2.24) using equation (2.19). In the general case of inhomogeneous boundary conditions, i.e. $\phi(t,0) = r(t)$, we get

$$f\big(u(t,0)\big) - \tau\partial_x u(t,0) = r(t) + \epsilon^2\tau\dot{r}(t)\,.$$

Proposition 2.1. *Suppose that $u \in \mathcal{C}^2(\mathcal{X}_T)$ is a solution of the IBVP (2.24) satisfying either boundary condition i) or ii). Assume moreover that the* compatibility condition

$$\int_0^L \Big[q(0,\xi) - h(\xi)\Big]\,\mathrm{d}\xi \ = \ 0 \tag{2.25}$$

is fulfilled in the case of ii). Then there exists a function $\phi \in \mathcal{C}^1(\mathcal{X}_T)$, such that (u,ϕ) becomes a solution of IBVP (2.23) satisfying boundary condition i) or ii) respectively. In the case of boundary condition i) ϕ is uniquely determined up to a purely time-dependent term.

Proof: Let us define ϕ in such a way that the second evolution equation in (2.23) is automatically satisfied. A slight reorganization of this equation yields

$$\partial_t\phi + \epsilon^{-2}\omega\phi = \epsilon^{-2}\omega f(u) - \epsilon^{-1}\partial_x u\,. \tag{2.26}$$

The solution of this ODE for ϕ with respect to t is given by the *variation of constants* formula

$$\begin{aligned}\phi(t,x) = &\left\{C + \int_0^x \Big[q(0,\xi) - h(\xi)\Big]\,\mathrm{d}\xi\right\}\mathrm{e}^{-\omega t/\epsilon^2} \\ &+ \epsilon^{-2}\omega\mathrm{e}^{-\omega t/\epsilon^2}\int_0^t \mathrm{e}^{\omega\theta/\epsilon^2}\Big[f\big(u(\theta,x)\big) - \tau\partial_x u(\theta,x)\Big]\,\mathrm{d}\theta\end{aligned} \tag{2.27}$$

where the initial value $p := \phi(0,\cdot)$ is set to

$$p(x) = C + \int_0^x \Big[q(0,\xi) - \overbrace{\partial_t u(0,\xi)}^{=h(\xi)}\Big]\,\mathrm{d}\xi \tag{2.28}$$

with a constant C to be specified later. Now we have to check that the first equation of (2.23) holds true with ϕ defined as above. By definition (2.27) we see that $\phi \in \mathcal{C}^1(\mathcal{X}_T)$ if $u \in \mathcal{C}^2(\mathcal{X}_T)$. Furthermore, it can be read off from the formula that the mixed derivative $\partial_t\partial_x\phi$ exists and is continuous. Therefore we are allowed to differentiate (2.26) with respect to x:

$$\epsilon^2\tau\partial_t(\partial_x\phi) + \partial_x\phi = f'(u)\partial_x u - \tau\partial_x^2 u\,.$$

We have multiplied the differentiated equation by $\epsilon^2\tau$ in order to compare it more easily with the slightly rearranged evolution equation of (2.24):

$$\epsilon^2\tau\partial_t(\underbrace{q - \partial_t u}) + \underbrace{q - \partial_t u} = f'(u)\partial_x u - \tau\partial_x^2 u\,.$$

Obviously, the quantity $q - \partial_t u$ satisfies the same ODE with respect to t as $\partial_x\phi$. Moreover we infer from (2.28) that $\partial_x\phi(0,x) = q(0,x) - h(x) = q(0,x) - \partial_t u(0,x)$. Due to the agreement of the initial conditions we can conclude from the *uniqueness* and existence theorem of *Picard-Lindelöf*[9] that

$$\partial_x\phi(t,x) = q(t,x) - \partial_t u(t,x) \qquad \text{for all } t \geq 0\,.$$

But this equation is equivalent to the first evolution equation of (2.23).
Let us now check the boundary conditions: In the case of BC i), ϕ is not subject to boundary conditions. Thus the integration constant C is freely selectable. Being multiplied by the factor $\mathrm{e}^{-\omega t/\epsilon^2}$, it is responsible for the time dependent term.
The case of BC ii) – corresponding to homogeneous *natural* boundary conditions for u – results in homogeneous Dirichlet conditions for ϕ, if $C = 0$ and if the compatibility condition (2.25) is fulfilled:

$$\phi(t,0) = \mathrm{e}^{-\omega t/\epsilon^2}\underbrace{\int_0^0 \Big[q(0,\xi) - h(\xi)\Big]\,\mathrm{d}\xi}_{=0\text{ (trivially)}} + \epsilon^{-2}\omega\int_0^t \mathrm{e}^{\omega(\theta-t)/\epsilon^2}\Big[\underbrace{f\big(u(t,0)\big) - \tau\partial_x u(t,0)}_{=0\text{ due to BC ii)}}\Big]\,\mathrm{d}\theta = 0,$$

$$\phi(t,L) = \mathrm{e}^{-\omega t/\epsilon^2}\underbrace{\int_0^L \Big[q(0,\xi) - h(\xi)\Big]\,\mathrm{d}\xi}_{=0\text{ by (2.25)}} + \epsilon^{-2}\omega\int_0^t \mathrm{e}^{\omega(\theta-t)/\epsilon^2}\Big[\underbrace{f\big(u(t,L)\big) - \tau\partial_x u(t,L)}_{=0\text{ due to BC ii)}}\Big]\,\mathrm{d}\theta = 0.$$

■

[9] The important theorem of Picard-Lindelöf states *grosso modo*, that the initial value problem $z(0) = z_0,\ \dot{z}(t) = F\big(t, z(t)\big)$ has a *unique* solution in a neighborhood of $t = 0$, if the function F is continuous w.r.t. its first and Lipschitz-continuous w.r.t. its second argument. Furthermore the theorem gives an lower estimate for the interval of existence in dependence of the Lipschitz constant. But this is here of no importance, since the equation is linear and the explicit formula (2.27) provides the solution for all $t \in [0,T]$. Therefore we only need the uniqueness statement for the above reasoning.

Remark 2.1. (Compatibility condition) Compatibility conditions for the prescribed data (sources, initial and boundary values) often arise to avoid possible conflicts with the evolution equation or other constraints.
For IBVP (2.24) the source q and the initial values for u and $\partial_t u$ are prescribed. From its solution u and the prescribed data we have constructed a function ϕ subject to homogeneous Dirichlet boundary conditions, such that (u, ϕ) solves IBVP (2.23). Now, the first evolution equation of (2.23) connects the quantities $\partial_x \phi, \partial_t u$ and q. Integrating this equation over the full space interval $[0, L]$ yields

$$\int_0^L \partial_\xi \phi(t,\xi)\,\mathrm{d}\xi = \phi(t,L) - \phi(t,0) = \int_0^L [q(t,\xi) - \partial_t u(t,\xi)]\mathrm{d}\xi\ .$$

For $t = 0$ we obtain that the difference of the initial boundary values of ϕ must be equal to $\int_0^L [q(0,\xi) - h(\xi)]\mathrm{d}\xi$. So we retrieve compatibility condition (2.25) in the special case of $\phi(t,0) = \phi(t,L) = 0$.
What is the physical meaning of this compatibility condition? To answer this, we take a glance on the target equation (2.16), that permits a clear physical interpretation in contrast to (2.23) and (2.24) if we identify u – for instance – with a mass density. By introducing the *physical flux* $\varphi := f(v) - \nu\partial_x v$, we can derive in the same manner as above the following equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^L v(t,\xi)\,\mathrm{d}\xi \equiv \int_0^L \partial_t v(t,\xi)\,\mathrm{d}\xi \ = \ \varphi(t,L) - \varphi(t,0) \ + \int_0^L q(t,\xi)\,\mathrm{d}\xi\ .$$

The left hand side is the time derivative of the total mass, i.e. the instantaneous change in total mass. Intuitively, it must be equal to what is flowing out or coming in at the boundary points $0, L$ (represented by $\varphi(t,L) - \varphi(t,0)$) plus the mass production rate represented by the space integral over the source function q. Thus the condition expresses conservation of mass.
Since u and ϕ approximate v and φ respectively, the existence of a similar compatibility condition involving u, ϕ is not surprising. Nevertheless, it is remarkable to hit on the same compatibility condition although φ is differently defined from ϕ via the ODE (2.19). Finally note, that (2.24) alone with BC ii) does not enforce the compatibility condition[10].

Remark 2.2. (Gauge freedom) The fact that ϕ can not be uniquely defined from (2.24) accompanied by BC i) reflects a property of the lattice-Boltzmann system (2.22) including BC i), that one might interpret as a sort of *gauge freedom* for the quite artificial population functions. If $(\mathsf{f}_1, \mathsf{f}_2)$ is a solution of the IBVP (2.22) subject to BC i), then $(\tilde{\mathsf{f}}_1, \tilde{\mathsf{f}}_2)$ with $\tilde{\mathsf{f}}_1 := \mathsf{f}_1 - c\mathrm{e}^{-\omega t/\epsilon^2}$ and $\tilde{\mathsf{f}}_2 := \mathsf{f}_2 + c\mathrm{e}^{-\omega t/\epsilon^2}$ solves (2.22) with BC i) as well but satisfies different initial conditions.
However, from the macroscopic perspective the crucial point is, that $(\mathsf{f}_1, \mathsf{f}_2)$ and $(\tilde{\mathsf{f}}_1, \tilde{\mathsf{f}}_2)$ have the same mass moment u. Hence they approximate the same v, which is of physical relevance as solution of the target equation. Moreover, the first moments $\phi = \frac{1}{\epsilon}(\mathsf{f}_2 - \mathsf{f}_1)$ and $\tilde{\phi} = \frac{1}{\epsilon}(\tilde{\mathsf{f}}_2 - \tilde{\mathsf{f}}_1) = \phi + 2c\epsilon^{-1}\mathrm{e}^{-\omega t/\epsilon^2}$ converge also for $\epsilon \to 0$ to the same function albeit being different.

[10] Instead we find: $\int_0^L [q(t,x) - \partial_t u(t,\xi)]\mathrm{d}\xi = -\epsilon^2\tau \int_0^L [\partial_t q(t,x) - \partial_t^2 u(t,\xi)]\mathrm{d}\xi = \mathrm{O}(\epsilon^2)$. This relation does not entail a condition for the prescribed data, since $\partial_t^2 u$ is not initialized.

This observation shows, that we cannot obtain from (2.24) a unique ϕ and thus unique population functions $\mathsf{f}_1, \mathsf{f}_2$, without having some extra information about their initial values.

Remark 2.3. (Periodic BC) The compatibility condition (2.25) emerges also in the case of *periodic* boundary conditions $\mathsf{f}(t,0) = \mathsf{f}(t,L)$, implying immediately periodic boundary conditions for both[11] moments: $u(t,0) = u(t,L)$ and $\phi(t,0) = \phi(t,L)$. In order to get classical periodic solutions, the derivatives occurring in the evolution equations are also required to be continuous periodic functions.
As far as the construction of a periodic ϕ is concerned (proof of proposition 2.1), we recognize in the explicit formula (2.27), that the only term[12] making trouble is the first integral with x as upper integration bound. In fact, the periodicity, i.e. the equality for $x = 0$ and $x = L$ enforces the integral to vanish over the full space domain $[0, L]$, leading thus to the compatibility condition. The constant C can be chosen arbitrarily.

2.2 A singularly perturbed initial value problem

In the previous section, two scalings for lattice-Boltzmann equations were presented using the D1P2 velocity model as example. Heuristically, it was shown how the scalings give rise to different target equations. In this section we want to inspect the limit more carefully, which is generated by the parabolic scaling. To start with, we take up the the scalar equation (2.21) to avoid the lattice-Boltzmann system. Further simplifications are obtained by considering the case of a trivial flux function f and absent sources q. Under these restrictions the IBVP (2.24) over the spatial domain $\mathcal{X} = [0, L]$ with homogeneous Dirichlet boundary conditions reads

$$\left.\begin{array}{ll} \text{EQ}: & \epsilon^2\tau\partial_t^2 u_\epsilon + \partial_t u_\epsilon - \tau\partial_x^2 u_\epsilon = 0, \qquad \tau > 0 \\ \text{BC}: & u_\epsilon(\cdot,0) = 0 = u_\epsilon(\cdot,L) \\ \text{IC}: & u_\epsilon(0,\cdot) = g \quad \wedge \quad \partial_t u_\epsilon(0,\cdot) = h \end{array}\right\}. \tag{2.29}$$

Here we want to focus on the behavior of the solution u_ϵ, if the parameter ϵ tends to zero. This is interesting, since the equation is losing its highest time derivative for $\epsilon = 0$. Thereby the equation undergoes an essential qualitative change[13]: the second initial condition prescribing $\partial_t u_\epsilon(0,\cdot)$ gets redundant (resulting in an overdetermination by too many initial conditions) due to an abrupt alteration of the equation type (hyperbolic $\rightarrow$ parabolic). Such a limit is called *singular*.

[11] Note the contrast to the bounce-back type boundary conditions affecting only one moment.

[12] Observe that the second integral is with respect to the time. That is why the assumed periodicity of u and $\partial_x u$ can directly pass over to ϕ.

[13] For example: second order differential equations are characterized by an oscillating behavior of their solutions (e.g. wave equation, harmonic oscillator), whereas solutions to first order equations exhibit rather a monotonic drop or growth (e.g. heat equation, radioactive decay). Furthermore, time evolution operators of parabolic equations are characterized by smoothing the regularity of the initial condition; this feature is not shared by hyperbolic equations. We will be confronted with this property in the course of this section.

In view of these malignancies it is by far not self-evident that the solution u_ϵ converges to the solution of the natural limit system,

$$\left.\begin{array}{ll} \mathrm{EQ}: & \partial_t u - \tau\partial_x^2 u = 0 \\ \mathrm{BC}: & u(\cdot,0) = 0 = u(\cdot,L) \\ \mathrm{IC}: & u(0,\cdot) = g \end{array}\right\} \tag{2.30}$$

where g is the same as in (2.29). This is all the more surprising, as the second initial condition does not need to be "consistent" with the limit, i.e. convergence happens even if $h \neq \partial_t u(0,\cdot)$. In the sequel we refer to (2.29) as the *perturbed problem*, while (2.30) denoted as the *limit problem*.

In this section we are going to prove the convergence of $u_\epsilon \to u$ employing *Fourier series*, that permit a quasi analytical computation of the solution for both IBVPs. Alternatively it is possible to approximate u_ϵ by a truncated expansion of the form $u_\epsilon = u^{(0)} + \epsilon u^{(1)} + ... + \epsilon^k u^{(k)} + \mathrm{O}(\epsilon^{k+1})$, finally using a stability result to infer convergence from consistency. Since this approach (which applies undoubtedly to a wider range of problems) is pursued in the subsequent chapters, we prefer here the classical Fourier technique for comparison[14].

2.2.1 The Fourier coefficient functions

Solution of the limit problem

IBVPs for *linear* PDEs are frequently attacked by an ansatz that is called *separation of variables*. The idea consists in writing the solution as a product, where one factor depends only on time while the other is purely space-dependent. This approach decomposes the original problem into a *boundary eigenvalue-problem* (BEVP) for the spatial function and into an *initial value problem* for the time-dependent part. Of course, not every solution of the original IBVP is so simply structured that it factorizes with respect to time and space. But thanks to the *linearity* of the problem, simple multiplicative solutions can be linearly combined, to synthesize more complicated solutions. In fact, under appropriate conditions, any solution can be expanded by means of simple solutions. Thus we arrive directly at what is known under the name of *generalized Fourier series*.

The solution of the IBVP (2.30) leads to the BEVP of the one-dimensional Laplacian (second derivative) with *homogeneous Dirichlet boundary conditions*. The eigenfunctions s_n to this problem are given by the sine

$$s_n(x) := \sqrt{\tfrac{2}{L}}\sin(n\pi x/L) \qquad \text{for } \ n \in \mathbb{N} \tag{2.31}$$

having an integral number of bellies between the boundary points 0 and L. More precisely these functions satisfy the equation

$$-\tau\partial_x^2 s_n = \lambda_n s_n \qquad \text{with the eigenvalues } \ \lambda_n := \tau\left(\tfrac{n\pi}{L}\right)^2. \tag{2.32}$$

[14] Apparently, the Fourier approach can be done with weaker regularity requirements.

The eigenfunctions s_n are normalized and pairwise orthogonal. Furthermore the set of eigenfunctions is *complete* in the Hilbert space $\mathcal{L}^2(0, L)$.

Let us suppose that the initial condition g is in $\mathcal{L}^2(0, L)$. Then it can be represented by a Fourier series

$$g = \sum_{n \in \mathbb{N}} \alpha_n s_n \qquad \text{with } \alpha_n := \int_0^L g(x) s_n(x)\, \mathrm{d}x\,. \tag{2.33}$$

The coefficients α_n are denoted as the *initial Fourier coefficients*.

Theorem 2.1. *For $g \in \mathcal{L}^2(0, L)$ the IBVP (2.30) has a unique solution*

$$\begin{aligned} u &\in \mathcal{C}\big([0,\infty), \mathcal{L}^2(0,L)\big) \cap \mathcal{C}^{1,2}\big((0,\infty)\times[0,L]\big) \qquad \textit{given by} \\ u(t,x) &= \sum_{n \in \mathbb{N}} \sigma_n(t) s_n(x) \qquad \textit{with} \quad \sigma_n(t) := \alpha_n \mathrm{e}^{-\lambda_n t}. \end{aligned} \tag{2.34}$$

Moreover $u \in \mathcal{C}^\infty\big((0,\infty)\times[0,L]\big)$.

The proof of this theorem can be performed by means of Fourier series where each coefficient σ_n comes out as solution of the IVP

$$\left.\begin{array}{ll} \text{EQ}: & \dot{\sigma}_n + \lambda_n \sigma_n = 0 \\ \text{IC}: & \sigma_n(0) = \alpha_n \end{array}\right\}. \tag{2.35}$$

Fourier coefficient functions for the perturbed problem

Since the spatial differential operator occurs in the perturbed equation (2.29) as well as in the limit equation (2.30), we will work with its eigenfunctions (2.31) to solve (2.29) too.
Let us assume that the second initial condition in (2.29) admits also a Fourier expansion

$$h = \sum_{n \in \mathbb{N}} \beta_n s_n \qquad \text{with } \beta_n := \int_0^L h(x) s_n(x)\, \mathrm{d}x\,. \tag{2.36}$$

In analogy to (2.34) we start with the ansatz

$$u_\epsilon(t) = \sum_{n \in \mathbb{N}} \sigma_{\epsilon,n}(t) s_n(x). \tag{2.37}$$

Plugging this expansion into (2.29), a formal computation yields the following IVP for each coefficient:

$$\left.\begin{array}{ll} \text{EQ}: & \epsilon^2 \tau \ddot{\sigma}_{\epsilon,n} + \dot{\sigma}_{\epsilon,n} + \lambda_n \sigma_{\epsilon,n} = 0 \\ \text{IC}: & \sigma_{\epsilon,n}(0) = \alpha_n \quad \wedge \quad \dot{\sigma}_{\epsilon,n}(0) = \beta_n \end{array}\right\} \tag{2.38}$$

In order to set up the *fundamental system* we need the zeros of the associated *characteristic polynomial*

$$z \mapsto \epsilon^2 \tau \Big(z^2 + \epsilon^{-2} \omega z + \epsilon^{-2} \omega \lambda_n \Big). \tag{2.39}$$

These are given by

$$\psi_{\epsilon,n} \;:=\; -\tfrac{1}{2}\epsilon^{-2}\omega + \sqrt{\tfrac{1}{4}\epsilon^{-4}\omega^2 - \epsilon^{-2}\omega\lambda_n} \;=\; -\tfrac{1}{2}\omega\epsilon^{-2}\left[1 - \sqrt{1 - 4\epsilon^2\tau\lambda_n}\right], \tag{2.40}$$

$$\chi_{\epsilon,n} \;:=\; -\tfrac{1}{2}\epsilon^{-2}\omega - \sqrt{\tfrac{1}{4}\epsilon^{-4}\omega^2 - \epsilon^{-2}\omega\lambda_n} \;=\; -\tfrac{1}{2}\omega\epsilon^{-2}\left[1 + \sqrt{1 - 4\epsilon^2\tau\lambda_n}\right]. \tag{2.41}$$

It will be convenient to abbreviate the square root:

$$W_{\epsilon,n} := \sqrt{1 - 4\epsilon^2\tau\lambda_n} \quad \Rightarrow \quad \begin{cases} \psi_{\epsilon,n} = -\tfrac{1}{2}\omega\epsilon^{-2}[1 - W_{\epsilon,n}] \\ \chi_{\epsilon,n} = -\tfrac{1}{2}\omega\epsilon^{-2}[1 + W_{\epsilon,n}] \end{cases}. \tag{2.42}$$

In general, two cases have to be distinguished concerning the fundamental system:

$$\begin{array}{lll} 1) \quad \text{two zeros} & \psi_{\epsilon,n} \neq \chi_{\epsilon,n} \Leftrightarrow W_{\epsilon,n} \neq 0: & t \mapsto \mathrm{e}^{\psi_{\epsilon,n} t}, \;\; t \mapsto \mathrm{e}^{\chi_{\epsilon,n} t} \\ 2) \quad \text{double zero} & \psi_{\epsilon,n} = \chi_{\epsilon,n} \Leftrightarrow W_{\epsilon,n} = 0: & t \mapsto \mathrm{e}^{\psi_{\epsilon,n} t}, \;\; t \mapsto t\mathrm{e}^{\psi_{\epsilon,n} t} \end{array}$$

The first case is the standard situation. If both zeros coincide, a *degenerated* case is obtained.

In order to avoid annoying repetitions of case distinctions, we henceforth adopt the subsequent agreement till explicit revocation.

Assumption 2.1. The real parameter $\epsilon \neq 0$ is thought as a null-sequence (converging towards zero) subject to the sole condition[15] that it has no common element with the null-sequence $n \mapsto \frac{L}{2\tau\pi n}$, i.e.:

$$\forall\, n \in \mathbb{N}: \quad \epsilon \neq \tfrac{L}{2\sqrt{\tau}\pi n} \quad \Leftrightarrow \quad 1 - 4\epsilon^2\tau\lambda_n \neq 0.$$

This condition implies $W_{\epsilon,n} \neq 0$, such that we are always in the regular case with two different zeros of the characteristic polynomial. Hence, the Fourier coefficient function $\sigma_{\epsilon,n}$ is a linear combination of $\mathrm{e}^{\psi_{\epsilon,n} t}$ and $\mathrm{e}^{\chi_{\epsilon,n} t}$ respecting the initial conditions in (2.38). A short calculation yields the formula

$$\sigma_{\epsilon,n}(t) := \frac{\beta_n - \alpha_n\chi_{\epsilon,n}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\mathrm{e}^{\psi_{\epsilon,n} t} \;+\; \frac{\alpha_n\psi_{\epsilon,n} - \beta_n}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\mathrm{e}^{\chi_{\epsilon,n} t}. \tag{2.43}$$

Although we have derived this expression for $\sigma_{\epsilon,n}$ starting with the ansatz (2.37), a rigorous treatment requires that we take this equation as a definition[16].
Sorting for terms containing either α_n or β_n we get

$$\sigma_{\epsilon,n}(t) = \alpha_n\frac{\psi_{\epsilon,n}\mathrm{e}^{\chi_{\epsilon,n} t} - \chi_{\epsilon,n}\mathrm{e}^{\psi_{\epsilon,n} t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}} \;+\; \beta_n\frac{\mathrm{e}^{\psi_{\epsilon,n} t} - \mathrm{e}^{\chi_{\epsilon,n} t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}. \tag{2.44}$$

[15] Notice that ϵ occurs always in squared form, that is why it is redundant to require $\epsilon > 0$.

[16] We do not yet know whether the resulting Fourier series will be convergent in $\mathcal{L}^2(0, L)$, not to mention any regularity properties which are necessary to justify the formal computation. Therefore we are obliged to go the other way round. So it has to be shown that the coefficients defined by the IVP (2.38) give rise to a converging Fourier series with a sufficiently smooth limit function.

Furthermore using the relation $\psi_{\epsilon,n} - \chi_{\epsilon,n} = \omega W_{\epsilon,n}/\epsilon^2$, this is transformed into

$$\begin{aligned}
\sigma_{\epsilon,n}(t) &= \tfrac{1}{2}\alpha_n \left[(1 - W_{\epsilon,n}^{-1})\mathrm{e}^{\chi_{\epsilon,n}t} + (1 + W_{\epsilon,n}^{-1})\mathrm{e}^{\psi_{\epsilon,n}t}\right] + \tfrac{1}{2}\beta_n \epsilon^2 \tau W_{\epsilon,n}^{-1}\left[\mathrm{e}^{\psi_{\epsilon,n}t} - \mathrm{e}^{\chi_{\epsilon,n}t}\right] \\
&= \tfrac{1}{2}\alpha_n \left[\left(\mathrm{e}^{\psi_{\epsilon,n}t} + \mathrm{e}^{\chi_{\epsilon,n}t}\right) + W_{\epsilon,n}^{-1}\left(\mathrm{e}^{\psi_{\epsilon,n}t} - \mathrm{e}^{\chi_{\epsilon,n}t}\right)\right] + \tfrac{1}{2}\beta_n \epsilon^2 \tau W_{\epsilon,n}^{-1}\left[\mathrm{e}^{\psi_{\epsilon,n}t} - \mathrm{e}^{\chi_{\epsilon,n}t}\right] \\
&= \alpha_n \rho_{n,\epsilon} + \beta_n \varrho_{n,\epsilon} \, , \qquad (2.45)
\end{aligned}$$

where we have set

$$\begin{aligned}
\rho_{\epsilon,n}(t) &:= \tfrac{1}{2}\left[\mathrm{e}^{\psi_{\epsilon,n}t} + \mathrm{e}^{\chi_{\epsilon,n}t}\right] + \tfrac{1}{2} W_{\epsilon,n}^{-1}\left[\mathrm{e}^{\psi_{\epsilon,n}t} - \mathrm{e}^{\chi_{\epsilon,n}t}\right] , \qquad (2.46) \\
\varrho_{\epsilon,n}(t) &:= \tfrac{1}{2}\epsilon^2 \tau W_{\epsilon,n}^{-1}\left[\mathrm{e}^{\psi_{\epsilon,n}t} - \mathrm{e}^{\chi_{\epsilon,n}t}\right] . \qquad (2.47)
\end{aligned}$$

In order to prove boundedness of these coefficient functions we prefer another representation. For this, recall the definition of $\cosh(z) = \frac{1}{2}(\mathrm{e}^z + \mathrm{e}^{-z})$ and $\sinh(z) = \frac{1}{2}(\mathrm{e}^z - \mathrm{e}^{-z})$. Using the representation of $\psi_{\epsilon,n}, \chi_{\epsilon,n}$ in (2.42) we get:

$$\begin{aligned}
\mathrm{e}^{\psi_{\epsilon,n}t} + \mathrm{e}^{\chi_{\epsilon,n}t} &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2 + \frac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2} + \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2 - \frac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2} \\
&= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\left[\mathrm{e}^{\frac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2} + \mathrm{e}^{-\frac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2}\right] \\
&= 2\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\cosh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right) .
\end{aligned}$$

Similarly we obtain:

$$\begin{aligned}
\mathrm{e}^{\psi_{\epsilon,n}t} - \mathrm{e}^{\chi_{\epsilon,n}t} &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\left[\mathrm{e}^{\frac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2} - \mathrm{e}^{-\frac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2}\right] \\
&= 2\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\sinh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right) .
\end{aligned}$$

This finally gives the following representation for the coefficient functions defined in (2.46) and (2.47)

$$\begin{aligned}
\rho_{\epsilon,n}(t) &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\cosh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right) + \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} W_{\epsilon,n}^{-1}\sinh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right) , \qquad (2.48) \\
\varrho_{\epsilon,n}(t) &= \epsilon^2 \tau \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} W_{\epsilon,n}^{-1}\sinh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right) . \qquad (2.49)
\end{aligned}$$

Boundedness of the Fourier coefficient functions

The boundedness of the Fourier coefficient functions may be related to stability estimates, that are necessary for alternative approaches (e.g. regular expansions). The subsequent lemma represents a preparative and indispensable intermediate step for proving further results.

Lemma 2.1. *The coefficient functions $\rho_{\epsilon,n}$ defined in (2.46) satisfy the uniform estimate*

$$|\rho_{\epsilon,n}(t)| \le 2$$

for all $t \ge 0$, $n \in \mathbb{N}$ and all admissible ϵ.

Proof: From (2.48) we obtain:

$$|\rho_{\epsilon,n}(t)| \leq \underbrace{\left|\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\cosh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right)\right|}_{=:A_{\epsilon,n}(t)\geq 0} + \underbrace{\left|\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} W_{\epsilon,n}^{-1}\sinh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right)\right|}_{=:B_{\epsilon,n}(t)\geq 0} .$$

We show in the sequel that $A_{\epsilon,n}(t), B_{\epsilon,n}(t) \leq 1$ independently of ϵ, n and t, such that $|\rho_{\epsilon,n}(t)| \leq A_{\epsilon,n}(t) + B_{\epsilon,n}(t) = 2$. There are two cases to be distinguished:

I) $1 - 4\epsilon^2\tau\lambda_n > 0 \Leftrightarrow W_{n,\epsilon} \in \mathbb{R}$ due to $\lambda_n, \tau, \epsilon > 0$ follows: $0 < W_{\epsilon,n} < 1$
II) $1 - 4\epsilon^2\tau\lambda_n < 0 \Leftrightarrow W_{n,\epsilon} \in i\mathbb{R}$ there exists $w_{\epsilon,n} \in \mathbb{R}^+$ with: $W_{\epsilon,n} = \mathrm{i}w_{\epsilon,n}$

Here 'i' denotes the *imaginary unit.*

Ad I) $0 < W_{\epsilon,n} < 1$

$$\begin{aligned}
A_{\epsilon,n}(t) &= \left|\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\cosh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right)\right| \\
&\leq \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\cosh\left(\tfrac{1}{2}\omega t/\epsilon^2\right) \\
&= \tfrac{1}{2}\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\left(\mathrm{e}^{\frac{1}{2}\omega t/\epsilon^2} + \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\right) \\
&= \tfrac{1}{2} + \tfrac{1}{2}\mathrm{e}^{-\omega t/\epsilon^2} \\
&\leq \tfrac{1}{2} + \tfrac{1}{2} = 1
\end{aligned}$$

For the first inequality we have employed that $W_{\epsilon,n}$ is positive and that cosh is monotonically increasing for positive arguments. The second inequality holds for $t \geq 0$ and $\omega > 0$.

In order to estimate $B_{\epsilon,n}(t)$ (note remark 2.4) we consider the function

$$f(a,x) := \mathrm{e}^{-a}\,\frac{\sinh(ax)}{x} \qquad \text{for} \quad 0 \leq x \leq 1 \;\wedge\; a \geq 0.$$

We assert that for a fixed $a > 0$ the function $(0,1) \ni x \mapsto f(a,x)$ is *monotonically increasing.* As the function is continuously differentiable, this is verified by checking the non-negativity of its derivative. Indeed:

$$\partial_x f(a,x) = \mathrm{e}^{-a}\,\frac{xa\cosh(ax) - \sinh(ax)}{x^2} > 0 \tag{2.50}$$

Due to its monotonicity the function $(0,1) \ni x \mapsto f(a,x)$ attains its extrema at the boundary i.e. for x tending to 0 or 1.

$$\begin{aligned}
\lim_{x\downarrow 0} f(a,x) &= \mathrm{e}^{-a}\lim_{x\downarrow 0}\tfrac{\sinh(ax)}{x} \overset{\text{L'Hospital}}{=} \mathrm{e}^{-a}\lim_{x\downarrow 0}\tfrac{a\cosh(ax)}{1} = \mathrm{e}^{-a}a \leq \mathrm{e}^{-1} < \tfrac{1}{2} \\
\lim_{x\uparrow 1} f(a,x) &= \mathrm{e}^{-a}\sinh(a) = \tfrac{1}{2}\mathrm{e}^{-a}\left(\mathrm{e}^{a} - \mathrm{e}^{-a}\right) = \tfrac{1}{2}(1 - \mathrm{e}^{-2a}) \leq \tfrac{1}{2}
\end{aligned}$$

Observe that the function $a \mapsto a\mathrm{e}^{-a}$ is maximal for $a = 1$ (see lemma 2.8). Since $a \geq 0$ is arbitrary (although fixed) we get $|f(a,x)| \leq \frac{1}{2} < 1$ for $0 < x < 1$. Identifying x with $W_{\epsilon,n}$ and setting $a = \frac{1}{2}\omega t/\epsilon^2$ we obtain

$$B_{\epsilon,n}(t) = \left|\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} W_{\epsilon,n}^{-1}\sinh\left(\tfrac{1}{2}\omega W_{\epsilon,n}t/\epsilon^2\right)\right| < 1 .$$

The correctness of equation (2.50) ensues from the fact that $y\cosh(y) > \sinh(y)$ for all $y > 0$. As the hyperbolic functions are real analytic, this can easily be seen by writing down the Taylor series:

$$\begin{aligned} y\cosh(y) &= y\big(1 + \tfrac{1}{2!}y^2 + \tfrac{1}{4!}y^4 + \tfrac{1}{6!}y^6 + \ldots\big) = y + \tfrac{1}{2!}y^3 + \tfrac{1}{4!}y^5 + \tfrac{1}{6!}y^7 + \ldots \\ \sinh(y) &= y + \tfrac{1}{3!}y^3 + \tfrac{1}{5!}y^5 + \tfrac{1}{7!}y^7 + \ldots . \end{aligned}$$

Ad II) $W_{\epsilon,n} = \mathrm{i}w_{\epsilon,n}, \quad w_{\epsilon,n} > 0$

In the following estimates we exploit the identities $\sinh(\mathrm{i}x) = \mathrm{i}\sin(x)$ and $\cosh(\mathrm{i}x) = \cos(x)$ for real x.

$$\begin{aligned} A_{\epsilon,n}(t) &= \left|\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot \cosh\left(\tfrac{1}{2}\omega W_{\epsilon,n} t/\epsilon^2\right)\right| \\ &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot \left|\cosh\left(\tfrac{1}{2}\omega i w_{\epsilon,n} t/\epsilon^2\right)\right| \\ &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot \left|\cos\left(\tfrac{1}{2}\omega w_{\epsilon,n} t/\epsilon^2\right)\right| \\ &\le \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot 1 \\ &\le 1 \end{aligned}$$

The first inequality results from the fact, that the cosine is bounded for real arguments in contrast to the cosine hyperbolicus. Therefore the exponential function, that was important for case I to damp the cosine hyperbolicus, does not play a crucial role this time.

$$\begin{aligned} B_{\epsilon,n}(t) &= \left|\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot W_{\epsilon,n}^{-1} \cdot \sinh\left(\tfrac{1}{2}\omega W_{\epsilon,n} t/\epsilon^2\right)\right| \\ &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot |\mathrm{i}w_{\epsilon,n}|^{-1} \cdot \left|\sinh\left(\tfrac{1}{2}\omega \mathrm{i} w_{\epsilon,n} t/\epsilon^2\right)\right| \\ &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot w_{\epsilon,n}^{-1} \cdot \left|\mathrm{i} \cdot \sin\left(\tfrac{1}{2}\omega w_{\epsilon,n} t/\epsilon^2\right)\right| \\ &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot \frac{\left|\sin\left(\frac{1}{2}\omega w_{\epsilon,n} t/\epsilon^2\right)\right|}{w_{\epsilon,n}} \\ &\le \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cdot \tfrac{1}{2}\omega t/\epsilon^2 \\ &\le 1 \end{aligned}$$

The first inequality emanates from the *mean value theorem*, which gives $\sin\left(\frac{1}{2}\omega w_{\epsilon,n} t/\epsilon^2\right) = \frac{1}{2}\omega t/\epsilon^2 \cdot \cos\left(\frac{1}{2}\omega \xi_{\epsilon,n} t/\epsilon^2\right) \cdot w_{\epsilon,n}$ for some $0 < \xi_{\epsilon,n} < w_{\epsilon,n}$. As the cosine function and the mapping $a \mapsto a\mathrm{e}^{-a}$ are bounded by 1 for non-negative arguments, we obtain the final estimate.

■

Remark 2.4. Under condition I) the mean value theorem is of no use to estimate $B_{\epsilon,n}$, since we get here a cosine hyperbolicus instead of a cosine. Notice, that the occurring exponential $\exp(-\frac{1}{2}\omega t/\epsilon^2)$ can only be utilized to "slay" either the cosine hyperbolicus or the factor $\frac{1}{2}\omega t/\epsilon^2$, which would come into play by the mean value theorem.

Corollary 2.1. *The coefficient functions $\varrho_{\epsilon,n}$ defined in (2.47) satisfy the uniform estimate*

$$|\varrho_{\epsilon,n}(t)| < \tau\epsilon^2$$

for all $t \geq 0$, $n \in \mathbb{N}$ and all admissible ϵ.

Proof: The representation of $\varrho_{\epsilon,n}$ in (2.49) corresponds to the second addend of (2.48) multiplied additionally by $\epsilon^2\tau$. Therefore $\varrho_{\epsilon,n}$ is estimated in the same way as $B_{\epsilon,n}$ in the proof of the previous lemma. ∎

Summarizing, we have obtained

$$|\sigma_{\epsilon,n}(t)| < 2|\alpha_n| + |\beta_n|\tau\epsilon^2 \ . \tag{2.51}$$

The uniform boundedness of the coefficient functions $\rho_{\epsilon,n}$ and $\varrho_{\epsilon,n}$ has two useful consequences.

Lemma 2.2. *The Fourier coefficient functions $\sigma_{\epsilon,n}(t)$ represent square-summable sequences with respect to n. Furthermore their $\ell^2(\mathbb{N})$-norm can be bounded independently*[17] *of $0 < \epsilon \leq 1$ and $t \geq 0$.*

Proof: Recall that the sequences $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ are square-summable as Fourier coefficients of the $\mathcal{L}^2$-functions g, h. Introducing $A^2 := \sum_{n\geq 1}|\alpha_n|^2$ and $B^2 := \sum_{n\geq 1}|\beta_n|^2$ and exploiting the estimates in lemma 2.1 and corollary 2.1 we obtain:

$$\begin{aligned}
\sum_{n\geq 1}|\sigma_{\epsilon,n}(t)|^2 &\overset{(2.45)}{=} \sum_{n\geq 1}\left|\alpha_n\rho_{\epsilon,n}(t) + \beta_n\varrho_{\epsilon,n}(t)\right|^2 \\
&\leq \sum_{n\geq 1}|\alpha_n|^2|\rho_{\epsilon,n}(t)|^2 + 2\sum_{n\geq 1}|\alpha_n||\beta_n||\rho_{\epsilon,n}(t)||\varrho_{\epsilon,n}(t)| + \sum_{n\geq 1}|\beta_n|^2|\varrho_{\epsilon,n}(t)|^2 \\
&\leq 4\sum_{n\geq 1}|\alpha_n|^2 + 4\tau\epsilon^2\sum_{n\geq 1}|\alpha_n||\beta_n| + \tau^2\epsilon^4\sum_{n\geq 1}|\beta_n|^2 \\
&\leq 4A^2 + 4\tau\epsilon^2 AB + \tau^2\epsilon^4 B^2 < 4A^2 + 4\tau AB + \tau^2 B^2 < \infty \quad \text{for } \epsilon \leq 1
\end{aligned}$$

Observe that we have employed the *Cauchy-Schwarz* inequality for infinite sums.to estimate the product sum $\sum_{n\geq 1}|\alpha_n||\beta_n|$. ∎

The sequences composed of the time-dependent Fourier coefficient functions permit not only a common bound for their ℓ^2-norms (independent of ϵ), but their sums converge also *uniformly.* The statement of the next lemma makes this more precise.

Lemma 2.3. *For every $\delta > 0$ there exists $N = N(\delta) \in \mathbb{N}$, such that*

$$\sum_{n>N}|\sigma_{\epsilon,n}(t)|^2 < \delta$$

independently of $t \geq 0$ and $0 < \epsilon \leq 1$.

[17] The condition $\epsilon \leq 1$ in the assumptions of this and the next lemma seems to be artificial and arbitrary, since other upper bounds are possible as well. Although the condition might appear superfluous, as we consider the limit $\epsilon \to 0$, we have added it here to be formally correct.

Proof: We remember again that $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ are square-summable. On account of the *Cauchy-Schwarz inequality*, the sequence $(\alpha_n\beta_n)_{n\in\mathbb{N}}$ is summable. Hence for every $\eta > 0$ there exist natural numbers $N_\alpha, N_\beta, N_{\alpha\beta} \in \mathbb{N}$ depending on η, such that

$$\sum_{n>N_\alpha} |\alpha_n|^2 < \eta, \qquad \sum_{n>N_\beta} |\beta_n|^2 < \eta, \qquad \sum_{n>N_{\alpha\beta}} |\alpha_n\beta_n| < \eta.$$

Let us set $N := \max\{N_\alpha, N_\beta, N_{\alpha\beta}\}$. The computation is now analogous to the estimate in the proof of the preceding lemma:

$$\begin{aligned}\sum_{n\geq N} |\sigma_{\epsilon,n}(t)|^2 &\leq 4\sum_{n\geq N} |\alpha_n|^2 + 4\tau\epsilon^2 \sum_{n\geq N} |\alpha_n||\beta_n| + \tau^2\epsilon^4 \sum_{n\geq N} |\beta_n|^2 \\ &\leq 4\eta + 4\tau\epsilon^2\eta + \tau^2\epsilon^4\eta \leq (4 + 4\tau + \tau^2)\eta \qquad \text{for } \epsilon \leq 1.\end{aligned}$$

If η is chosen to be $(4 + 4\tau + \tau^2)^{-1}\delta$ the assertion follows. ■

Remark 2.5. This lemma turns out to be the crucial tool in dealing with the infinite Fourier series for u_ϵ, that we have formally written down in (2.37). It permits us to show that the Fourier coefficients functions $\sigma_{\epsilon,n}$ bequeath properties – like continuity in time or convergence with respect to ϵ – to the whole series.

Estimate of the time derivative

In subsection 2.2.3 we need an estimate similar to (2.51) also for $\dot\sigma_{\epsilon,n}$, that shall be anticipated here.

Lemma 2.4. *The time derivative of the Fourier coefficient functions satisfies the estimate*

$$|\dot\sigma_{\epsilon,n}(t)| \leq |\alpha_n|\lambda_n + 2|\beta_n|$$

for all $t \geq 0$, $n \in \mathbb{N}$ and any admissible $\epsilon > 0$.

Proof: Computing the time derivative of the definition (2.43)

$$\begin{aligned}\dot\sigma_{\epsilon,n}(t) &= \alpha_n \frac{\mathrm{d}}{\mathrm{d}t}\frac{\psi_{\epsilon,n}\mathrm{e}^{\chi_{\epsilon,n}t} - \chi_{\epsilon,n}\mathrm{e}^{\psi_{\epsilon,n}t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}} + \beta_n \frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathrm{e}^{\psi_{\epsilon,n}t} - \mathrm{e}^{\chi_{\epsilon,n}t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}} \\ &= \alpha_n\psi_{\epsilon,n}\chi_{\epsilon,n}\underbrace{\frac{\mathrm{e}^{\chi_{\epsilon,n}t} - \mathrm{e}^{\psi_{\epsilon,n}t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}}_{=-\varrho_{\epsilon,n}(t)} + \beta_n\underbrace{\frac{\psi_{\epsilon,n}\mathrm{e}^{\psi_{\epsilon,n}t} - \chi_{\epsilon,n}\mathrm{e}^{\chi_{\epsilon,n}t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}}_{=:\tilde\rho_{\epsilon,n}(t)},\end{aligned}$$

we recognize in the first addend the coefficient function $\varrho_{\epsilon,n}$ (see (2.47)), while the second addend contains a factor $\tilde\rho_{\epsilon,n}$ being similar to the coefficient function $\rho_{\epsilon,n}$ (see (2.46)). With

$$\begin{aligned}\psi_{\epsilon,n}\chi_{\epsilon,n} &= \Big(-\tfrac{1}{2}\omega\epsilon^{-2}(1 - W_{\epsilon,n})\Big)\cdot\Big(-\tfrac{1}{2}\omega\epsilon^{-2}(1 + W_{\epsilon,n})\Big) = \tfrac{1}{4}\omega^2\epsilon^{-4}(1 - W^2) \\ &= \tfrac{1}{4}\omega^2\epsilon^{-4}\Big(1 - (1 - 4\epsilon^2\tau\lambda_n)\Big) = \omega\epsilon^{-2}\lambda_n\end{aligned}$$

we obtain

$$\dot{\sigma}_{\epsilon,n}(t) = \alpha_n \omega \epsilon^{-2} \lambda_n \varrho_{\epsilon,n}(t) + \beta_n \tilde{\rho}_{\epsilon,n}(t).$$

Applying corollary (2.1), we find the estimate

$$\begin{aligned} |\dot{\sigma}_{\epsilon,n}(t)| &\leq |\alpha_n| \omega \epsilon^{-2} \lambda_n |\varrho_{\epsilon,n}(t)| + |\beta_n||\tilde{\rho}_{\epsilon,n}(t)| \\ &\leq |\alpha_n| \lambda_n + |\beta_n||\tilde{\rho}_{\epsilon,n}(t)| \, . \end{aligned} \tag{2.52}$$

In order to estimate $|\tilde{\rho}_{\epsilon,n}(t)|$, a further computation using $\psi_{\epsilon,n} - \chi_{\epsilon,n} = \omega W_{\epsilon,n}/\epsilon^2$ yields:

$$\begin{aligned} \tilde{\rho}_{\epsilon,n}(t) &= \tfrac{1}{2}(1 - W_{\epsilon,n}^{-1}) \mathrm{e}^{\psi_{\epsilon,n} t} + \tfrac{1}{2}(1 + W_{\epsilon,n}^{-1}) \mathrm{e}^{\chi_{\epsilon,n} t} \\ &= \tfrac{1}{2}\big(\mathrm{e}^{\psi_{\epsilon,n} t} - \mathrm{e}^{\chi_{\epsilon,n}}\big) + \tfrac{1}{2} W_{\epsilon,n}^{-1} \big(\mathrm{e}^{\chi_{\epsilon,n} t} + \mathrm{e}^{\psi_{\epsilon,n}}\big) \\ &= \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \cosh(\tfrac{1}{2}\omega W_{\epsilon,n} t/\epsilon^2) - W_{\epsilon,n} \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \sinh(\tfrac{1}{2}\omega W_{\epsilon,n} t/\epsilon^2) \, . \end{aligned}$$

A comparison with (2.46) shows, that $\tilde{\rho}_{\epsilon,n}$ is the difference of those terms, which give as sum $\rho_{\epsilon,n}$. Since each of the two terms has been estimated separately in lemma 2.1, we conclude

$$|\tilde{\rho}_{\epsilon,n}(t)| \leq 2.$$

With (2.52) this gives the assertion. ■

Convergence of the Fourier coefficient functions for $\epsilon \searrow 0$

Let us first consider how the zeros in (2.40) and (2.41) behave, when ϵ tends towards zero. The characteristic polynomial is of the form

$$p(z) = \epsilon^2 z^2 + az + b \qquad \text{with } a, b \in \mathbb{R} \setminus \{0\}. \tag{2.53}$$

Obviously the quadratic polynomial degenerates to a linear function for vanishing ϵ. Its zeros are given by

$$z_\epsilon^+ := \frac{-\frac{1}{2}a + \sqrt{\frac{1}{4}a^2 - \epsilon^2 b}}{\epsilon^2} \quad \text{and} \quad z_\epsilon^- := \frac{-\frac{1}{2}a - \sqrt{\frac{1}{4}a^2 - \epsilon^2 b}}{\epsilon^2} \, . \tag{2.54}$$

Observe that both zeros are real for sufficiently small ϵ.

Lemma 2.5. *The zeros in 2.54 have the following asymptotic comportment:*

$$\begin{array}{llllll} i) & a > 0: & z_\epsilon^+ \xrightarrow{\epsilon \downarrow 0} -\frac{b}{a} & \textit{and} & z_\epsilon^- \xrightarrow{\epsilon \downarrow 0} -\infty \\ ii) & a < 0: & z_\epsilon^+ \xrightarrow{\epsilon \downarrow 0} \infty, & \textit{and} & z_\epsilon^- \xrightarrow{\epsilon \downarrow 0} -\frac{b}{a} \end{array}$$

Proof: ad i) As the square root is by definition positive (for sufficiently small ϵ) and as $a > 0$ by hypothesis, we see that the numerator of z_ϵ^- converges to $-a$, while the positive denominator ϵ^2 tends to 0. Hence z_ϵ^- goes to minus infinity.
In order to compute the limit of z_ϵ^+ we use *L'Hospital's rule* with

$$f(\epsilon) := -\tfrac{1}{2}a + \sqrt{\tfrac{1}{4}a^2 - \epsilon^2 b} \quad \text{and} \quad g(\epsilon) := \epsilon^2.$$

Since $f(0) = 0$ and $g(0) = 0$ we compute the first derivatives:

$$f'(\epsilon) = -\epsilon b \left(\tfrac{1}{4}a^2 - \epsilon^2 b\right)^{-\frac{1}{2}} \quad \text{and} \quad g'(\epsilon) = 2\epsilon$$

Once more we obtain $f'(0) = 0$ and $g'(0) = 0$, such that we have to pass over to the second derivatives:

$$f''(\epsilon) = -b \left(\tfrac{1}{4}a^2 - \epsilon^2 b\right)^{-\frac{1}{2}} - \epsilon^2 b(\tfrac{1}{4}a^2 - \epsilon^2 b)^{-\frac{3}{2}} \quad \text{and} \quad g''(\epsilon) = 2$$

Now we get: $\lim\limits_{\epsilon\to 0} z_\epsilon^+ = \lim\limits_{\epsilon\downarrow 0} \frac{f(\epsilon)}{g(\epsilon)} = \lim\limits_{\epsilon\downarrow 0} \frac{f'(\epsilon)}{g'(\epsilon)} = \lim\limits_{\epsilon\downarrow 0} \frac{f''(\epsilon)}{g''(\epsilon)} = \frac{f''(0)}{g''(0)} = \frac{-b(a^2/4)^{-\frac{1}{2}}}{2} = -\frac{b}{a}$.

ad ii) The second case comes out in complete analogy where the term $(a^2/4)^{\frac{1}{2}}$ must be evaluated with care, as it is $-a/2$ for $a < 0$. ∎

The finite limit $-\frac{b}{a}$ is just the zero of the affine-linear function, that is obtained by setting $\epsilon = 0$ in (2.53).

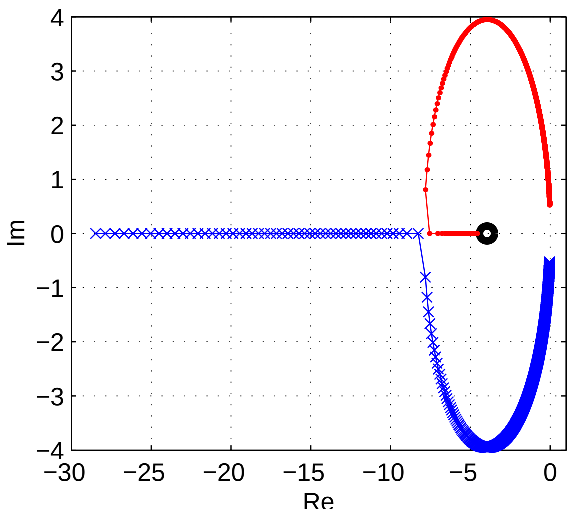


Figure 2.1: The figure illustrates the position of the characteristic roots $\psi_{\epsilon,n}$ (red dots) and $\chi_{\epsilon,n}$ (blue crosses) in the complex plane for different values of ϵ. For this example the parameters are $L = 1$, $\tau = 0.1$, $n = 2$ while ϵ varies between 12 and 0.55 in steps of 0.005. For relatively large ϵ the roots are complex. Due to the real coefficients of the characteristic polynomial (2.39), $\psi_{\epsilon,n}$ and $\chi_{\epsilon,n}$ are complex conjugated to each other. If ϵ falls below a certain threshold value, $\psi_{n,\epsilon}$ converges on the real axis from left to $-\lambda_n$ (thick bullet) while $\chi_{\epsilon,n}$ tends monotonically to $-\infty$.

Let us apply the lemma to the case of our interest. By identifying a with ω and b with $\omega\lambda_n$ we conclude immediately:

Corollary 2.2. *The roots (2.40) and (2.41) of the characteristic polynomial associated to (2.38) display the following asymptotic behavior:*

$$\psi_{\epsilon,n} \xrightarrow{\epsilon\downarrow 0} -\lambda_n \quad \textit{and} \quad \chi_{\epsilon,n} \xrightarrow{\epsilon\downarrow 0} -\infty\,.$$

Now we are enabled to verify the convergence of the Fourier coefficient functions $\sigma_{\epsilon,n}$. Besides the boundedness and the presented implications, this is the second ingredient entering the convergence proof of theorem 2.3.

Lemma 2.6. *If ϵ tends to zero, the Fourier coefficient functions $\sigma_{\epsilon,n}$ defined in (2.43) converge pointwise for all $t \geq 0$ to the Fourier coefficient functions $\sigma_n(t)$ pertaining to the solution of the limit problem (2.30):*

$$\sigma_{\epsilon,n}(t) \xrightarrow{\epsilon\downarrow 0} \sigma_n(t) = \alpha_n \mathrm{e}^{-\lambda_n t}.$$

Proof: For $t = 0$ the initial condition yields: $\sigma_{\epsilon,n}(0) = \alpha_n = \sigma_n(0)$. So the case $t > 0$ remains to be considered. Let $n \in \mathbb{N}$ be fixed. Since $\chi_{\epsilon,n}$ converges to minus infinity for $\epsilon \downarrow 0$, there is $\epsilon_0 = \epsilon_0(n)$ such that $\chi_{\epsilon,n} \neq 0$ for all admissible $\epsilon < \epsilon_0$. From corollary 2.2 we conclude that $\psi_{\epsilon,n}/\chi_{\epsilon,n} \to 0$ and $\beta_n/\chi_{\epsilon,n} \to 0$ if $\epsilon \to 0$. Continuity of the exponential function entails $\mathrm{e}^{\psi_{\epsilon,n}t} \to \mathrm{e}^{-\lambda_n t}$ and $\mathrm{e}^{\chi_{\epsilon,n}t} \to 0$ for fixed $t > 0$.

$$\begin{aligned}\sigma_n(t) &= \alpha_n \mathrm{e}^{-\lambda_n t} \\ &= \frac{\lim\limits_{\epsilon\downarrow 0}\left(\frac{\beta_n}{\chi_{\epsilon,n}}\right) - \alpha_n}{\lim\limits_{\epsilon\downarrow 0}\left(\frac{\psi_{\epsilon,n}}{\chi_{\epsilon,n}}\right) - 1} \cdot \lim_{\epsilon\downarrow 0} \mathrm{e}^{\psi_{\epsilon,n}t} + \frac{\alpha_n \lim\limits_{\epsilon\downarrow 0}\left(\frac{\psi_{\epsilon,n}}{\chi_{\epsilon,n}}\right) - \lim\limits_{\epsilon\downarrow 0}\left(\frac{\beta_n}{\chi_{\epsilon,n}}\right)}{\lim\limits_{\epsilon\downarrow 0}\left(\frac{\psi_{\epsilon,n}}{\chi_{\epsilon,n}}\right) - 1} \cdot \lim_{\epsilon\downarrow 0} \mathrm{e}^{\chi_{\epsilon,n}t}.\end{aligned}$$

As all limits exist, we are allowed to apply the *limit rules* for arithmetic operations:

$$\sigma_n(t) = \lim_{\epsilon\downarrow 0}\left[\frac{\frac{\beta_n}{\chi_{\epsilon,n}} - \alpha_n}{\frac{\psi_{\epsilon,n}}{\chi_{\epsilon,n}} - 1} \cdot \mathrm{e}^{\psi_{\epsilon,n}t} + \frac{\alpha_n \frac{\psi_{\epsilon,n}}{\chi_{\epsilon,n}} - \frac{\beta_n}{\chi_{\epsilon,n}}}{\frac{\psi_{\epsilon,n}}{\chi_{\epsilon,n}} - 1} \cdot \mathrm{e}^{\chi_{\epsilon,n}t}\right] = \lim_{\epsilon\downarrow 0} \sigma_{\epsilon,n}(t)\,.$$

Multiplying denominator and numerator by $\chi_{\epsilon,n}$ yields (2.43) and hence the assertion. ∎

2.2.2 Solution of the perturbed problem and convergence

In (2.37) we have chosen for u_ϵ a Fourier series ansatz without knowing, whether this will lead to a reasonable result. In order to impart a meaning to (2.37) and to justify thereby the ansatz, we have to show that the sequence is convergent in some sense. This is the intention of the next proposition.

Proposition 2.2. *Let the coefficient functions $\sigma_{\epsilon,n}$ be as in (2.43). Then*

$$u_\epsilon(t,x) := \sum_{n\in\mathbb{N}} \sigma_{\epsilon,n}(t) s_n(x)$$

defines for all $t \geq 0$ a square-integrable function, i.e. $u_\epsilon(t,\cdot) \in \mathcal{L}^2(0,L)$.

Proof: Generally, a series of pairwise orthogonal elements converges in a Hilbert space if and only if the sum of the their squared norms is convergent. As the s_n's are normalized in $\mathcal{L}^2(0,L)$, we have to check, whether the sum of the squared moduli $|\sigma_{\epsilon,n}(t)|^2$ is finite. But this is just affirmed by lemma 2.2. ∎

The proposition does not yet classify u_ϵ as a function of time. This is the object of the next theorem.

Theorem 2.2. *The function u_ϵ defined in proposition 2.2 depends continuously on the time, more exactly:*

$$u_\epsilon \in \mathcal{C}\big([0,\infty), \mathcal{L}^2(0,L)\big)$$

As already insinuated in remark 2.5, the proof of this theorem relies on lemma 2.3. Since we do not dispose of equicontinuity in time for the whole sequence $(\sigma_{n,\epsilon})_{n\in\mathbb{N}}$, the Fourier series representing the difference $u_\epsilon(t,\cdot)-u_\epsilon(t_0,\cdot)$ is split into two parts. The first portion comprises only finitely many addends up to a certain index. These terms are handled by exploiting the continuity of the coefficient functions to be transferred on the whole series. The tail of the series, is controlled in a general way by means of the uniform decay provided by lemma 2.3.

Proof: We have to show that for all $t_0 \in [0,\infty)$ and for any $\delta > 0$ there exists a real number $\theta > 0$, such that

$$\|u_\epsilon(t,\cdot) - u_\epsilon(t_0,\cdot)\|_2^2 < \delta \qquad \text{if} \quad |t-t_0| < \theta.$$

Using *Parseval's theorem* and the orthonormality of the s_n's , the squared $\mathcal{L}^2$-norm can be expressed in terms of the Fourier coefficient functions:

$$\begin{aligned}
\|u_\epsilon(t,\cdot) - u(t,\cdot)\|_2^2 &= \Big\| \sum_{n\geq 1} \big[\sigma_{\epsilon,n}(t) - \sigma_{\epsilon,n}(t_0)\big] \cdot s_n \Big\|_2^2 \\
&= \sum_{n\geq 1} \big|\sigma_{\epsilon,n}(t) - \sigma_{\epsilon,n}(t_0)\big|^2 \cdot \|s_n\|_2^2 \\
&= \sum_{n\geq 1} \big|\sigma_{\epsilon,n}(t) - \sigma_{\epsilon,n}(t_0)\big|^2 \ . \qquad (2.55)
\end{aligned}$$

Lemma 2.3 guarantees the existence of $M = M(\delta) \in \mathbb{N}$, such that

$$\sum_{n>M} |\sigma_{\epsilon,n}(t)|^2 < \tfrac{\delta}{3} \qquad \text{and} \qquad \sum_{n>M} |\sigma_{\epsilon,n}(t_0)|^2 < \tfrac{\delta}{3}\ .$$

As each of the coefficient functions is continuous in time, we can find real numbers $\theta_1,...,\theta_M > 0$ with

$$|\sigma_{\epsilon,n}(t) - \sigma_{\epsilon,n}(t_0)| < \tfrac{\delta}{3M} \qquad \text{if} \quad |t-t_0| < \theta_n \quad \text{and} \ \ n \in \{1,...,M\}.$$

Choosing finally $\theta := \min\{\theta_1,...,\theta_M\}$ we obtain from (2.55) by applying the triangle inequality

$$\begin{aligned}
\|u_\epsilon(t,\cdot) - u(t,\cdot)\|_2^2 &\leq \sum_{n=1}^{M} \big|\sigma_{\epsilon,n}(t) - \sigma_{\epsilon,n}(t_0)\big|^2 + \sum_{n>M} |\sigma_{\epsilon,n}(t)|^2 + \sum_{n>M} |\sigma_{\epsilon,n}(t_0)|^2 \\
&\leq M\,\tfrac{\delta}{3M} + \tfrac{\delta}{3} + \tfrac{\delta}{3} = \delta
\end{aligned}$$

whenever $|t-t_0| < \theta$. Note that $\theta_1,...,\theta_M$ and hence θ depend *a priori* (for given initial Fourier coefficients defining $\sigma_{\epsilon,n}$) on the choice of δ and M. If one tries to choose M as small as possible, M depends on δ and therefore θ depends actually only on δ. ∎

At this point it would be natural to expect a statement establishing u_ϵ as the solution of the perturbed problem (2.29). However, in opposition to the limit problem (2.30), there is no result equivalent to theorem 2.1. The reason[18] for the different behavior lies in the second time derivative that makes (2.29, EQ) to a sort of wave equation, that is of hyperbolic nature despite the attenuating action of the first time derivative. It is a characteristic property of hyperbolic equations that their associated time evolution operator does not smooth the regularity of the initial data. In contrast, the solution of a parabolic equation like the diffusion equation in (2.30) acquires $\mathcal{C}^\infty$-regularity as soon as $t > 0$ – even if started by an $\mathcal{L}^2$-function. Therefore it is not possible to assign to u_ϵ the meaning of a classical solution for the perturbed IVBP under the assumption of $\mathcal{L}^2$ initial data. Nevertheless it is possible to interpret u_ϵ as solution in a wider, generalized sense (weak or distributional solution).

In view of these facts, the convergence of u_ϵ towards the $\mathcal{C}^\infty$-solution u of the limit problem might appear questionable. On the other hand, having verified the convergence of the Fourier coefficient functions in lemma 2.6, it is natural to assume that this fact should have some repercussions. A first convergence result is given by the following theorem. Its proof is based on the same idea of splitting to control the Fourier series that has been applied to verify theorem 2.2.

Theorem 2.3. *For fixed time $t \geq 0$, $u_\epsilon(t,\cdot)$ converges to $u(t,\cdot)$ with respect to the $\mathcal{L}^2$-norm as ϵ tends to zero (satisfying assumption 2.1):*

$$\|u_\epsilon(t,\cdot) - u(t,\cdot)\|_2 \xrightarrow{\epsilon \downarrow 0} 0 \, .$$

Proof: Thanks to proposition 2.2 and theorem 2.1 both of the functions u_ϵ, u are in $\mathcal{L}^2\big((0,L)\big)$ for fixed $t \geq 0$, so that the assertion makes sense. We have to show, that for any $\delta > 0$ there exists an $\eta > 0$, such that the estimate

$$\|u_\epsilon(t,\cdot) - u(t,\cdot)\|_2^2 < \delta$$

holds for all admissible ϵ with $0 < \epsilon < \eta$. We can proceed as in the proof of theorem 2.2 with some slight adaptations concerning the formulation.
Lemma 2.3 can be extended to the sequence $\sigma_n(t)$ corresponding formally to the limit case $\epsilon = 0$. This ensures the existence of $M = M(\delta) \in \mathbb{N}$, such that

$$\sum_{n>M} |\sigma_{\epsilon,n}(t)|^2 < \tfrac{\delta}{3} \qquad \text{and} \qquad \sum_{n>M} |\sigma_n(t)|^2 < \tfrac{\delta}{3} \, .$$

Furthermore we know from lemma 2.6, that each of the coefficient functions $\sigma_{n,\epsilon}(t)$ converges to $\sigma_n(t)$. Therefore we can find real numbers $\eta_1, ..., \eta_M > 0$ with

$$|\sigma_{\epsilon,n}(t) - \sigma_n(t)| < \tfrac{\delta}{3M} \qquad \text{if} \quad \epsilon < \eta_n \quad \text{and} \quad n \in \{1, ..., M\}.$$

[18] We refer to the appendix for a more thorough discussion.

By setting $\eta := \min\{\eta_1, ..., \eta_M\}$ we have found the searched η, because

$$\begin{aligned}\|u_\epsilon(t,\cdot) - u(t,\cdot)\|_2^2 &= \sum_{n\geq 1} \left|\sigma_{\epsilon,n}(t) - \sigma_n(t)\right|^2 \\ &\leq \sum_{n=1}^{M} \left|\sigma_{\epsilon,n}(t) - \sigma_n(t)\right|^2 + \sum_{n>M} |\sigma_{\epsilon,n}(t)|^2 + \sum_{n>M} |\sigma_n(t)|^2 \\ &\leq M\,\tfrac{\delta}{3M} + \tfrac{\delta}{3} + \tfrac{\delta}{3} = \delta\end{aligned}$$

for all $0 < \epsilon < \eta$. ∎

The degenerate case

Let us briefly discuss the degenerate case excluded so far by assumption 2.1. This case occurs, if $\epsilon = \frac{L}{2\sqrt{\tau}\pi n}$ for some $n \in \mathbb{N}$ which implies

$$W_{\epsilon,n} = 0 \quad \text{and equivalently} \quad \psi_{\epsilon,n} = \chi_{\epsilon,n} = -\tfrac{1}{2}\omega/\epsilon^2.$$

Then the fundamental system of (2.38, EQ) is given by

$$t \mapsto \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \qquad \text{and} \qquad t \mapsto t\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\,.$$

Taking the initial condition (2.38,IC) into account we obtain for $\sigma_{\epsilon,n}$:

$$\begin{aligned}\sigma_{\epsilon,n}(t) &= \alpha_n \mathrm{e}^{\psi_{\epsilon,n} t} + (\beta_n - \alpha_n \psi_{\epsilon,n})\, t\mathrm{e}^{\psi_{\epsilon,n} t} \\ &= \alpha_n \mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} + (\beta_n + \tfrac{1}{2}\alpha_n \omega/\epsilon^2)\, t\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \\ &= \alpha_n (1 - \tfrac{1}{2}\omega t/\epsilon^2)\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} + \beta_n t\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}\,.\end{aligned}$$

Hence $\rho_{\epsilon,n}(t) = (1-\frac{1}{2}\omega t/\epsilon^2)\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}$ and $\varrho_{\epsilon,n}(t) = t\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2}$. The triangle inequality and lemma 2.8 lead to the estimates

$$|\rho_{\epsilon,n}(t)| = |1 - \tfrac{1}{2}\omega t/\epsilon^2|\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \leq 2 \tag{2.56}$$

$$|\varrho_{\epsilon,n}(t)| = 2\tau\epsilon^2 \cdot (\tfrac{1}{2}\omega t/\epsilon^2)\mathrm{e}^{-\frac{1}{2}\omega t/\epsilon^2} \leq 2\tau\epsilon^2, \tag{2.57}$$

that agree with those in lemma 2.1 and corollary 2.1, except for an additional factor[19] 2 in (2.57).

Let us suppose now that ϵ represents a general null-sequence, which does not fulfill assumption 2.1. It turns out that the convergence proof of theorem 2.3 works without any modification. Let us see why:

Remember that $M = M(\delta) \in \mathbb{N}$ was chosen at first for given $\delta > 0$. By means of this natural number the infinite series is decomposed into two parts. The leading part for the indices $n \in \{1, ..., M\}$ is estimated utilizing the convergence of the Fourier coefficients $\sigma_{\epsilon,n}(t) \to \sigma_n(t)$. If $\delta > 0$ is small enough, this requires usually $W_{\epsilon,n} > 0$ (otherwise $\psi_{\epsilon,n}$ is still relatively far from λ_n away.) Therefore the case $W_{\epsilon,n} = 0$ should occur for $n > M$, i.e. in the tail of the series.

[19]Fortunately, this is of no harm. The factor can also be added in corollary 2.1 to have a common estimate.

In order to enforce this, we additionally impose $\epsilon < L/(2\sqrt{\tau}M\pi)$ implying$W_{n,\epsilon} > 0$ for $n \in \{1, ..., M\}$. So, η is defined by

$$\eta := \min\{\tfrac{L}{2\sqrt{\tau}M\pi}, \eta_1, ..., \eta_M\}.$$

Thus, the degenerate case can only occur "in the tail" of the series. But here we benefit from the boundedness (uniform in t and n) of the coefficients functions $\rho_{n,\epsilon}$ and $\varrho_{n,\epsilon}$, so that we have not to distinguish between the *standard* and the *degenerate* case.
In retrospect, assumption 2.1 was not really necessary as "measure of precaution", as the following function is continuous with respect to ϵ:

$$\epsilon \mapsto \sigma_{\epsilon,n}(t) = \begin{cases} \alpha_n \frac{\psi_{\epsilon,n} e^{\chi_{\epsilon,n}t} - \chi_{\epsilon,n} e^{\psi_{\epsilon,n}t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}} + \beta_n \frac{e^{\psi_{\epsilon,n}t} - e^{\chi_{\epsilon,n}t}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}} & \text{if } \epsilon \neq \frac{L}{2\sqrt{\tau}M\pi} \\ \alpha_n e^{\psi_{\epsilon,n}t} + (\beta_n - \alpha_n \psi_{\epsilon,n})\, t e^{\psi_{\epsilon,n}t} & \text{if } \epsilon = \frac{L}{2\sqrt{\tau}M\pi} \end{cases}.$$

2.2.3 Uniform convergence and convergence rate

In the preceding subsection we have proved pointwise convergence of u_ϵ in time under the weak hypothesis of $\mathcal{L}^2$ initial data. The goal of this section is to extend this result. So we will prove *uniform* temporal convergence and derive a convergence rate. The latter one requires however certain smoothness assumptions concerning the initial data. Another aspect to be discussed here concerns the convergence behavior of the time derivative $\partial_t u_\epsilon$, where we will make a first acquaintance with the *initial layer* to be discussed in section 2.3.

The singular IVP for the Fourier coefficient functions

The original problem (2.29) involving a PDE is reformulated by means of Fourier expansion into a family (parametrized by n) of equally structured ODE initial value problems (2.38). Before we can ameliorate the result of theorem 2.3, we have to study their solutions in more detail.
For this, let us return to the computation of the characteristic roots given in (2.40) and (2.41). Taylor expansion yields for $1 > x$

$$\sqrt{1-x} = 1 - \tfrac{1}{2}x - \tfrac{1}{8}(1-\xi)^{-3/2}x^2,$$

where $\xi \in (0, x)$ denotes an appropriate intermediate position (Lagrange's remainder). Applying this to the root occurring in (2.42) we get

$$W_{\epsilon,n} := \sqrt{1 - 4\epsilon^2\tau\lambda_n} = 1 - 2\epsilon^2\tau\lambda_n - c_{\epsilon,n}\epsilon^4 \qquad \text{if} \quad 4\epsilon^2\tau\lambda_n < 1. \tag{2.58}$$

Here $c_{\epsilon,n}$ is a *positive* number depending on ϵ and on n via λ_n. If the argument $4\epsilon^2\tau\lambda_n$ stays away from 1, we can compute an upper bound for $c_{\epsilon,n}$. For instance:

$$0 < 4\tau\lambda_n\epsilon^2 < \tfrac{5}{9} \qquad \Rightarrow \qquad 0 < c_{\epsilon,n} < \tfrac{1}{8}(1 - \tfrac{5}{9})^{-3/2} \cdot (4\tau\lambda_n)^2 < 8\tau^2\lambda_n^2. \tag{2.59}$$

The expansion of $W_{\epsilon,n}$ immediately yields the following expressions for the characteristic roots:

$$\left.\begin{array}{rcl} \psi_{\epsilon,n} & = & -\lambda_n - \frac{1}{2}\omega c_{\epsilon,n}\epsilon^2 \\ \chi_{\epsilon,n} & = & -\omega\epsilon^{-2} + \lambda_n + \frac{1}{2}\omega c_{\epsilon,n}\epsilon^2 \end{array}\right\} \quad \Rightarrow \quad \psi_{\epsilon,n} - \chi_{\epsilon,n} = \omega\epsilon^{-2} - 2\lambda_n - \omega c_{\epsilon,n}\epsilon^2.$$

Thence:

$$\frac{\beta_n - \alpha_n\chi_{\epsilon,n}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}} = \frac{\alpha_n - \alpha_n\lambda_n\tau\epsilon^2 + \beta_n\tau\epsilon^2 - \frac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4}{1 - 2\lambda_n\tau\epsilon^2 - c_{\epsilon,n}\epsilon^4} = \alpha_n + \mathrm{O}(\epsilon^2)\,, \tag{2.60}$$

$$\frac{\alpha_n\psi_{\epsilon,n} - \beta_n}{\psi_{\epsilon,n} - \chi_{\epsilon,n}} = \frac{(-\alpha_n\lambda_n - \beta_n)\tau\epsilon^2 - \frac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4}{1 - 2\lambda_n\tau\epsilon^2 - c_{\epsilon,n}\epsilon^4} = \mathrm{O}(\epsilon^2)\,. \tag{2.61}$$

Let us define for given $n \in \mathbb{N}$ and $\lambda_n = \tau\frac{n^2\pi^2}{L^2}$

$$\eta_n := \tfrac{1}{3}\tfrac{1}{\sqrt{\tau\lambda_n}}. \tag{2.62}$$

In the sequel we will often tacitly assume $\epsilon \le \eta_n$. We underline, that in this case the root $W_{\epsilon,n}$ permits the expansion in (2.58) with $c_{\epsilon,n}$ satisfying the estimate (2.59). In particular, this implies also that we can expand the denominator in (2.60) and (2.61) by the geometric series, providing thus a justification of the claimed asymptotic behavior for $\epsilon \to 0$. Additionally, we obtain the following useful estimate:

$$\frac{1}{1 - 2\lambda_n\tau\epsilon^2 - c_{\epsilon,n}\epsilon^4} \le \frac{1}{1 - \frac{1}{2}\cdot\frac{5}{9} - \frac{8}{81}} < 2. \tag{2.63}$$

Furthermore we will make use of the crude estimates $\mathrm{e}^{\psi_{\epsilon,n}t} < 1$ and $\mathrm{e}^{\chi_{\epsilon,n}t} < 1$ for $t > 0$ and $\epsilon < \eta_n$. In the proof of proposition 2.4 the latter exponential, however, requires a much finer estimate, which is supplied already now.

Lemma 2.7. *Let η_n be defined as in (2.62). Then for any $\theta > 0$ and all $\epsilon \in (0, \eta_n)$*

$$\mathrm{e}^{\chi_{\epsilon,n}t} < \tau\theta^{-1}\epsilon^2 \qquad \text{if } t \in [\theta, \infty).$$

Proof: Note that $W_{\epsilon,n}$ and $\chi_{\epsilon,n}$ are real for $\epsilon \in (0, \eta_n)$. Then $\chi_{\epsilon,n}$ is dominated by $-\frac{1}{2}\omega\epsilon^{-2}$ according to (2.42). Now we seek a constant $M_\theta > 0$, such that

$$\mathrm{e}^{\chi_{\epsilon,n}t} < \mathrm{e}^{-\omega t/2\epsilon^2} \overset{!}{\le} M_\theta\epsilon^2 \quad \Leftrightarrow \quad f(\epsilon) := \frac{\mathrm{e}^{-\omega t/2\epsilon^2}}{\epsilon^2} \le M_\theta.$$

In order to define M_θ suitably, we have to determine the maximum of the function f. A short computation reveals, that its derivative has a single positive zero at $\sqrt{\omega t/2}$. Since $f(0) = f(\infty) = 0$ and $f(\sqrt{\omega t/2}) = (\mathrm{e}\omega t/2)^{-1} > 0$ we have obtained the maximum, which is obviously decreasing with respect to t. So we choose $M_\theta = \sup_{t\in[\theta,\infty)} f(\sqrt{\omega t/2}) = (\mathrm{e}\omega\theta/2)^{-1} = \frac{2\tau}{\mathrm{e}\theta} < \frac{\tau}{\theta}$. ∎

After these preparations the first proposition represents a significant supplement of lemma 2.6. Although (2.38) is a singular perturbation of problem (2.35), this has to all appearances no impact on the quality of mutual approximation between their solutions $\sigma_{\epsilon,n}$ and σ_n. Similarly to a regularly perturbed problem, the factor ϵ^2 in front of the second time derivative in (2.38) seems to determine the approximation speed in a natural way.

Proposition 2.3. *If ϵ tends to 0, the Fourier coefficient function $t \mapsto \sigma_{\epsilon,n}(t)$, defined in (2.43), converges uniformly on $[0,\infty)$ towards $\sigma_n(t) = \alpha_n\mathrm{e}^{-\lambda_n t}$.*
There exists a constant C_n depending on τ and n via λ_n and the initial conditions α_n, β_n, such that

$$\sup_{t\in[0,\infty)} |\sigma_{\epsilon,n}(t) - \sigma_n(t)| < C_n\epsilon^2 \qquad \text{for } 0 < \epsilon < \eta_n.$$

Proof: It suffices to prove the above estimate, as uniform convergence ensues from it immediately. Using the definition of $\sigma_{\epsilon,n}(t)$ in (2.43) and applying the triangle inequality yields

$$|\sigma_{\epsilon,n}(t)-\sigma_n(t)| \le \underbrace{\left|\frac{\beta_n-\alpha_n\chi_{\epsilon,n}}{\psi_{\epsilon,n}-\chi_{\epsilon,n}}\mathrm{e}^{\psi_{\epsilon,n}t}-\alpha_n\mathrm{e}^{-\lambda_n t}\right|}_{=:E_1}+\underbrace{\left|\frac{\alpha_n\psi_{\epsilon,n}-\beta_n}{\psi_{\epsilon,n}-\chi_{\epsilon,n}}\mathrm{e}^{\chi_{\epsilon,n}t}\right|}_{=:E_2}.$$

Let us treat each of these expressions separately:

$$E_1 \overset{(2.60)}{=} \left|\frac{(\alpha_n-\alpha_n\lambda_n\tau\epsilon^2+\beta_n\tau\epsilon^2-\frac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4)\mathrm{e}^{\psi_{\epsilon,n}t}-\big(1-2\lambda_n\tau\epsilon^2-c_{\epsilon,n}\epsilon^4\big)\alpha_n\mathrm{e}^{-\lambda_n t}}{1-2\lambda_n\tau\epsilon^2-c_{\epsilon,n}\epsilon^4}\right|$$

We estimate the denominator by 2 as in (2.63) and apply the triangle inequality to the numerator:

$$\begin{aligned}
E_1 &\le 2\left|\alpha_n\mathrm{e}^{\psi_{\epsilon,n}t}-\alpha_n\mathrm{e}^{-\lambda_n t}\right|+2\epsilon^2\left|\beta_n\tau-\alpha_n\lambda_n\tau-\tfrac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^2+2\alpha_n\lambda_n\tau+\alpha_n c_{\epsilon,n}\epsilon^2\right|\\
&\le 2\mathrm{e}^{-\lambda_n t}|\alpha_n|\cdot\left|\mathrm{e}^{-\frac{1}{2}\omega c_{\epsilon,n}\epsilon^2 t}-1\right| + 2\epsilon^2\left|\beta_n\tau+\alpha_n\lambda_n\tau+\tfrac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^2\right|\\
&\le 2\mathrm{e}^{-\lambda_n t}|\alpha_n|\cdot\tfrac{1}{2}\omega c_{\epsilon,n}\epsilon^2 t + 2\epsilon^2\big(|\beta_n\tau+\alpha_n\lambda_n\tau|+\tfrac{1}{2}|\alpha_n|c_{\epsilon,n}\epsilon^2\big).
\end{aligned}$$

In the last step we have applied the mean value theorem to obtain ϵ^2 as a factor also in the first addend:

$$|\mathrm{e}^{-\frac{1}{2}\omega c_{\epsilon,n}\epsilon^2 t}-1| = \tfrac{1}{2}\omega c_{\epsilon,n}\epsilon^2\underbrace{\mathrm{e}^{-\frac{1}{2}\omega c_{\epsilon,n}\epsilon^2\vartheta}}_{\le 1}t \qquad \text{with } 0<\vartheta<t.$$

In order to get rid of the factor t, we remark, that on account of lemma 2.8 and $\lambda_n>0$, we get $\lambda_n t\mathrm{e}^{-\lambda_n t}<1$ for all $t\ge 0$. So we end up with

$$\begin{aligned}
E_1 &\overset{(2.59)}{\underset{(2.62)}{\le}} 2\mathrm{e}^{-\lambda_n t}|\alpha_n|\cdot\tfrac{1}{2}\omega 8\tau^2\lambda_n^2 t\epsilon^2 + 2\epsilon^2\tau\big(|\beta_n+\alpha_n\lambda_n|+\tfrac{1}{2}|\alpha_n|8\lambda_n^2\tau^2\tfrac{1}{9}\tfrac{1}{\tau\lambda_n}\big)\\
&\le \Big(9|\alpha_n|\lambda_n+2|\beta_n+\alpha_n\lambda_n|\Big)\tau\,\epsilon^2. \qquad (2.64)
\end{aligned}$$

Similarly we treat E_2:

$$\begin{aligned}
E_2 &\overset{(2.61)}{=} \left|\frac{(-\alpha_n\lambda_n-\beta_n)\tau\epsilon^2-\frac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4}{1-2\lambda_n\tau\epsilon^2-c_{\epsilon,n}\epsilon^4}\mathrm{e}^{\chi_{\epsilon,n}t}\right| \le 2\left|(\alpha_n\lambda_n+\beta_n)\tau\epsilon^2+\tfrac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4\right|\\
&\le 2\Big(|\alpha_n\lambda_n+\beta_n|+\tfrac{1}{2}|\alpha_n|\lambda_n\Big)\tau\,\epsilon^2.
\end{aligned}$$

Thus, both expressions E_1, E_2 vanish quadratically with ϵ and we find the assertion verified with

$$C_n := \Big(14|\alpha_n|\lambda_n+4|\beta_n|\Big)\tau\,. \qquad (2.65)$$

■

The next proposition manifests the "singular" character of the limit. We have already seen that the choice of the second initial condition β_n does not affect the convergence $\sigma_{\epsilon,n} \xrightarrow{\epsilon\downarrow 0} \sigma_n$. This means that convergence occurs indifferently of whether β_n is chosen *compatibly*[20] or not. On the contrary, it turns out, that even the time derivative $\dot\sigma_{\epsilon,n}$ converges to $\dot\sigma_n$ for $\epsilon \downarrow 0$. Of course, in any neighborhood of $t = 0$, this convergence cannot be uniform if β_n is not compatible, i.e. $\beta_n = \dot\sigma_{\epsilon,n}(0) \neq \dot\sigma_n(0)$. However uniform convergence takes place outside any vicinity of the time origin $t = 0$. This circumstance clearly signalizes the existence of an initial layer.

Proposition 2.4. *Let ϵ tend to zero. Then the time derivative of the Fourier coefficient $\sigma_{\epsilon,n}$ converges for any $\theta > 0$ uniformly on $[\theta,\infty)$ towards $\dot\sigma_n(t) = -\alpha_n\lambda_n \mathrm{e}^{-\lambda_n t}$. Moreover, there exists a constant $K_{\theta,n} > 0$ such that*

$$\sup_{t\in[\theta,\infty)} |\dot\sigma_{\epsilon,n}(t) - \dot\sigma_n(t)| < K_{\theta,n}\epsilon^2 \qquad \text{for } 0 < \epsilon < \eta_n.$$

For $\beta_n = -\alpha_n\lambda_n$ the above estimate and a fortiori *uniform convergence hold even in the case of $\theta = 0$.*

Proof: Recall that $\epsilon < \eta_n$ entails $c_{\epsilon,n} < 8\lambda_n^2\tau^2$ and (2.63). Furthermore $\mathrm{e}^{\chi_{\epsilon,n}t} < 1$ by lemma 2.7 as well as $\mathrm{e}^{\psi_{\epsilon,n}t} < 1$ for $t \geq \theta > 0$. We proceed now analogously to the proof of the previous proposition:

$$|\dot\sigma_{\epsilon,n}(t) - \dot\sigma_n(t)| \leq \underbrace{\left|\frac{\beta_n - \alpha_n\chi_{\epsilon,n}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\psi_{\epsilon,n}\mathrm{e}^{\psi_{\epsilon,n}t} + \alpha_n\lambda_n\mathrm{e}^{-\lambda_n t}\right|}_{=:F_1} + \underbrace{\left|\frac{\alpha_n\psi_{\epsilon,n} - \beta_n}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\chi_{\epsilon,n}\mathrm{e}^{\chi_{\epsilon,n}t}\right|}_{=:F_2}.$$

Using $\psi_{\epsilon,n} = -\lambda_n - \frac{1}{2}\omega c_{\epsilon,n}\epsilon^2$ we find

$$\begin{aligned} F_1 \;&\leq\; \lambda_n\left|\frac{\beta_n - \alpha_n\chi_{\epsilon,n}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\mathrm{e}^{\psi_{\epsilon,n}t} - \alpha_n\mathrm{e}^{-\lambda_n t}\right| \;+\; \tfrac{1}{2}\omega c_{\epsilon,n}\epsilon^2 \underbrace{\left|\frac{\beta_n - \alpha_n\chi_{\epsilon,n}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\mathrm{e}^{\psi_{\epsilon,n}t}\right|}_{=:G_1} \\ &\leq\; \lambda_n E_1 + 4\tau\lambda_n^2\epsilon^2 G_1 \leq\; \Big(21|\alpha_n|\lambda_n^2 + 8|\beta_n|\lambda_n\Big)\tau\,\epsilon^2\,, \end{aligned}$$

where G_1 is estimated as follows:

$$\begin{aligned} G_1 &\leq \left|\frac{\beta_n - \alpha_n\chi_{\epsilon,n}}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\right| \;\leq\; 2\Big|\alpha_n - \alpha_n\lambda_n\tau\epsilon^2 + \beta_n\tau\epsilon^2 - \tfrac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4\Big| \\ &\leq 2|\alpha_n| + \tfrac{2}{9}|\alpha_n| + \tfrac{2}{9}|\beta_n|\lambda_n^{-1} + \tfrac{8}{81}|\alpha_n| \;\leq\; \tfrac{7}{3}|\alpha_n| + \tfrac{1}{3}|\beta_n|\lambda_n^{-1}. \end{aligned}$$

[20] β_n is *compatible* to the solution of the limit problem (2.35) if $\beta_n = \dot\sigma_n(0) = -\alpha_n\lambda_n$.

Similarly, employing $\chi_{\epsilon,n} = -\omega\epsilon^{-2} + \lambda_n + \omega c_{\epsilon,n}\epsilon^2$ we get for F_2:

$$\begin{aligned}
F_2 &\le \omega\epsilon^{-2}\Big|\frac{\alpha_n\psi_{\epsilon,n} - \beta_n}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\mathrm{e}^{\chi_{\epsilon,n}t}\Big| \;+\; (\lambda_n + \omega c_{\epsilon,n}\epsilon^2)\,\underbrace{\Big|\frac{\alpha_n\psi_{\epsilon,n} - \beta_n}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\mathrm{e}^{\chi_{\epsilon,n}t}\Big|}_{=:G_2} \\
&\le \omega\epsilon^{-2}\,2\,\big|(\alpha_n\lambda_n + \beta_n)\tau\epsilon^2 + \tfrac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4\big|\mathrm{e}^{\chi_{\epsilon,n}t} \;+\; (\lambda_n + \omega c_{\epsilon,n}\epsilon^2)G_2 \\
&\le 2|\alpha_n\lambda_n + \beta_n|\mathrm{e}^{\chi_{\epsilon,n}t} \;+\; 8|\alpha_n|\lambda_n^2\tau\epsilon^2 + 2\lambda_n G_2 \\
&\le 2|\alpha_n\lambda_n + \beta_n|\mathrm{e}^{\chi_{\epsilon,n}t} \;+\; \tau\Big(14|\alpha_n|\lambda_n^2 + 4|\beta_n|\lambda_n\Big)\epsilon^2\,.
\end{aligned}$$

where G_2 has been replaced by the estimate

$$\begin{aligned}
G_2 &\le \Big|\frac{\alpha_n\psi_{\epsilon,n} - \beta_n}{\psi_{\epsilon,n} - \chi_{\epsilon,n}}\Big| \;\le\; 2\big|(\alpha_n\lambda_n + \beta_n)\tau\epsilon^2 + \tfrac{1}{2}\alpha_n c_{\epsilon,n}\epsilon^4\big| \\
&\le 2\Big(|\alpha_n\lambda_n + \beta_n| + \tfrac{1}{2}|\alpha_n|\lambda_n\Big)\tau\,\epsilon^2\,.
\end{aligned}$$

Summarizing, we have derived the estimate

$$|\dot\sigma_{\epsilon,n}(t) - \dot\sigma_n(t)| \le 2|\alpha_n\lambda_n + \beta_n|\mathrm{e}^{\chi_{\epsilon,n}t} + \overbrace{\Big(35|\alpha_n|\lambda_n^2 + 12|\beta_n|\lambda_n\Big)\tau}^{K_n:=}\;\epsilon^2$$

If $\beta_n = -\alpha_n\lambda_n$ the first addend vanishes and the asserted estimate holds for all $t \ge 0$ with $K_{\theta,n} = K_n$. Otherwise if the initial condition of the perturbed problem is not consistent with the limit problem i.e. $|\alpha_n\lambda_n + \beta_n| \ne 0$, we take recourse to lemma 2.7. Thus we obtain the assertion by setting

$$K_{\theta,n} = 2\big(|\alpha_n|\lambda_n + \beta_n\big)\tau/\theta + K_n\ .$$

■

Main convergence results

Uniform convergence of u_ϵ can be obtained as corollary of the previous results.

Corollary 2.3. *Let the initial data $g, h \in \mathcal{L}^2(0, L)$. Then u_ϵ as defined in proposition 2.2 converges to the solution u of (2.30) in the space*[21] $\mathcal{C}_b\big([0,\infty), \mathcal{L}^2(0,L)\big)$*, i.e. uniformly in time:*

$$\sup_{t\in[0,\infty)} \|u_\epsilon(t,\cdot) - u(t,\cdot)\|_2 \xrightarrow{\epsilon\downarrow 0} 0.$$

Proof: The assertion can be concluded from proposition 2.3 in the same way as theorem 2.3 has been inferred from lemma 2.6. The splitting of the Fourier series works, because all estimates in subsection 2.2.1 hold uniformly in time. ■

[21] $\mathcal{C}_b\big([0,\infty), \mathcal{L}^2(0,L)\big)$ denotes the space of bounded continuous function over $[0,\infty)$. Observe that the boundedness is necessary in the context of unbounded domains to define the norm reasonably.

In contrast, the transfer of the estimate in proposition 2.3 turns out to be more subtle and requires an additional regularity condition.

Assumption 2.2. We suppose in the sequel that the initial Fourier coefficients are not only square-summable, but also satisfy the stronger condition

$$A_\lambda := \sum_{n\geq 1} |\alpha_n|\lambda_n < \infty \qquad B := \sum_{n\geq 1} |\beta_n| < \infty \,. \tag{2.66}$$

What does this condition mean with respect to the initial conditions g and h? As the series of the β_n's is absolutely convergent, the Fourier series of h converges normally, what gives continuity of h thanks to proposition 2.6. In order to interpret the first condition, we must be aware of $\lambda_n \propto n^2$. Since $(n^2\alpha_n)_n$ and *a fortiori* $(n\alpha_n)_n$ and $(\alpha_n)_n$ are absolutely convergent, g is twice continuously differentiable[22].

Theorem 2.4. *Let the Fourier coefficients of the initial data satisfy condition 2.66. Then there exists a constant $C > 0$ depending only on the initial data, L and τ such that*

$$\sup_{t\in[0,\infty)} \|u_\epsilon(t,\cdot) - u(t,\cdot)\|_\infty < C\epsilon^2 \qquad \textit{for} \quad 0 < \epsilon < \eta_1$$

with η_1 defined in (2.62).

Proof: For $0 < \epsilon < \eta_1$ given, define $N_\epsilon := \max\{n \in \mathbb{N} : n \leq \frac{1}{3}\frac{L}{\tau\pi\epsilon}\}$. Condition (2.66) ensures by proposition 2.6 that $u_\epsilon, u \in \mathcal{C}_b\big([0,\infty)\times[0,L]\big)$.
Applying the triangle inequality and using $\|s_n\|_\infty \leq \sqrt{2/L}$ we obtain

$$\sup_{t\in[0,\infty)} \big\|u_\epsilon(t,\cdot) - u(t,\cdot)\big\|_\infty = \sup_{t\in[0,\infty)} \Big\|\sum_{n\geq 1} \big(\sigma_{\epsilon,n}(t) - \sigma_n(t)\big)s_n\Big\|_\infty$$
$$\leq \underbrace{\sqrt{\tfrac{2}{L}} \sum_{n=1}^{N_\epsilon} \sup_{t\in[0,\infty)} |\sigma_{\epsilon,n}(t) - \sigma_n(t)|}_{=:S_1} + \underbrace{\sqrt{\tfrac{2}{L}} \sum_{n>N_\epsilon} \sup_{t\in[0,\infty)} |\sigma_{\epsilon,n}(t) - \sigma_n(t)|}_{=:S_2} .$$

Note that $\eta_n \geq \epsilon$ for $1 \leq n \leq N_\epsilon$ because

$$\eta_n = \frac{1}{3}\frac{1}{\sqrt{\tau\lambda_n}} = \frac{1}{3}\frac{\omega L}{\pi n} \geq \frac{1}{3}\frac{\omega L}{\pi N_\epsilon} \geq \frac{1}{3}\frac{\omega L}{\pi}\left(\frac{1}{3}\frac{L}{\tau\pi\epsilon}\right)^{-1} = \epsilon.$$

Therefore S_1 can be estimated by means of proposition 2.3 with the constants C_n specified in (2.65).

$$S_1 < \sum_{n=1}^{N_\epsilon} C_n\epsilon^2 = \Big(14\sum_{n=1}^{N_\epsilon} |\alpha_n|\lambda_n + 4\sum_{n=1}^{N_\epsilon} |\beta_n|\Big)\tau\epsilon^2 < \big(14A_\lambda + 4B\big)\tau\epsilon^2 .$$

The boundedness of the sums in (2.66) permits us to estimate the factor in front of $\tau\epsilon^2$ independently of N_ϵ and hence of ϵ.

[22] It must be shown that differentiation and summation of the Fourier series can be interchanged in this case.

Lemma 2.1 and corollary 2.1 provide an estimate for $\sigma_{\epsilon,n}(t)$:

$$\begin{aligned}\sup_{t\in[0,\infty)} |\sigma_{\epsilon,n}(t)| \;&\leq\; |\alpha_n| \sup_{t\in[0,\infty)} |\rho_{\epsilon,n}(t)| \;+\; |\beta_n| \sup_{t\in[0,\infty)} |\varrho_{\epsilon,n}(t)| \\ &\leq\; 2|\alpha_n| \;+\; \tau\epsilon^2|\beta_n| \;.\end{aligned}$$

Thus we get for the second addend S_2:

$$\begin{aligned} S_2 &\leq \sum_{n>N_\epsilon} \sup_{t\in[0,\infty)} |\sigma_{\epsilon,n}(t)| \;+\; \sum_{n>N_\epsilon} \sup_{t\in[0,\infty)} |\sigma_n(t)| \\ &\leq \sum_{n>N_\epsilon} \big(2|\alpha_n| + \tau\epsilon^2|\beta_n|\big) \;+\; \sum_{n>N_\epsilon} |\alpha_n| \sup_{t\in[0,\infty)} \mathrm{e}^{-\lambda_n t} \\ &\leq\; 3 \sum_{n>N_\epsilon} |\alpha_n| \;+\; B\tau\epsilon^2 \;. \end{aligned}$$

In order to show that the tail of the sum over the $|\alpha_n|$'s can be estimated by a constant times ϵ^2 we exploit the hypothesis (2.66) once more. Recall that according to (2.32) $\lambda_n^{-1} = \frac{\omega L^2}{n^2\pi^2}$ with $\omega = \tau^{-1}$.

$$\begin{aligned} \sum_{n>N_\epsilon} |\alpha_n| &= \sum_{n>N_\epsilon} \lambda_n^{-1}\lambda_n|\alpha_n| \\ &<\; \lambda_{(N_\epsilon+1)}^{-1} \sum_{n>N_\epsilon} |\alpha_n|\lambda_n \\ &<\; \frac{\omega L^2}{\pi^2(N_\epsilon+1)^2} A_\lambda \\ &<\; \frac{\omega L^2}{\pi^2}\left(\frac{1}{3}\frac{\omega L}{\pi\epsilon}\right)^{-2} A_\lambda \;=\; 9A_\lambda\tau\,\epsilon^2 \;. \end{aligned} \tag{2.67}$$

Observe that the definition of N_ϵ entails: $N_\epsilon \leq \frac{\omega L}{3\pi\epsilon} < N_\epsilon + 1$.

Combining the estimate of S_1 and S_2 we conclude

$$\sup_{t\in[0,\infty)} \|u_\epsilon(t,\cdot) - u(t,\cdot)\|_\infty < \underbrace{\sqrt{2/L}\big(41A_\lambda + 5B\big)\tau}_{=:C}\,\epsilon^2 \;.$$

■

Remark 2.6. Note that condition (2.66) could be formulated in a slightly weaker form[23] for the price of staying in the $\mathcal{L}^2$-setting, i.e. working in theorem 2.4 with the $\|\cdot\|_2$-norm instead of the $\|\cdot\|_\infty$-norm.
As the square-summability of $|\alpha_n|\lambda_n \propto n^2|\alpha_n|$ implies the summability of $n|\alpha_n|$ (Hölder inequality + square-summability of the harmonic sequence $\frac{1}{n}$), we conclude that g is at least a $\mathcal{C}^1$-function even under the alleviated hypothesis. Therefore we have preferred to switch directly into the setting of continuous functions.

[23] No additional condition for the β_n's and $|\alpha_n|\lambda_n$ to be square-summable, i.e. $\sum_{n\geq 1} |\alpha_n|^2\lambda_n^2 < \infty$.

Remark 2.7. The question might arise, why the estimate in theorem 2.4 needs stronger regularity assumptions than the preceding convergence results. Heuristically, we can understand this by the following argument:
Formally identifying[24] the solution u of the limit problem with $u_0 = u_\epsilon|_{\epsilon=0}$, we may ask for the behavior of the mapping $\epsilon \mapsto u_\epsilon$ at the origin $\epsilon = 0$. Now an estimate of the form $\|u_\epsilon - u_0\| < C\epsilon$ is equivalent with saying, that the mapping is Lipschitz continuous at the origin, what is actually quite close[25] to being differentiable. In fact, Lipschitz continuity is easily checked if differentiability is provided. As ϵ (or better ϵ^2) mostly appears in combination with the time (see t/ϵ^2 occurring in (2.48) and (2.49)), differentiability with respect to ϵ requires differentiability with respect to t. Via the differential equation, time derivatives are related to spatial derivatives. Sending t to zero, the required differentiability of the solution with respect to the time translates into spatial smoothness requirements for the initial condition, resulting thus in the assumption concerning the initial Fourier coefficients.
Let us point out, that the intrinsic parameter is not ϵ but ϵ^2. Therefore we obtain quadratic convergence with respect to ϵ, which is – properly said – linear convergence with respect to ϵ^2.

Finally, let us throw a glance at the convergence behavior of the time derivative $\partial_t u_\epsilon$. Under the hypothesis that the Fourier series

$$\partial_t u_\epsilon = \sum_{n\geq 1} \dot{\sigma}_{\epsilon,n} s_n$$

converges in the $\mathcal{L}^2$-sense and defines then $\partial_t u_\epsilon$, we can formulate a uniform convergence result in analogy to corollary 2.3.

Corollary 2.4. *Let $\big(|\alpha_n|\lambda_n + 2|\beta_n|\big)_{n\in\mathbb{N}}$ be square-summable[26]. Then $\partial_t u_\epsilon$ converges to $\partial_t u$ in $\mathcal{C}_b\big([\theta,\infty), \mathcal{L}^2(0,L)\big)$ for any $\theta > 0$ if ϵ tends to zero.*

Proof: The uniform convergence of the time-differentiated Fourier coefficient functions (proposition 2.4) can be transferred to u_ϵ similarly to the proof of theorem 2.3. Notice, that on account of the hypothesis and due to lemma 2.4 there is an analogon to lemma 2.3 for $\dot{\sigma}_\epsilon$. ∎

In order to derive a convergence rate for $\partial_t u_\epsilon$ we need roughly speaking $\mathcal{C}^2$-regularity: the existence of the second derivative is consumed to get a sort of Lipschitz estimate for $\partial_t u_\epsilon$ with respect to ϵ.

Assumption 2.3. For the rest of this section we suppose

$$A_{\lambda^2} := \sum_{n\geq 1} |\alpha_n|\lambda_n^2 < \infty \qquad B_\lambda := \sum_{n\geq 1} |\beta_n|\lambda_n < \infty\ , \tag{2.68}$$

which tightens condition (2.66).

[24]This is justified by theorem 2.3 and corollary 2.3.
[25]Lipschitz continuity implies differentiability almost everywhere (Rademacher's theorem).
[26]Equivalently, one can separately require $(|\alpha_n|\lambda_n)_{n\in\mathbb{N}}$ and $(|\beta_n|)_{n\in\mathbb{N}}$ to be square-summable.

Theorem 2.5. *Let the Fourier coefficients of the initial data satisfy condition* (2.68). *Then there exists a constant $K > 0$ depending only on the initial data, L, θ and τ such that*

$$\sup_{t\in[\theta,\infty)} \|\partial_t u_\epsilon(t,\cdot) - \partial_t u(t,\cdot)\|_\infty < K\epsilon^2 \qquad \textit{for} \quad 0 < \epsilon < \eta_1.$$

Proof: As in the proof of theorem (2.4) we decompose the Fourier series in two parts. The first one is handled by proposition 2.4 providing uniform convergence of the temporally differentiated Fourier coefficient functions. The second part, which corresponds to the tail of the Fourier series, vanishes also with ϵ^2. Notice that condition (2.68) is required for the estimate of both parts.

$$\begin{aligned}
&\sup_{t\in[\theta,\infty)} \|\partial_t u_\epsilon(t,\cdot) - \partial_t u(t,\cdot)\|_\infty \\
&\qquad \le \sqrt{\tfrac{2}{L}} \underbrace{\sum_{n=1}^{N_\epsilon} \sup_{t\in[\theta,\infty)} \left|\dot\sigma_{\epsilon,n}(t) - \dot\sigma_n(t)\right|}_{=:S_1} + \sqrt{\tfrac{2}{L}} \underbrace{\sum_{n\ge N_\epsilon} \sup_{t\in[\theta,\infty)} \left|\dot\sigma_{\epsilon,n}(t) - \dot\sigma_n(t)\right|}_{=:S_2}
\end{aligned}$$

For the first part we obtain:

$$\begin{aligned}
S_1 &\le \sum_{n\ge N_\epsilon} K_{\theta,n}\epsilon^2 = \sum_{n\ge N_\epsilon} 2\theta^{-1}\Big(|\alpha_n|\lambda_n + |\beta_n|\Big)\tau\epsilon^2 + \Big(35|\alpha_n|\lambda_n^2 + 12|\beta_n|\lambda_n\Big)\tau\epsilon^2 \\
&\le \Big(2\theta^{-1}A_\lambda + 2\theta^{-1}B + 35A_{\lambda^2} + 12B_\lambda\Big)\tau\epsilon^2\,.
\end{aligned}$$

Applying the triangle inequality to each term in the second sum S_2 and employing lemma 2.4 we get:

$$S_2 \;\le\; \sum_{n>N_\epsilon} \big(|\alpha_n|\lambda_n + 2|\beta_n|\big) + |\alpha_n|\lambda_n\,.$$

Multiplying each addend by $1 = \lambda_n^{-1}\lambda_n$ and estimating λ_n^{-1} from above by $\lambda_{(N_\epsilon+1)}^{-1}$, we can utilize the same trick as in (2.67).

$$S_2 \le\; 2\lambda_{(N_\epsilon+1)}^{-1} \sum_{n>N_\epsilon} |\alpha_n|\lambda_n^2 + |\beta_n|\lambda_n \;\le\; 18(A_{\lambda^2} + B_\lambda)\tau\epsilon^2$$

Hence the assertion follows with $K := \sqrt{\frac{2}{L}}\Big(2\theta^{-1}A_\lambda + 2\theta^{-1}B + 53A_{\lambda^2} + 30B_\lambda\Big)\tau$. ■

Appendix

Here we collect some prerequisites that have been used throughout this section.

First, let us prove the fact that exponential decay is "stronger" than polynomial and especially linear growth, which has been exploited several times.

Lemma 2.8. *Consider for fixed $n \in \mathbb{N}$ the function $f(x) = x^n e^{-1}$ on $[0, \infty)$. Then:*

$$\lim_{x\to\infty} x^n \mathrm{e}^{-x} = 0 \quad \textit{and} \quad f(x) \leq n^n \mathrm{e}^{-n} \quad \textit{for } x \in \mathbb{R}_0^+.$$

Proof: Obviously $f(0) = 0$. In order to compute the limit we consider the Taylor series of the exponential function, that may also serve as its definition.

$$x^n \mathrm{e}^{-x} = \frac{x^n}{\mathrm{e}^x} = \frac{x^n}{\sum\limits_{k\geq 0} \frac{1}{k!} x^k} < \frac{x^n}{\frac{1}{n!}x^n + \frac{1}{(n+1)!}x^{n+1}} = \frac{(n+1)!}{(n+1)+x} \xrightarrow{x\to\infty} 0$$

From $f(0) = \lim\limits_{x\to\infty} f(x)$ we conclude that f must attain its extrema in the interior of its domain of definition, i.e. in $(0, \infty)$. Therefore we can apply *Rolle's* theorem to ascertain the extrema. Computing the zeros of the derivative $f'(x) = x^{n-1}(n - x)\mathrm{e}^{-x}$ shows, that there is only one extremum at n. Since $f(n) = n^n \mathrm{e}^{-n} > 0$, it must be the global maximum. ■

It was already mentioned that IBVP (2.29) and (2.30) can be solved by means of Fourier series. To justify this ansatz and to prove theorem 2.1 for instance, convergence and regularity of the resulting Fourier series must be verified. For this the following propositions turn out to be helpful.

Proposition 2.5. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of continuously differentiable functions in $\mathcal{C}^1\big([a,b];\mathbb{R}\big)$. Suppose that for $n \to \infty$*

1) *f_n converges in some point $x_0 \in [a,b]$ to some value $y_0 \in \mathbb{R}$,*
2) *f_n' converges uniformly to some function g.*

Then f_n converges uniformly to a function $f \in \mathcal{C}^1\big([a,b];\mathbb{R}\big)$ with $f' = g$.

This means that under the mentioned conditions the operations of taking the limit and computing the derivative can be permuted. A proof of this theorem can be found in [56] on page 152.

Definition 2.1. *A series of functions $\sum_k f_k$ is called* normally convergent *on $\mathcal{X} \subset \mathbb{R}$, if each addend $f_k : \mathcal{X} \to \mathbb{R}$ is bounded and the series of norms is convergent: $\sum_k \|f_k\|_{\mathcal{L}^\infty(\mathcal{X})} < \infty$*

In particular a *normally* convergent series of functions is *absolutely* convergent in any point.

Proposition 2.6. *Let $\sum_k f_k$ be a* normally *convergent series. If all f_k are continuous then the limit function is continuous too.*

A proof is found in [42] on page 86.

Proposition 2.7. ***(Normal convergence implies uniform convergence)*** *Let the series $\sum_k f_k$ converge* normally *on $\mathcal{X} \subset \mathbb{R}$ to some function f. Then the series converges also* uniformly *to f.*

A proof is found in [42] on page 294.

Setting

$$S_N(t,x) := \sum_{n=1}^N \alpha_n \mathrm{e}^{-\lambda_n t} s_n(x) = \sum_{n=1}^N \alpha_n \mathrm{e}^{-\tau n^2\pi^2 t/L^2} \sin(n\pi x/L)$$

and computing the derivatives

$$\begin{aligned}
\partial_t S_N(t,x) &= \sum_{n=1}^N -\lambda_n \alpha_n \mathrm{e}^{-\lambda_n t} s_n(x) &&= -\tau \frac{\pi^2}{L^2} \sum_{n=1}^N \alpha_n n^2 \mathrm{e}^{-\tau n^2\pi^2 t/L^2} \sin(n\pi x/L) \\
\partial_x S_N(t,x) &= \sum_{n=1}^N \alpha_n \mathrm{e}^{-\lambda_n t} s_n'(x) &&= \frac{\pi}{L} \sum_{n=1}^N \alpha_n n\, \mathrm{e}^{-\tau n^2\pi^2 t/L^2} \cos(n\pi x/L) \\
\partial_x^2 S_N(t,x) &= \sum_{n=1}^N \alpha_n \mathrm{e}^{-\lambda_n t} s_n''(x) &&= -\frac{\pi^2}{L^2} \sum_{n=1}^N \alpha_n n^2 \mathrm{e}^{-\tau n^2\pi^2 t/L^2} \sin(n\pi x/L)
\end{aligned}$$

it can be checked that all sums define $\mathcal{C}^\infty$-functions in the limit $N \to \infty$ for $t > 0$ even if $(\alpha_n)_n$ is only square-summable, i.e. $g \in \mathcal{L}^2(0, L)$. Since the limit $N \to \infty$ can be interchanged with the spatial and temporal derivative, $\lim_{N\to\infty} S_N$ represents the solution of (2.30).

This nice behavior is mainly due to the circumstance that the sequence $(n^p \mathrm{e}^{-nt})_n$ is summable for arbitrary $p \in \mathbb{N}$ if $t > 0$.

$$\sum_{n\in\mathbb{N}} n^p \mathrm{e}^{-nt} = (-1)^p \frac{\mathrm{d}^p}{\mathrm{d}x^p} \frac{\mathrm{e}^{-t}}{1 - \mathrm{e}^{-t}} < \infty$$

In principle, the perturbed IBVP (2.29) is solved similarly; however the initial conditions must fulfill stronger regularity assumptions in order to obtain a smooth function by the Fourier series. Otherwise the Fourier series must be interpreted in a weak or distributional sense.

Remark 2.8. If $g \in \mathcal{L}^2(0, L)$ or equivalently $(\alpha_n)_n \in \ell^2$, the time derivative $\lim_{N\to\infty} \partial_t S_N$ does not represent an $\mathcal{L}^2$-function for $t = 0$ since generally $(n^2\alpha_n)_n \notin \ell^2$. Even the condition $g \in \mathcal{H}^1_0(0, L)$ entailing $(\alpha_n)_n \in \ell^2$ and $(n\alpha_n)_n \in \ell^2$ does not suffice to guarantee $(n^2\alpha_n)_n \in \ell^2$ whence $\lim_{N\to\infty} \partial_t S_N|_{t=0}$ is no $\mathcal{L}^2$-function. However, if $u \in \mathcal{H}^1_0(0, L)$ denotes a testfunction with Fourier coefficients $(\beta_n)_n \in \ell^2$, then the Cauchy-Schwarz inequality yields $(n^2\alpha_n\beta_n)_n \in \ell^1$ as $(n\alpha_n)_n \in \ell^2$ and $(n\beta_n)_n \in \ell^2$. So the formal Fourier series for $\lim_{N\to\infty} \partial_t S_N|_{t=0}$ can be used to define a functional on $\mathcal{H}^1_0(0, L)$. But this means that it should be interpreted as an element of the dual space $\mathcal{H}^{-1}(0, L)$. (Compare with [17] page 352, Sobolev space setting for parabolic equations.)

The following figures illustrate how the roots $\psi_{\epsilon,n}$ and $\chi_{\epsilon,n}$ of the characteristic polynomial (2.39) pertaining to (2.38) depend on ϵ and $\lambda_n \propto n^2$. While ϵ and n are sufficiently small, $\psi_{n,\epsilon}$ and $\chi_{n,\epsilon}$ are real. If n is increased, $W_{\epsilon,n}$ finally becomes imaginary such that the roots are complex-valued. The asymptotic behavior for $n \to \infty$

$$|\psi_{\epsilon,n}| \in \mathrm{O}(n), \qquad |\psi_{\epsilon,n}| \in \mathrm{O}(n)$$

can be recognized.

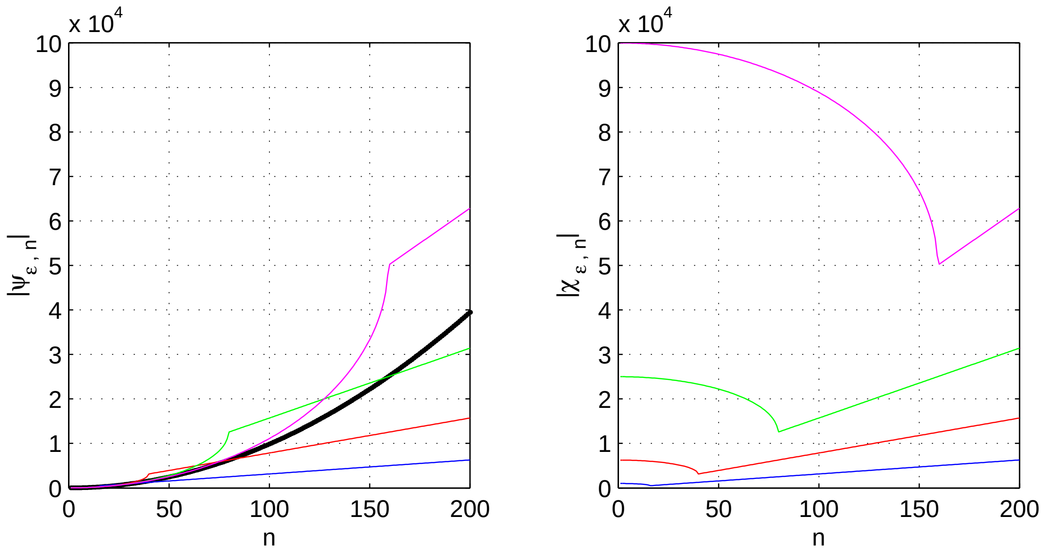


Figure 2.2: The diagrams show the moduli of $\psi_{\epsilon,n}$ (left) and $\chi_{\epsilon,n}$ (right) plotted versus n for different values of $\epsilon \in \{0.01,\ 0.02,\ 0.04\ ,0.1\}$ with $\tau = 0.1$ and $L = 1$. At the kink $\psi_{\epsilon,n}$ and $\chi_{\epsilon,n}$ become complex-valued; their asymptotically linear dependence on $n \to \infty$ is clearly visible. The thick line in the left figure represents the eigenvalue $\lambda_n = \tau(\frac{n\pi^2}{L})$ as function of n.

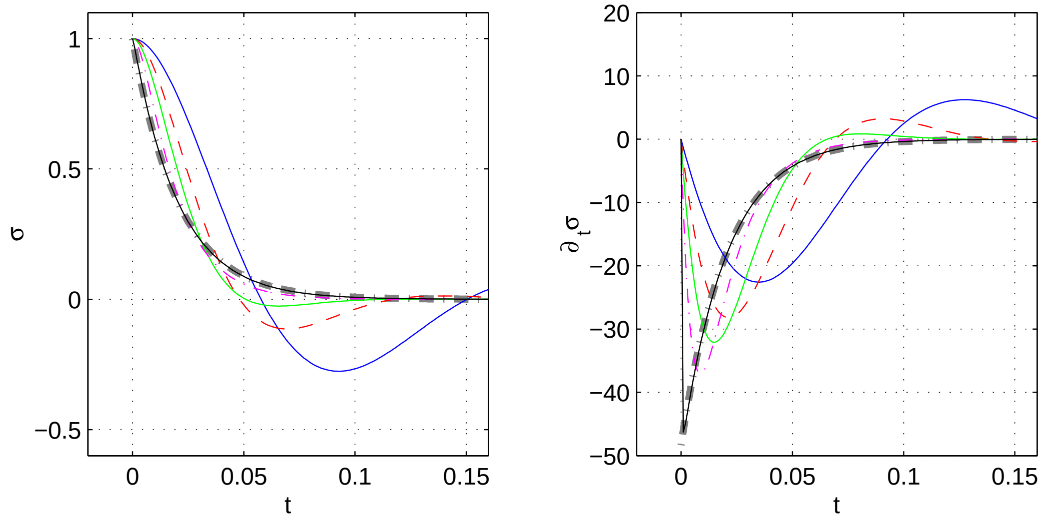


Figure 2.3: Time evolution of a Fourier coefficient function (left) and its temporal derivative (right). The thick, dashed line represents σ_n, while the thinner lines correspond to $\sigma_{\epsilon,n}$ (and their time derivatives respectively) for different values of ϵ: 0.6(blue-solid), 0.4(red-dashed) 0.3(green-solid), 0.2(magenta-dashdot) and 0.05(black-solid). For the two largest values of ϵ one can recognize an oscillating behavior, since $W_{\epsilon,n}$ is still imaginary. The left figure suggests, that $\sigma_{\epsilon,n}$ converges *uniformly* to σ_n. In contrast, the time derivative displays an *initial layer*, that is generated by the incompatibility of the second initial condition $\beta_n \neq \dot{\sigma}_n(0) = -\alpha_n\lambda_n$. The initial layer becomes the narrower the smaller ϵ is chosen, its width is of the magnitude $\mathrm{O}(\epsilon^2)$. Hence, the time derivative $\dot{\sigma}_{\epsilon,n}$ converges uniformly outside any neighborhood of the time origin. Choice of other parameters: $n = 7, \tau = 0.1$.

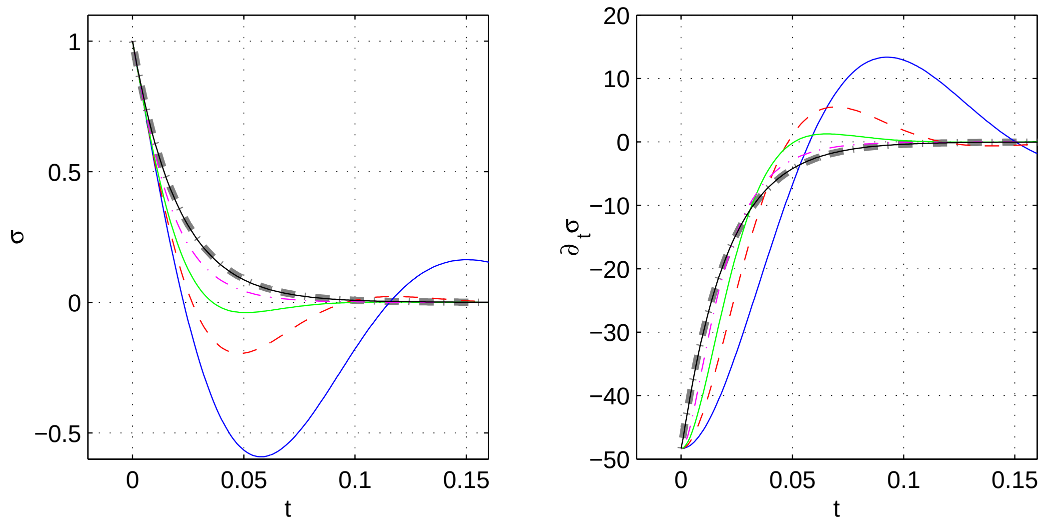


Figure 2.4: Unlike figure 2.3, the initial layer has been “switched off” by setting $\beta_n = -\alpha_n\lambda_n = \dot{\sigma}_n(0)$. In this case $\dot{\sigma}_{\epsilon,n} \equiv \partial_t\sigma_{\epsilon,n}$ converges uniformly on the entire non-negative real axis.

2.3 Two-scale expansion and resolution of the initial layer

Hitherto, our analysis of the singular IBVP (2.29) is based on the classical tool of Fourier series. Even though this technique offers the advantage of providing quasi-explicit solutions, analytic representations may be sometimes rather cumbersome to extract qualitative properties.
In this section we attack the singular IBVP by asymptotic expansions with respect to the perturbation parameter ϵ. As mentioned at the beginning of the previous section, this approach works independently of the possibility to apply Fourier series and is particularly suitable for non-linear problems. In contrast to the Fourier approach, the asymptotic expansion we will choose below has the virtue to offer much insight into the "internal dynamics" of the perturbed equation. The temporal evolution of the *deviation* $u_\epsilon - u$ is in the limelight of discussion to understand the convergence in detail. By means of the asymptotic expansion it is possible to derive equations telling us, where and why the error is generated and by what sources it is fed.

Although the perturbed problem (2.29) could be tackled directly by asymptotic expansions, we prefer to profit from the Fourier expansion in so far as it reduces the PDE problem to a bunch of structurally equal ODE problems. Since the qualitative behavior of the singular limit is fully contained in each of the Fourier coefficient functions, it is enough to concentrate here on a single Fourier mode. Hence we drop from now on the index n. Recall that every Fourier coefficient function of the perturbed problem solves an IVP of the type

$$\left.\begin{array}{ll} \mathrm{EQ:} & \epsilon^2\tau\ddot{\sigma}_\epsilon + \dot{\sigma}_\epsilon + \lambda\sigma_\epsilon = 0 \\ \mathrm{IC:} & \sigma_\epsilon(0) = \alpha \quad \wedge \quad \dot{\sigma}_\epsilon(0) = \beta \end{array}\right\} \tag{2.69}$$

with the limit problem

$$\left.\begin{array}{ll} \mathrm{EQ:} & \dot{\sigma} + \lambda\sigma = 0 \\ \mathrm{IC:} & \sigma(0) = \alpha \end{array}\right\}. \tag{2.70}$$

In order to analyze how the solution of the IVP (2.69) behaves for $\epsilon \downarrow 0$, we attempt to represent σ_ϵ by an asymptotic expansion. The simplest ansatz is a *regular expansion* corresponding to a (truncated) power series in ϵ

$$\sigma_\epsilon(t) = \varsigma^{(0)}(t) + \epsilon^2\varsigma^{(2)}(t) + \epsilon^4\varsigma^{(4)}(t) + ... \tag{2.71}$$

with the asymptotic order functions $\varsigma^{(2k)},\ \ k = 0, 1, 2, ...$, as time depending coefficients. Plugging (2.71) into (2.69) and equating terms of the same order in ϵ, yields the subsequent IVPs for the order functions:

$$\begin{array}{llrcl} \epsilon^0: & \varsigma^{(0)}(0) = \alpha & \dot{\varsigma}^{(0)} + \lambda\varsigma^{(0)} & = & 0 \\ \epsilon^2: & \varsigma^{(2)}(0) = 0 & \dot{\varsigma}^{(2)} + \lambda\varsigma^{(2)} & = & -\tau\ddot{\varsigma}^{(0)} \\ \epsilon^{2k}: & \varsigma^{(2k)}(0) = 0 & \dot{\varsigma}^{(2k)} + \lambda\varsigma^{(2k)} & = & -\tau\ddot{\varsigma}^{(2k-2)} \end{array} \tag{2.72}$$

Obviously the leading order function $\varsigma^{(0)}$ satisfies (2.70), that is why we obtain $\varsigma^{(0)} = \sigma$. This confirms the convergence result of proposition 2.3, namely:

$$\sigma_\epsilon = \varsigma^{(0)} + \epsilon^2 \varsigma^{(2)} + ... = \sigma + \mathrm{O}(\epsilon^2) \; . \tag{2.73}$$

Observe, that the second initial condition $\dot{\sigma}_\epsilon(0) = \beta$ has been completely ignored to derive (2.72). This condition cannot be taken into account, since otherwise the IVPs in (2.72) would forfeit their well-posedness. This indicates that the regular expansion (2.71) is unable to reflect the full behavior of σ_ϵ.

Motivation & computation of a general two-scale expansion

To get an idea how to improve ansatz (2.71) we take a look at the right plot of figure 2.3, which compares the evolution of $\dot{\sigma}_\epsilon$ and $\dot{\sigma}$. Especially for small values of ϵ, the curves representing $\dot{\sigma}_\epsilon$ appear as superpositions of two processes.

Despite of being not compatibly initialized, $\dot{\sigma}_\epsilon$ tries to attain the initial value of $\dot{\sigma}$. This process, which dominates the initial phase, corresponds to a quick relaxation, starting from the incompatible initial value 0 and tending the more rapidly towards $\dot{\sigma}(0)$ the smaller ϵ is chosen. Simultaneously, another process is triggered that seems to counteract the first one. This second process mimics the diffusive decay of the limit function $\dot{\sigma}$. In contrast to the relaxation, the diffusion evolves relatively slowly and becomes clearly visible only after the relaxation has lost its prevalence.

While the curves representing $\dot{\sigma}_\epsilon$ converge towards $\dot{\sigma}$ outside some vicinity of 0, they behave differently near 0. It seems as if they approximate a vertical line that cannot be described as graph of a single-valued function. This means that the limit function becomes discontinuous, jumping instantaneously from the actual incompatible initial value 0 to $\dot{\sigma}(0)$.

This indicates that the relaxation process gets accelerated as ϵ tends to zero. To visualize the process better, $\dot{\sigma}_\epsilon$ is plotted versus the scaled time t/ϵ^2 in figure 2.5 c) where a usual convergence behavior can be observed. So we are led to the conclusion, that apparently $\dot{\sigma}_\epsilon$ exhibits two time-scales similar to the function $t \mapsto t^2 + \sin(t/\epsilon^2)$ representing a parabola modulated by a sinusoidal oscillation. If we look at the long-term behavior of this function, then we recover the parabola. Zooming into a time interval of length $\epsilon \ll 1$, the quadratic growth of the parabola becomes hardly perceptible, whereas the oscillations get very striking.

Inspired by these observations we are inclined to assume that $\dot{\sigma}_\epsilon$ and hence σ_ϵ as well depend on the time in a two-fold manner such that they are described by a two-scale function[27]. Concretely we suppose that the variable t occurs either alone (slow time) or in combination with the factor ϵ^{-2}, i.e. t/ϵ^2 (fast time[28]). Therefore

[27]Generally, the definition of *multi-scale dependence* remains somewhat vague, because this notion combines some degree of arbitrariness. For instance, the identity function $t \mapsto t$ can be written quite artificially in the form $t = \frac{1}{5}t + \frac{4}{5}\epsilon^2(t/\epsilon^2)$, which looks like a veritable multiple-scale function.

[28]If ϵ is a small parameter less than 1, then $t = \epsilon^2$ implies $t/\epsilon^2 = 1 > \epsilon^2$. This means that t/ϵ^2 grows much faster than t.

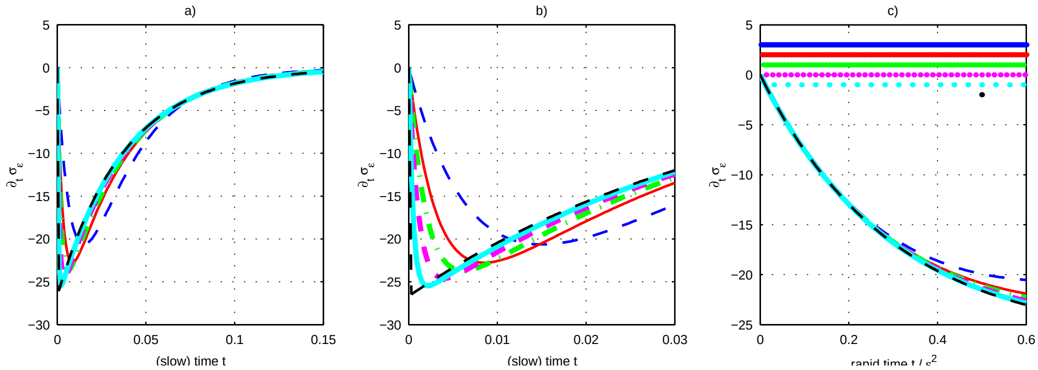


Figure 2.5: Inspection of the initial layer: Subplot a) shows the long-term behavior (diffusive decay) of the derivative $\dot{\sigma}_\epsilon$ as a function of time for different values of ϵ. It can be seen clearly, that the curves coincide almost outside the range of the initial layer manifesting the convergence. Subplot b) yields a close-up of the so called initial layer, that represents the short-term relaxation process to the initial value of the limit function $\dot{\sigma}$. In subplot c) $\dot{\sigma}_\epsilon$ is plotted versus the scaled time t/ϵ^2 being referred to as the fast time. Comparing c) with a), the curves reveal now the converse behavior, because they agree near 0 and diverge as the rapid time increases. The fat dots indicate the time dilation depending on ϵ. For instance, the time dilation factor is 16 times larger for the smallest value of ϵ (dashed black) than for the second smallest one (solid cyan). Parameter setting: $\tau = 0.3, \lambda = 9\pi^2\tau$, $\epsilon = 0.15$(dashed blue), 0.1(solid red), 0.08(dash-dotted green), 0.06(dashed magenta), 0.04(solid cyan), 0.01(dashed black).

we introduce two new variables: t/ϵ^2 is replaced by r – the *rapid* time variable – whereas t alone is substituted by s – the *slow* time variable. Although this setting decouples both of the time scales in an unnatural[29] way, as r and s are independent of each other while t/ϵ^2 and t are not, it is reasonable to do so in order to simplify the analysis of the problem.

$$\sigma_\epsilon(t) = \sigma^{(0)}(t/\epsilon^2, t) + \epsilon^2\sigma^{(2)}(t/\epsilon^2, t) + \epsilon^4\sigma^{(4)}(t/\epsilon^2, t)\dots \tag{2.74}$$

As usual, the $\sigma^{(2k)}$'s are referred to as *asymptotic order functions*. Expansion (2.74) makes sense only if we can apply the *comparison of coefficients*[30]. This requires the $\sigma^{(2k)}$'s to remain bounded if ϵ tends to 0. Therefore we postulate the existence of a constant $C > 0$ such that

$$\forall\, t \in [0, T], \quad \forall\, \epsilon \in (0, 1): \qquad |\sigma^{(2k)}(t/\epsilon^2, t)| < C^{2k}. \tag{2.75}$$

This *condition of boundedness*[31] turns out to be crucial, when we extract the equations governing the evolution of the order functions.

[29] From a mathematical point of view this is just a change of coordinates from the t-ϵ plane to the r-s plane.

[30] Two (finite or infinite) expansions of the type (2.74) are equal, if and only if all corresponding asymptotic order functions are equal. This can be seen successively by sending ϵ to 0. Thanks to the *condition of boundedness* we obtain the equality of the zeroth order. Subtracting this, dividing by ϵ^2 and sending ϵ once more to zero we get the equality of the second order and so forth.

[31] Observe that this condition is less restrictive than uniform boundedness with respect to k. If

By computing the first and second derivative of $\sigma^{(2k)}$ with respect to the time t

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}\sigma^{(2k)}(t/\epsilon^2,t) &= \left(\epsilon^{-2}\partial_r+\partial_s\right)\sigma^{(2k)}(t/\epsilon^2,t)\\ \tfrac{\mathrm{d}^2}{\mathrm{d}t^2}\sigma^{(2k)}(t/\epsilon^2,t) &= \left(\epsilon^{-4}\partial_r^2+\epsilon^{-2}2\partial_r\partial_s+\partial_s^2\right)\sigma^{(2k)}(t/\epsilon^2,t)\end{aligned}$$

we recognize that differentiation will mix the orders of ansatz (2.74): hence the derivative of an order function is not equal to the order function of the derivative $\dot\sigma_\epsilon$ as in the case of the regular expansion. Of course, this is due to the fact, that the asymptotic order functions depend on ϵ, if they are written as functions of t. Plugging (2.74) into (2.69), collecting terms of equal order in ϵ and performing the standard comparison of coefficients we obtain the following equations

$$\begin{aligned}\epsilon^{-2}:&\ \partial_r^2\sigma^{(0)}+\omega\partial_r\sigma^{(0)} = 0\\ \epsilon^{0}:&\ \partial_r^2\sigma^{(2)}+\omega\partial_r\sigma^{(2)} = -2\partial_r\partial_s\sigma^{(0)}-\omega\partial_s\sigma^{(0)}-\omega\lambda\sigma^{(0)}\\ \epsilon^{2k}:&\ \partial_r^2\sigma^{(2k+2)}+\omega\partial_r\sigma^{(2k+2)} = -2\partial_r\partial_s\sigma^{(2k)}-\omega\partial_r\sigma^{(2k)}-\omega\lambda\sigma^{(2k)}-\partial_s^2\sigma^{(2k-2)}\end{aligned}\tag{2.76}$$

for $k\in\mathbb{N}$. We interpret each equation as evolution equation for the highest occurring asymptotic order function. Then (2.76) constitutes a sequence of equations, which are successively solvable – starting with the lowest order function $\sigma^{(0)}$ appearing as source for $\sigma^{(2)}$ and so forth.
The second order ODEs[32] are accompanied by initial conditions, that are derived by plugging (2.74) into the initial conditions of (2.69) and performing once more the comparison of coefficients.

$$\begin{aligned}\sigma^{(0)}(0,0)&=\alpha\ , & \partial_r\sigma^{(0)}(0,0)&= 0\\ \sigma^{(2)}(0,0)&=0\ , & \partial_r\sigma^{(2)}(0,0)&=\beta-\partial_s\sigma^{(0)}(0,0)\\ \sigma^{(2k+2)}(0,0)&=0\ , & \partial_r\sigma^{(2k+2)}(0,0)&=-\partial_s\sigma^{(2k)}(0,0)\end{aligned}\tag{2.77}$$

Formally $\sigma^{(2k)}, k\in\mathbb{N}_0$, is calculated by transforming the scalar second-order differential equation into a first order system, where *Duhamel's formula* (variation of constants) is applied. So we set

$$\mathbf{z}^{(2k)}(r,s)\equiv\begin{pmatrix}z_1^{(2k)}(r,s)\\ z_2^{(2k)}(r,s)\end{pmatrix}:=\begin{pmatrix}\sigma^{(2k)}(r,s)\\ \partial_r\sigma^{(2k)}(r,s)\end{pmatrix}.\tag{2.78}$$

Then the corresponding evolution system becomes

$$\left.\begin{aligned}\partial_r z_1^{(2k)} &= z_2^{(2k)}\\ \partial_r z_2^{(2k)} &= -\omega z_2^{(2k)}+Q^{(2k)}\end{aligned}\right\}\Leftrightarrow\frac{\partial}{\partial r}\begin{pmatrix}z_1^{(2k)}\\ z_2^{(2k)}\end{pmatrix}=\underbrace{\begin{pmatrix}0 & 1\\ 0 & -\omega\end{pmatrix}}_{=:A}\begin{pmatrix}z_1^{(2k)}\\ z_2^{(2k)}\end{pmatrix}+\begin{pmatrix}0\\ Q^{(2k)}\end{pmatrix}.$$

the expansion has infinitely many terms then it is convergent for $\epsilon<1/C$, as it is dominated by the geometric series belonging to $C\epsilon<1$.

[32] It is remarkable, that the resulting system (2.76) of differential equations needs not to be considered as a system of PDEs (in the variable r,s), whose solution would be much more complicated.

Here $Q^{(2k)}$ represents the associated source term occurring in the equations of (2.76). In matrix notation this system is equivalently expressed by

$$\partial_r \mathbf{z}^{(2k)} = A\mathbf{z}^{(2k)} + Q^{(2k)} \binom{0}{1} \qquad k \in \mathbb{N}_0. \tag{2.79}$$

Duhamel's formula reads now

$$\mathbf{z}^{(2k)}(r,s) = \mathrm{e}^{rA}\mathbf{y}^{(2k)}(s) + \int_0^r Q^{(2k)}(\vartheta, s)\, \mathrm{e}^{(r-\vartheta)A} \binom{0}{1} \,\mathrm{d}\vartheta\,, \tag{2.80}$$

where the exponential matrix is explicitly given by

$$\mathrm{e}^{rA} = \begin{pmatrix} 1 & \tau - \tau\mathrm{e}^{-\omega r} \\ 0 & \mathrm{e}^{-\omega r} \end{pmatrix} .$$

With $\mathbf{y}^{(2k)}$ we denote the initial condition depending generally on the slow time variable s that plays the subordinate role of a mere parameter in the differential equation (2.79). Let us finally write down (2.76) in the new notation to see, how the sources look like:

$$\begin{aligned}
\partial_s \mathbf{z}^{(0)} \quad - \quad A\mathbf{z}^{(0)} \quad &= \quad 0\,, \\
\partial_s \mathbf{z}^{(2)} \quad - \quad A\mathbf{z}^{(2)} \quad &= -\big[2\partial_s z_2^{(0)} + \omega\partial_s z_1^{(0)} + \omega\lambda z_1^{(0)}\big] \tbinom{0}{1}\,, \\
\partial_s \mathbf{z}^{(2k+2)} - A\mathbf{z}^{(2k+2)} &= -\big[2\partial_s z_2^{(2k)} + \omega\partial_s z_1^{(2k)} + \omega\lambda z_1^{(2k)} + \partial_s^2 z_1^{(2k-2)}\big] \tbinom{0}{1}\,.
\end{aligned} \tag{2.81}$$

These evolution equations are supplemented by the initial conditions

$$\mathbf{z}^{(0)}(0,0) = \tbinom{\alpha}{0}, \quad \mathbf{z}^{(2)}(0,0) = \tbinom{0}{\beta - \partial_s z_1^{(0)}(0,0)} \text{ and } \mathbf{z}^{(2k+2)}(0,0) = \tbinom{0}{-\partial_s z_1^{(2k+2)}(0,0)}$$

becoming in terms of $\mathbf{y}^{(0)}, \mathbf{y}^{(2)}$ and $\mathbf{y}^{(2k+1)}$

$$\mathbf{y}^{(0)}(0) = \tbinom{\alpha}{0}, \quad \mathbf{y}^{(2)}(0) = \tbinom{0}{\beta - \dot{y}_1^{(0)}(0)} \quad \text{and} \quad \mathbf{y}^{(2k+2)}(0) = \tbinom{0}{-\dot{y}_1^{(2k)}(0)}, \tag{2.82}$$

where dots indicate derivatives with respect to s.

Computation of $\mathbf{z}^{(0)}(r,s)$. Since the equation for $\mathbf{z}^{(0)}$ is homogeneous ($Q^{(0)} \equiv 0$), we obtain from (2.80)

$$\mathbf{z}^{(0)}(r,s) = \mathrm{e}^{rA}\mathbf{y}^{(0)}(s) = \begin{pmatrix} y_1^{(0)}(s) + \tau y_2^{(0)}(s) - \tau\mathrm{e}^{-\omega r} y_2^{(0)}(s) \\ \mathrm{e}^{-\omega r} y_2^{(0)}(s) \end{pmatrix} . \tag{2.83}$$

In order to determine $\mathbf{y}^{(0)}(s)$ we consider the solution formula for $\mathbf{z}^{(2)}$:

$$\mathbf{z}^{(2)}(r,s) = \mathrm{e}^{rA}\mathbf{y}^{(2)}(s) - \underbrace{\int_0^r \Big[2\partial_s z_2^{(0)} + \omega\partial_s z_1^{(0)} + \omega\lambda z_1^{(0)}\Big](\vartheta, s)\, \mathrm{e}^{(r-\vartheta)A} \tbinom{0}{1} \,\mathrm{d}\vartheta}_{=:J^{(0)}}\,.$$

Let us compute the integral $J^{(0)}$ more explicitly. For this we replace $z_1^{(0)}$ and $z_2^{(0)}$ by their representation given in (2.83). Furthermore $J^{(0)}$ is decomposed into a sum of four integrals, whose specific dependence on the rapid time r is indicated at the right.

$$\begin{aligned}
J^{(0)} = \quad & \binom{\tau}{0} \int_0^r \left[\omega \dot{y}_1^{(0)} + \dot{y}_2^{(0)} + \omega\lambda y_1^{(0)} + \lambda y_2^{(0)}\right] \mathrm{d}\vartheta && \mid \propto r \\
+ & \binom{\tau}{0} \int_0^r \left[\dot{y}_2^{(0)} - \lambda y_2^{(0)}\right] \mathrm{e}^{-\omega\vartheta} \, \mathrm{d}\vartheta && \mid \propto \tau(1-\mathrm{e}^{-\omega r}) \\
+ & \binom{-\tau}{1} \int_0^r \left[\omega \dot{y}_1^{(0)} + \dot{y}_2^{(0)} + \omega\lambda y_1^{(0)} + \lambda y_2^{(0)}\right] \mathrm{e}^{\omega(\vartheta - r)} \, \mathrm{d}\vartheta && \mid \propto \tau(1-\mathrm{e}^{-\omega r}) \\
+ & \binom{-\tau}{1} \int_0^r \left[\dot{y}_2^{(0)} - \lambda y_2^{(0)}\right] \mathrm{e}^{-\omega r} \, \mathrm{d}\vartheta && \mid \propto r\mathrm{e}^{-\omega r}
\end{aligned}$$

Since $\mathbf{y}^{(0)}$ only depends on the slow time s, the rectangular braces are independent of the integration variable ϑ. In particular, the integrands of the first and fourth integral are independent of ϑ.
In the case of the first integral, the integration yields a term of the form $rf(s)$ with f being the s-dependent integrand $\omega \dot{y}_1^{(0)} + \dot{y}_2^{(0)} + \omega\lambda y_1^{(0)} + \lambda y_2^{(0)}$. After resubstitution of the time, $r \to t/\epsilon^2$ and $s \to t$, we get $tf(t)/\epsilon^2$. For $f \not\equiv 0$ this term can not be uniformly bounded with respect to $\epsilon \in (0,1)$ if $t > 0$. Since this term occurs in the first component of $\mathbf{z}^{(2)}$, being equal to $\sigma^{(2)}$, the asymptotic order function $\sigma^{(2)}$ violates the condition of boundedness (2.75) if $f \not\equiv 0$. Therefore, to comply with (2.75), we must require f to vanish. This yields one of the desired equations for determining the components of $\mathbf{y}^{(0)}$

$$\omega \dot{y}_1^{(0)} + \dot{y}_2^{(0)} + \omega\lambda y_1^{(0)} + \lambda y_2^{(0)} = 0 \,. \tag{2.84}$$

Unfortunately, the ansatz does not supply the second equation to determine $y_1^{(0)}$ and $y_2^{(0)}$ in a unique way. In fact, we should not expect, that these variables come out univocally, as we have replaced one unknown $\sigma^{(0)}$ by two unknowns $y_1^{(0)}, y_2^{(0)}$.

Regarding the second integral proportional to $1 - \mathrm{e}^{-\omega r}$, it is clear, that the argument from above does not work to produce another equation. In this respect, the fourth integral looks more promising. Integration produces a term of the type $r\mathrm{e}^{-\omega r} f(s)$, where f stands now for $\dot{y}_2^{(0)} - \lambda y_2^{(0)}$. After resubstitution of the time we get $t/\epsilon^2 \mathrm{e}^{-\omega t/\epsilon^2} f(t)$. However, due to the exponential function this term is bounded for $\epsilon \in (0,1)$ and $t > 0$, although ϵ^{-2} occurs as an delusive external factor. So, none of the other integrals gives rise to a second equation.
As we dispose only of one equation for two unknowns $y_1^{(0)}, y_2^{(0)}$, we can define one of them arbitrarily (in accordance with initial conditions and regularity requirements) and compute the other by equation (2.84). In particular we can choose $y_2^{(0)}$ to satisfy

$$\dot{y}_2^{(0)} - \lambda y_2^{(0)} = 0 \;. \tag{2.85}$$

Observe that this *choice* is *a priori* not distinguished, except that it makes the integral $J^{(0)}$ vanish, what simplifies the further computations.

So we end up with the following IVPs for $y_2^{(0)}$ and $y_1^{(0)}$:

$$\left.\begin{array}{rcl} \dot{y}_2^{(0)} - \lambda y_2^{(0)} &=& 0 \\ y_2^{(0)}(0) &=& 0 \end{array}\right\} \quad \Rightarrow \quad y_2^{(0)} \equiv 0 \tag{2.86}$$

$$\left.\begin{array}{rcl} \dot{y}_1^{(0)} + \lambda y_1^{(0)} &=& -\tau \dot{y}_2^{(0)} - \tau\lambda y_2^{(0)} = 0 \\ y_1^{(0)}(0) &=& \alpha \end{array}\right\} \quad \Rightarrow \quad y_1(s) = \alpha \mathrm{e}^{-\lambda s} \tag{2.87}$$

The initial conditions have been read from equation (2.82).

Remark 2.9. We emphasize that the additional imposition of equation (2.85) entailing $y_2^{(0)} \equiv 0$ is not compelling. Therefore let us suppose $y_2^{(0)} \not\equiv 0$. Due to the initial condition $y_2^{(0)}(0) = 0$, the component $y_2^{(0)}(s)$ can be written as $s\tilde{y}_2^{(0)}(s)$ with some $\tilde{y}_2^{(0)}(s)$. Using (2.83) we obtain

$$\sigma^{(0)}(t/\epsilon^2, t) = \underbrace{y_1^{(0)}(t) + \tau y_2^{(0)}(t)}_{=\mathrm{O}(1)} \; - \underbrace{\epsilon^2 \, \underbrace{\tau \mathrm{e}^{-\omega t/\epsilon^2} t/\epsilon^2 \tilde{y}_2^{(0)}(t)}_{=\mathrm{O}(1)}}_{\mathrm{O}(\epsilon^2)} \; .$$

Thus, $\sigma^{(0)}$ is not only composed of O(1)-terms but also contains an $\mathrm{O}(\epsilon^2)$-term, that is expected to appear in the second order function $\sigma^{(2)}$. Even though this is not forbidden, as parts of higher order functions may be incorporated into lower order functions, it is against the idea of the expansion. This justifies our choice as particularly reasonable, since it disentangles the zeroth and second order.
It might be interesting to compute, how the appearance of the $\mathrm{O}(\epsilon^2)$-term in the zeroth order function $\sigma^{(0)}$ affects the second order function $\sigma^{(2)}$. Either this term belongs really to the expansion of σ_ϵ or it comes into play artificially and appears with inverse sign in $\sigma^{(2)}$.

Computation of $\mathbf{z}^{(2)}(r,s)$. The computation proceeds analogously to the previous one, observing that the source $Q^{(2)}$ does vanish *a posteriori* thanks to equation (2.86) and (2.87). So we obtain from (2.80)

$$\mathbf{z}^{(2)}(r,s) = \mathrm{e}^{rA}\mathbf{y}^{(2)}(s) = \begin{pmatrix} y_1^{(2)}(s) + \tau y_2^{(2)}(s) - \tau \mathrm{e}^{-\omega r} y_2^{(2)}(s) \\ \mathrm{e}^{-\omega r} y_2^{(2)}(s) \end{pmatrix} . \tag{2.88}$$

In order to determine $\mathbf{y}^{(2)}$ we must appeal to the representation of $\mathbf{z}^{(4)}$ by Duhamel's formula. Inserting (2.88) into

$$\mathbf{z}^{(4)}(r,s) = \mathrm{e}^{rA}\mathbf{y}^{(4)}(s) - \underbrace{\int_0^r \left[2\partial_s z_2^{(2)} + \omega \partial_s z_1^{(2)} + \omega\lambda z_1^{(2)} + \partial_s^2 z_1^{(0)}\right](\vartheta, s)\, \mathrm{e}^{(r-\vartheta)A} \binom{0}{1} \,\mathrm{d}\vartheta}_{=:J^{(2)}}$$

yields for the integral $J^{(2)}$

$$\begin{aligned}
J^{(2)} = \quad & \binom{\tau}{0} \int_0^r \left[\omega \dot{y}_1^{(2)} + \dot{y}_2^{(2)} + \omega\lambda y_1^{(2)} + \lambda y_2^{(2)} + \ddot{y}_1^{(0)}\right](s)\,\mathrm{d}\vartheta \\
+ & \binom{\tau}{0} \int_0^r \left[\dot{y}_2^{(2)} - \lambda y_2^{(2)}\right](s)\mathrm{e}^{-\omega\vartheta}\,\mathrm{d}\vartheta \\
+ & \binom{-\tau}{1} \int_0^r \left[\omega \dot{y}_1^{(2)} + \dot{y}_2^{(2)} + \omega\lambda y_1^{(2)} + \lambda y_2^{(2)} + \ddot{y}_1^{(0)}\right](s)\,\mathrm{e}^{\omega(\vartheta - r)}\,\mathrm{d}\vartheta \\
+ & \binom{-\tau}{1} \int_0^r \left[\dot{y}_2^{(2)} - \lambda y_2^{(2)}\right](s)\,\mathrm{e}^{-\omega r}\,\mathrm{d}\vartheta\,.
\end{aligned}$$

With the same arguments as in the computation of $\mathbf{z}^{(0)}$ we infer, that the integrand of the first integral must vanish. This results into equation (2.90) for $y_1^{(2)}$, that presupposes $y_2^{(2)}$ as given. Like (2.86), equation (2.89) is due to a special selection of $y_2^{(2)}$, such that the second and fourth integral on the right and hence $J^{(2)}$ become zero as well.

$$\left.\begin{aligned}
\dot{y}_2^{(2)} - \lambda y_2^{(2)} &= 0 \\
y_2^{(2)}(0) &= \beta - \dot{y}_1^{(0)}(0)
\end{aligned}\right\} \tag{2.89}$$

$$\left.\begin{aligned}
\dot{y}_1^{(2)} + \lambda y_1^{(2)} &= -\tau\dot{y}_2^{(2)} - \tau\lambda y_2^{(2)} - \tau\ddot{y}_1^{(0)} \\
y_1^{(2)}(0) &= 0
\end{aligned}\right\} \tag{2.90}$$

Computation of $\mathbf{z}^{(2k)}(r,s), k \geq 2$. In principle, the computation is repeatable in every order. Assuming that $\mathbf{y}^{(2k-2)}$ has been defined in such a way, that $Q^{(2k)}$ vanishes, equation (2.80) yields

$$\mathbf{z}^{(2k)}(r,s) = \mathrm{e}^{rA}\mathbf{y}^{(2k)}(s) = \begin{pmatrix} y_1^{(2k)}(s) + \tau y_2^{(2k)}(s) - \tau\mathrm{e}^{-\omega r} y_2^{(2k)}(s) \\ \mathrm{e}^{-\omega r} y_2^{(2k)}(s) \end{pmatrix}. \tag{2.91}$$

Plugging this into the representation of $\mathbf{z}^{(2k+2)}$ given once more by (2.80) we obtain

$$\begin{aligned}
\mathbf{z}^{(2k+2)}(r,s) &= \mathrm{e}^{rA}\mathbf{y}^{(2k+2)}(s) \\
&\quad - \underbrace{\int_0^r \left[2\partial_s z_2^{(2k)} + \omega\partial_s z_1^{(2k)} + \omega\lambda z_1^{(2k)} + \partial_s^2 z_1^{(2k-2)}\right](\vartheta,s)\,\mathrm{e}^{(r-\vartheta)A}\binom{0}{1}\,\mathrm{d}\vartheta}_{=:J^{(2k)}}\,.
\end{aligned}$$

Evidently, the integral $J^{(2k)}$ has the same structure as $J^{(0)}$ and $J^{(2)}$. This allows us to use the reasoning from above to extract the subsequent equations.

$$\left.\begin{aligned}
\dot{y}_2^{(2k+2)} - \lambda y_2^{(2k+2)} &= \tau\ddot{y}_2^{(2k)} \\
y_2^{(2k+2)}(0) &= -\dot{y}_1^{(2k)}(0)
\end{aligned}\right\} \tag{2.92}$$

$$\left.\begin{aligned}
\dot{y}_1^{(2k+2)} + \lambda y_1^{(2k+2)} &= -\tau\dot{y}_2^{(2k+2)} - \tau\lambda y_2^{(2k+2)} - \tau\ddot{y}_1^{(2k)} - \tau^2\ddot{y}_2^{(2k)} \\
y_1^{(2k+2)}(0) &= 0
\end{aligned}\right\} \tag{2.93}$$

It should be reemphasized, that the equation for $y_2^{(2k+2)}$ is not enforced by the ansatz. Notice also, that from the fourth order on, the preceding order function appears as source term in the equations of both components. Since (2.93) and (2.92) imply that the integral $J^{(2k)}$ becomes zero, the condition $Q^{(2k+2)} = 0$ is assured, which is necessary to go on likewise with the next order.

Interpretation of the two-scale expansion

In order to construe the equations obtained by the asymptotic expansion, it is advantageous to introduce new variables being more natural and thus simplifying the hierarchic evolution equations derived for the $\mathbf{y}^{(2k)}$s. For $k \in \mathbb{N}_0$ we set:

$$\zeta^{(2k)} := y_1^{(2k)} + \tau y_2^{(2k)}, \qquad \phi^{(2k)} := -\tau y_2^{(2k)} \ .$$

According to (2.78) and (2.91) this yields

$$\begin{aligned}\sigma^{(2k)}(r,s) = z_1^{(2k)}(r,s) &= y_1^{(2k)}(s) + \tau y_2^{(2k)}(s) - \tau \mathrm{e}^{-\omega r} y_2^{(2k)}(s) \\ &= \zeta^{(2k)}(s) + \mathrm{e}^{-\omega r}\phi^{(2k)}(s)\,.\end{aligned}$$

After resubstitution of the time t we get finally

$$\sigma^{(2k)}(t/\epsilon^2, t) = \underbrace{\zeta^{(2k)}(t)}_{\text{regular}} + \underbrace{\mathrm{e}^{-\omega t/\epsilon^2}\phi^{(2k)}(t)}_{\text{irregular}} \ .$$

This representation shows that each asymptotic order function can be written as the sum of a *regular* and an *irregular part*. The regular part depends only on the slow time $s = t$ like the order functions of the regular expansion. In contrast, the irregular part depends on both time scales; furthermore it reveals a multiplicative structure, where the dependence on the rapid time scale $r = t/\epsilon^2$ is given explicitly by the exponential function. Even if the second factor $\phi^{(2k)}$ is of exponential growth (cf. evolution equation for $\phi^{(2)}$ in (2.95)), for ϵ sufficiently small the exponential function effectuates a decay of the irregular part, that becomes the faster the smaller ϵ is chosen. Due to this circumstance the irregular part is called *initial layer*[33], because it is only "visible" as long as t is of the size ϵ^2. Therefore it is almost irrelevant for the behavior of $\sigma^{(2k)}$ over a time interval of magnitude 1.

Let us now express the IVPs determining the $\mathbf{y}^{(2k)}$s in terms of the regular part $\zeta^{(2k)}$ and the *initial layer function* $\phi^{(2k)}$. For $\phi^{(0)}$ and $\zeta^{(0)}$ we obtain from (2.86) and (2.87)

$$\phi^{(0)} \equiv 0 \qquad \begin{cases} \dot{\zeta}^{(0)} + \lambda\zeta^{(0)} = 0 \\ \zeta^{(0)}(0) = \alpha \end{cases} . \tag{2.94}$$

[33] In order to get convergence of the irregular part (initial layer) $\mathrm{e}^{-\omega t/\epsilon^2}\phi^{(2k)}(t)$ to a regular function for $\epsilon \to 0$, we must write it as a function of the rapid time. Replacing t by $\epsilon^2 r$ yields $\mathrm{e}^{-\omega r}\phi^{(2k)}(\epsilon^2 r) \xrightarrow{\epsilon\to 0} \mathrm{e}^{-\omega r}\phi^{(2k)}(0)$, so that the modulating function ϕ is reduced to its initial value. If $\sigma_\epsilon(t)$ is plotted versus the rapid time (cf. figure 2.5), the curves nearly coincide with the graph of $r \mapsto \mathrm{e}^{-\omega r}\phi^{(2k)}(0)$. Therefore it is reasonable to refer to $\mathrm{e}^{-\omega r}$ as the initial layer (profile).

Similarly, we obtain from (2.89) and (2.90)

$$\begin{cases}\dot\phi^{(2)} - \lambda\phi^{(2)} = 0 \\ \qquad \phi^{(2)}(0) = \tau\dot\zeta^{(0)}(0) - \tau\beta\end{cases} \qquad \begin{cases}\dot\zeta^{(2)} + \lambda\zeta^{(2)} = -\tau\ddot\zeta^{(0)} \\ \qquad \zeta^{(2)}(0) = -\phi^{(2)}(0)\end{cases} . \tag{2.95}$$

Eventually, the general case (2.92) and (2.93) leads to:

$$\begin{cases}\dot\phi^{(2k)} - \lambda\phi^{(2k)} = \tau\ddot\phi^{(2k-2)} \\ \qquad \phi^{(2k)}(0) = \tau\dot\phi^{(2k-2)}(0) + \tau\dot\zeta^{(2k-2)}(0)\end{cases} \qquad \begin{cases}\dot\zeta^{(2k)} + \lambda\zeta^{(2k)} = -\tau\ddot\zeta^{(2k-2)} \\ \qquad \zeta^{(2k)}(0) = -\phi^{(2k)}(0)\end{cases} \tag{2.96}$$

These IVPs can be solved[34]subsequently as well, computing first $\zeta^{(0)}$, then $\phi^{(2)}$ before $\zeta^{(2)}$, thereafter $\phi^{(4)}$ and $\zeta^{(4)}$ and so on.

Convergence of σ_ϵ. In contrast to the higher order-functions, $\sigma^{(0)}$ consists only of a regular part, which is the solution of the unperturbed IVP (2.70). Of course, this implies $\zeta^{(0)} = \sigma$. So we end up with the expansion

$$\sigma_\epsilon(t) = \underbrace{\zeta^{(0)}(t)}_{=\sigma(t)} + \epsilon^2\Big(\zeta^{(2)}(t) \ + \ \mathrm{e}^{-\omega t/\epsilon^2}\phi^{(2)}(t)\Big) \ + \ \ldots . \tag{2.97}$$

Similarly to the result obtained by the regular expansion (cf. equation (2.73)), this affirms the assertion of proposition 2.3. It becomes immediately clear, why σ_ϵ should approximate σ with a deviation of second order:

$$\sigma_\epsilon(t) - \sigma(t) = \epsilon^2\Big(\zeta^{(2)}(t) \ + \ \mathrm{e}^{-\omega t/\epsilon^2}\phi^{(2)}(t)\Big) \ + \ \ldots \quad \Rightarrow \quad |\sigma_\epsilon(t) - \sigma(t)| = \mathrm{O}(\epsilon^2) .$$

Convergence of $\dot\sigma_\epsilon$. Differentiating (2.97) with respect to the time yields

$$\dot\sigma_\epsilon(t) = \dot\zeta^{(0)} - \omega\mathrm{e}^{-\omega t/\epsilon^2}\phi^{(2)} + \epsilon^2\Big(\dot\zeta^{(2)} + \mathrm{e}^{-\omega t/\epsilon^2}\dot\phi^{(2)} - \omega\mathrm{e}^{-\omega t/\epsilon^2}\phi^{(4)}\Big) + \ \ldots \ . \tag{2.98}$$

Differentiation of the initial layer seems to add a term to the zeroth order. However, as the exponential function $\epsilon \mapsto \mathrm{e}^{-\omega t/\epsilon^2}$ fades down much faster than any power of ϵ for $t > 0$, this term can actually be counted to the second order. Similarly to lemma 2.7 one can prove[35]

$$\mathrm{e}^{-\omega t/\epsilon^2} \le \tfrac{\tau}{t}\epsilon^2 .$$

Therefore, $\dot\sigma_\epsilon$ approximates $\dot\sigma$ with second order accuracy, as claimed by proposition 2.4, if t stays away from a neighborhood of 0. Hence we get

$$|\dot\sigma_\epsilon(t) - \dot\sigma(t)| = \mathrm{O}(\epsilon^2) \qquad \text{uniformly for all } t > \theta,$$

where θ is some fixed positive time. A special case happens, if $\phi^{(2)}$ vanishes identically. Because of the absence of the initial layer in the second order, uniform approximation occurs then on the entire non-negative time-ray. This exceptional case is given if and only if $\phi^{(2)}$ is initialized by 0 or equivalently $\tau\dot\zeta^{(0)}(0) - \tau\beta = 0$ which is satisfied for $\dot\sigma_\epsilon(0) = \beta \overset{!}{=} -\alpha\lambda = \dot\sigma(0)$. So the initial layer (in the second order) is suppressed if σ_ϵ is initialized *compatibly.*

[34]As far as the successive solution of these IVPs is concerned, we need not care about questions of regularity. Since the starting equations for $\zeta^{(0)}$ and $\phi^{(2)}$ are homogeneous, their solutions are $\mathcal{C}^\infty$ and this entails the same smoothness for all the other functions pertaining to higher orders.

[35]Observe: $\mathrm{e}^{-\omega t/\epsilon^2} < M\epsilon^2 \ \Leftrightarrow \ \epsilon^{-2}\mathrm{e}^{-\omega t/\epsilon^2} =: f(\epsilon^2) \le M$. The constant M must be less or equal than the maximum of the function $f(\epsilon^2)$.

Further comments on the structure of the deviation $\sigma_\epsilon - \sigma$. Now we turn to the regular part of the leading error term $\sigma^{(2)}$. The linearity of the problem permits us once more to decompose easily $\zeta^{(2)}$ into two parts, that are of different provenance

$$\zeta^{(2)} = \underbrace{\zeta_N^{(2)}}_{\substack{\text{natural}\\\text{deviation}}} + \underbrace{\zeta_I^{(2)}}_{\substack{\text{“interaction”}\\\text{with initial}\\\text{layer}}} .$$

$\zeta_N^{(2)}$ and $\zeta_I^{(2)}$ are defined by the following IVPs:

$$\begin{cases} \dot{\zeta}_N^{(2)} + \lambda\zeta_N^{(2)} = -\tau\ddot{\zeta}^{(0)} \\ \zeta_N^{(2)}(0) = 0 \end{cases} \qquad \begin{cases} \dot{\zeta}_I^{(2)} + \lambda\zeta_I^{(2)} = 0 \\ \zeta_I^{(2)}(0) = -\phi^{(2)}(0) \end{cases}$$

Since $\zeta_N^{(2)}$ is driven by the solution of the limit problem, it represents the constituent of the deviation, that is “naturally” expected by writing the perturbed equation (2.69) in the form of the limit equation plus a small source, viz. $\dot{\sigma}_\epsilon + \lambda\sigma_\epsilon = -\epsilon^2\tau\ddot{\sigma}_\epsilon$. As this kind of error is encountered in many approximation processes, it is not specific for the singular nature of (2.69).
Like the initial layer, $\zeta_I^{(2)}$ is generated by incompatible initialization $\beta \neq -\alpha\lambda$. However, it reveals the same decay rate as the solution of the limit problem $\zeta^{(0)}$. This shows that an incompatible initialization does not only spark the rapidly vanishing initial layer but also affects the long term behavior of σ_ϵ.
Just as the deviation terms in the second order are driven by the exact solution and the initial incompatibility, they feed terms of fourth order, that appear again as initial conditions and sources for the sixth order and so forth. Asymptotically, the prevailing terms are, of course, those of second order in ϵ.
Observe that $\zeta_N^{(2)}$ is equal to $\varsigma^{(2)}$ of the regular expansion (2.71), as both satisfy the same IVP. Further terms in the second order of the two-scale expansion (2.97) are masked out by the regular expansion in (2.71). This shows clearly that a full analysis of the singular limit problem requires a two-scale expansion. Nevertheless, the main feature, namely the convergence of σ_ϵ to σ, is correctly predicted even by the regular expansion.
It should be pointed out, that the different constituents of the error (initial layer, natural part, interaction), that we have identified, can be found less precisely in the estimates performed to prove proposition 2.3 and 2.4.

Smooth initialization up to arbitrary order. The compatible initialization $\dot{\sigma}_\epsilon(0) = -\alpha\lambda$ suppresses only the initial layer $\phi^{(2)}$ in second order, while the initial layers in higher orders, i.e. $\phi^{(4)}, \phi^{(6)}, ...$ persist in general. These become noticeable, if we consider the convergence of higher derivatives $\ddot{\sigma}_\epsilon \to \ddot{\sigma}, \dddot{\sigma}_\epsilon \to \dddot{\sigma}, ...$ If we initialize $\dot{\sigma}_\epsilon$ by a truncated expansion

$$\dot{\sigma}_\epsilon(0) = \beta^{(0)} + \epsilon^2\beta^{(2)} + \epsilon^4\beta^{(4)} + ... , \tag{2.99}$$

it is possible to “switch off” arbitrarily many initial layers by a suitable choice of the coefficients $\beta^{(2k)}$. In the sequel we determine the initial condition which removes the initial layers up to order $2n+2$. We call such an initialization smooth up to

order $2n+2$. This ensures that the leading derivatives of σ_ϵ converge uniformly in a neighborhood of 0 and hence on $[0,\infty)$.
Inserting (2.97) into (2.99), we obtain by equating terms of same order in ϵ (comparison of coefficients):

$$\begin{aligned} \epsilon^0 \;: &\qquad \phi^{(2)}(0) = \tau\Big(\dot\zeta^{(0)}(0) - \beta^{(0)}\Big) \\ \epsilon^{2k}: &\qquad \phi^{(2k+2)}(0) = \tau\Big(\dot\zeta^{(2k)}(0) + \dot\phi^{(2k)}(0) - \beta^{(2k)}\Big) \end{aligned} \tag{2.100}$$

Thus we discover, how the coefficients $\beta^{(2k)}$ should be adjusted. First, the known compatibility condition $\beta^{(0)} = \dot\zeta^{(0)}$ making $\phi^{(2)}$ vanish is recovered. Observing that this implies in particular $\dot\phi^{(2)}(0) = 0$ we set $\beta^{(2)} = \dot\zeta^{(2)}(0)$, such that $\phi^{(4)}(0) = 0$. So the initial layer function $\phi^{(4)}$ is initialized by zero, while the source $\ddot\phi^{(2)}$ appearing in its evolution equation vanishes too. Hence $\phi^{(4)}$ must also vanish, as a linear homogeneous ODE with zero initial condition admits only the trivial solution. This procedure can be continued by setting $\beta^{(4)} = \dot\zeta^{(4)}(0)$ to enforce $\phi^{(6)}$ to stay zero and so forth.
However, there remains a problem, because $\dot\zeta^{(0)}(0), \dot\zeta^{(2)}(0), \dot\zeta^{(4)}(0), \ldots$ are not given explicitly. In some cases, like this example, it is possible to compute these values from the given initial data, without solving the IVPs for the regular asymptotic order functions. In fact, as $\zeta^{(0)}(0)$ is given, we obtain by means of the evolution equation for $\zeta^{(0)}$:

$$\dot\zeta^{(0)}(0) = -\lambda\zeta^{(0)}(0) = -\alpha\lambda \; .$$

Differentiating the evolution equation of $\zeta^{(0)}$ yields

$$\ddot\zeta^{(0)}(0) = -\lambda\dot\zeta^{(0)}(0) = \alpha\lambda^2 \quad \text{and by induction } \tfrac{\mathrm{d}^n}{\mathrm{d}t^n}\zeta^{(0)}(0) = (-1)^n\alpha\lambda^n \; .$$

This enables us to compute the derivatives of $\zeta^{(2)}(t)$ for $t=0$. By differentiating its evolution equation n times and setting $t=0$ we arrive at the recursion relation

$$\tfrac{\mathrm{d}^{n+1}}{\mathrm{d}t^{n+1}}\zeta^{(2)}(0) = -\lambda\tfrac{\mathrm{d}^n}{\mathrm{d}t^n}\zeta^{(2)}(0) - \tau\tfrac{\mathrm{d}^{n+2}}{\mathrm{d}t^{n+2}}\zeta^{(0)}(0) \; ,$$

with the initial condition for $n=0$: $\zeta^{(2)}(0) = -\phi^{(2)}(0) = 0$, as we suppose that $\phi^{(2)}$ vanishes by the choice of $\beta^{(0)}$. The solution of this recurrence equation is given by

$$\tfrac{\mathrm{d}^n}{\mathrm{d}t^n}\zeta^{(2)}(0) = (-1)^n n\tau\alpha\lambda^{n+1}.$$

Proceeding in this way one can compute iteratively the derivatives of all regular asymptotic orders at the initial time $t=0$. Let us demonstrate this once more for $\zeta^{(4)}$: from its evolution we obtain the recursive equation

$$\tfrac{\mathrm{d}^{n+1}}{\mathrm{d}t^{n+1}}\zeta^{(4)}(0) = -\lambda\tfrac{\mathrm{d}^n}{\mathrm{d}t^n}\zeta^{(4)}(0) - \tau\tfrac{\mathrm{d}^{n+2}}{\mathrm{d}t^{n+2}}\zeta^{(2)}(0) \; .$$

Assuming that $\beta^{(2)}$ compensates $\dot\zeta^{(2)}(0)$, such that $\phi^{(4)}$ becomes identically zero, this entails $\zeta^{(4)}(0) = -\phi^{(4)}(0) = 0$ as start value for $n=0$. By induction the following explicit representation is verified

$$\tfrac{\mathrm{d}^n}{\mathrm{d}t^n}\zeta^{(4)}(0) = (-1)^n\tfrac{1}{2}(n^2+3n)\tau^2\alpha\lambda^{n+2} \; .$$

Let us summarize the result of this paragraph in the subsequent proposition.

plot	fine & solid line (red)	bold & dashed line (blue)
a)	$\sigma^{(0)}$	σ_ϵ
b)	$\sigma^{(2)}$	$\frac{1}{\epsilon^2}\left[\sigma_\epsilon - \sigma^{(0)}\right]$
c)	$\sigma^{(4)}$	$\frac{1}{\epsilon^4}\left[\sigma_\epsilon - \sigma^{(0)} - \epsilon^2\sigma^{(2)}\right]$
d)	$\sigma^{(6)}$	$\frac{1}{\epsilon^6}\left[\sigma_\epsilon - \sigma^{(0)} - \epsilon^2\sigma^{(2)} - \epsilon^4\sigma^{(4)}\right]$
e)	$\dot\zeta^{(0)} - \omega \mathrm{e}^{-\omega t/\epsilon^2}\phi^{(2)}$	$\dot\sigma_\epsilon$
f)	$\dot\zeta^{(2)} + \mathrm{e}^{-\omega t/\epsilon^2}\left(\dot\phi^{(2)} - \omega\phi^{(4)}\right)$	$\frac{1}{\epsilon^2}\left[\dot\sigma_\epsilon - \dot\zeta^{(0)} + \mathrm{e}^{-\omega t/\epsilon^2}\left(\dot\phi^{(0)} - \omega\phi^{(2)}\right)\right]$
g)	$\dot\zeta^{(4)} + \mathrm{e}^{-\omega t/\epsilon^2}\left(\dot\phi^{(4)} - \omega\phi^{(6)}\right)$	$\frac{1}{\epsilon^4}\left[\dot\sigma_\epsilon - \sum_{k=0}^{1}\epsilon^{2k}\dot\zeta^{(2k)} + \epsilon^{2k}\mathrm{e}^{-\omega t/\epsilon^2}\left(\dot\phi^{(2k)} - \omega\phi^{(2k+2)}\right)\right]$
h)	$\dot\zeta^{(6)} + \mathrm{e}^{-\omega t/\epsilon^2}\left(\dot\phi^{(6)} - \omega\phi^{(8)}\right)$	$\frac{1}{\epsilon^6}\left[\dot\sigma_\epsilon - \sum_{k=0}^{2}\epsilon^{2k}\dot\zeta^{(2k)} + \epsilon^{2k}\mathrm{e}^{-\omega t/\epsilon^2}\left(\dot\phi^{(2k)} - \omega\phi^{(2k+2)}\right)\right]$

Table 2.2: Quantities plotted in figure 2.6 and 2.7

.

Proposition 2.8. *The following initial conditions of the IVP (2.69)*

$$\sigma_\epsilon(0) = \alpha, \qquad \dot\sigma_\epsilon(0) = \dot\zeta^{(0)}(0) + \epsilon^2\dot\zeta^{(2)}(0) + ... + \epsilon^{2n}\dot\zeta^{(2n)}(0)$$

imply $\phi^{(2)} \equiv \phi^{(4)} \equiv ... \equiv \phi^{2n+2} \equiv 0$. *Hence there is no initial layer up to the order* $2n+2$ *inclusively. The* a priori *unknown values* $\dot\zeta^{(0)}(0), \dot\zeta^{(2)}(0), ..., \dot\zeta^{(2n)}(0)$ *are available without solving any differential equation as described above.*

Additive two-scale ansatz. It should be mentioned for completeness, that beside the general two-scale expansion ansatz pursued in this section, more specialized expansions are imaginable. In particular, the observations concerning the nature of the initial layer motivate an *additive* two-scale expansion, the more so as the IVP (2.69) is linear. Inserting the ansatz

$$\sigma_\epsilon(t) = \xi^{(0)}(t) + \epsilon^2\Big(\xi^{(2)}(t) + \psi^{(2)}(t/\epsilon^2)\Big) + \epsilon^4\Big(\xi^{(4)}(t) + \psi^{(4)}(t/\epsilon^2)\Big) + ... \tag{2.101}$$

into (2.69) and performing the comparison of coefficients leads to the following equations, where the dots over the ψ's refer to their argument being t/ϵ^2 and not t.

$$\begin{aligned} \epsilon^0 &: \quad \dot\xi^{(0)} + \lambda\xi^{(0)} + \tau\ddot\psi^{(2)} + \dot\psi^{(2)} = 0 \\ \epsilon^{2k} &: \quad \dot\xi^{(2k)} + \lambda\xi^{(2k)} + \tau\ddot\xi^{(2k-2)} + \tau\ddot\psi^{(2k)} + \dot\psi^{(2k)} + \lambda\psi^{(2k-2)} = 0 \end{aligned} \tag{2.102}$$

and

$$\begin{aligned} \epsilon^0 &: \quad \xi^{(0)}(0) = \alpha \qquad & \dot\xi^{(0)}(0) + \dot\psi^{(2)}(0) &= \beta \\ \epsilon^{2k} &: \quad \xi^{(2k)}(0) + \psi^{(2k)}(0) = 0 \qquad & \dot\xi^{(2k)}(0) + \dot\psi^{(2k+2)}(0) &= 0 \end{aligned} \tag{2.103}$$

In order to solve these equations it is reasonable to split them in such a way, that each equation contains only one function as unknown. Of course, this step is not compelling and represents only one possibility to solve the above equations. Doing so we obtain again a hierarchic system of IVPs, that can be solved in a subsequent manner:

$$\left\{\begin{aligned} \dot\xi^{(0)} + \lambda\xi^{(0)} &= 0 \\ \xi^{(0)}(0) &= \alpha \end{aligned}\right. \qquad \left\{\begin{aligned} \ddot\psi^{(2)} + \omega\dot\psi^{(2)} &= 0 \\ \psi^{(2)}(0) &= 0 \\ \dot\psi^{(2)}(0) &= \beta - \dot\xi^{(0)}(0) \end{aligned}\right.$$

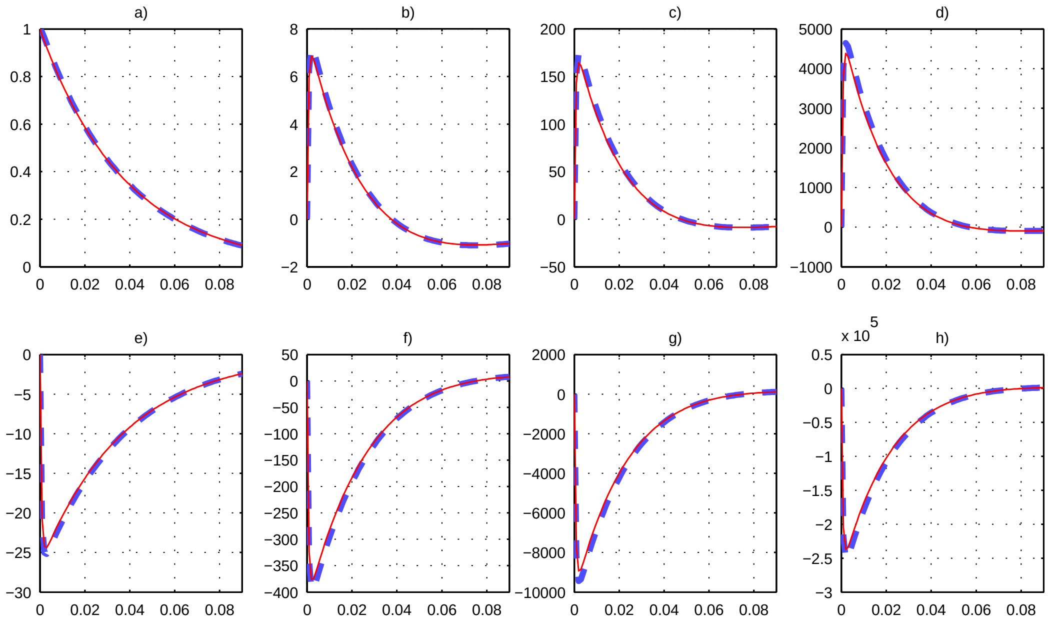


Figure 2.6: The plots illustrate the expansion for σ_ϵ (upper row a-d) and its time derivative $\dot{\sigma}_\epsilon$ (lower row e-h). A detailed listing of the plotted quantities is given in table 2.2. Observe the initial layers occurring in all plots but a). Furthermore the growth of the asymptotic order functions should be noticed. Chosen parameters: $\epsilon = 0.045$, $\tau = 0.3$, $\lambda = 9\pi^2\tau$, $\alpha = 1$ and $\beta = 0$. The order functions of the expansion, i.e. the solutions of (2.94), (2.95) and (2.96), have been computed with Maple.

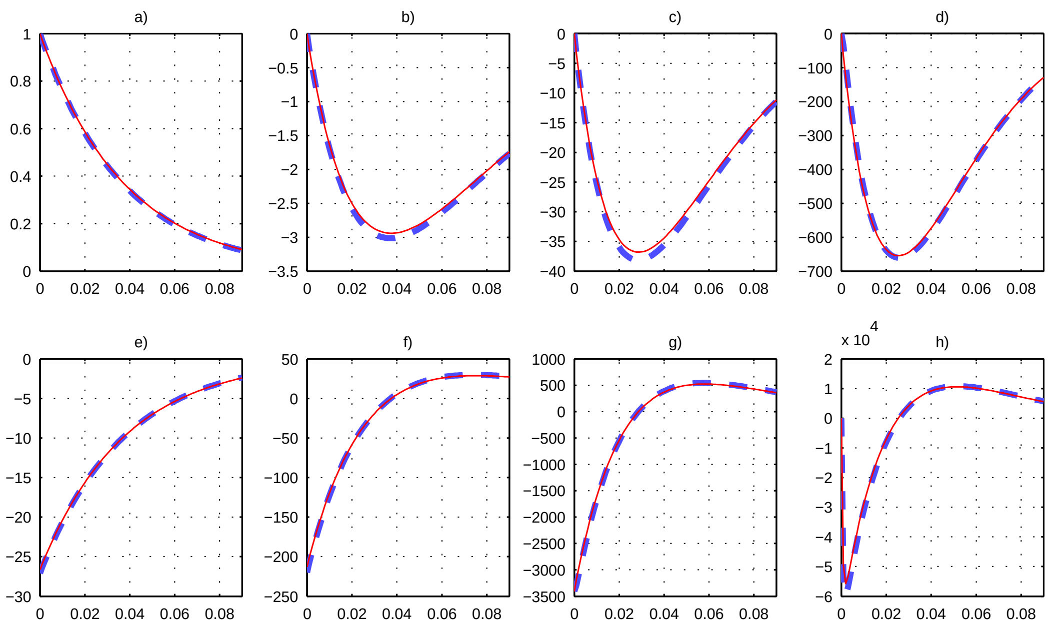


Figure 2.7: With exception of the second initial value β, the curves refer to the same parameter setting as in figure 2.6. In order to exemplify the effect of the smooth initialization as described above, the following value for β has been taken: $\beta = \beta^{(0)} + \epsilon^2\beta^{(2)} + \epsilon^4\beta^{(4)}$ with $\beta^{(0)} = \dot{\zeta}^{(0)}(0) = -\alpha\lambda$, $\beta^{(2)} = \dot{\zeta}^{(2)}(0) = -\tau\alpha\lambda^2$, $\beta^{(4)} = \dot{\zeta}^{(4)}(0) = -2\tau^2\alpha\lambda^3$. Apparently, the initial layer disappear up to the third order. The only initial layer to be detected in plot h) stems from the fourth order.

$$k \geq 1: \begin{cases} \dot{\xi}^{(2k)} + \lambda \xi^{(2k)} &= -\tau \xi^{(2k-2)} \\ \xi^{(2k)}(0) &= 0 \end{cases} \qquad k \geq 2: \begin{cases} \ddot{\psi}^{(2k)} + \omega \dot{\psi}^{(2k)} &= -\lambda \omega \psi^{(2k-2)} \\ \psi^{(2k)}(0) &= 0 \\ \dot{\psi}^{(2k)}(0) &= -\dot{\xi}^{(2k-2)}(0) \end{cases}$$

A priori, these equations convey a different picture from that obtained by (2.94), (2.95) and (2.96). In particular, the irregular part does not couple to the regular one. Moreover the equations hide the internal structure of the initial layer and it is not definitely clear, that the $\psi^{(2k)}$'s are really functions of the rapid time t/ϵ^2. To answer this, a deeper analysis of the equations should be done.

Towards rigor

The aim of this subsection is to present a rigorous justification of the two-scale expansion (2.74). It will be shown, that (2.74) derived so far by formal arguments converges to the solution σ_ϵ of (2.69), if the number of asymptotic order functions is increased. For this we introduce the *comparison function* of order $n \in \mathbb{N}$

$$\hat{\sigma}_\epsilon^{[n]}(t) := \zeta^{(0)} + \sum_{k=1}^{n} \epsilon^{2k} \big[\zeta^{(2k)} + \mathrm{e}^{-\omega t/\epsilon^2} \phi^{(2k)}(t) \big], \tag{2.104}$$

that corresponds to the two-scale expansion truncated behind the order $2n$ (remember that $\zeta^{(2k)}$ and $\phi^{(2k)}$ have been defined as solutions of (2.94), (2.95) and (2.96)). Due to the construction of the expansion, it is no surprise that (2.104) solves (2.69) approximately leaving a residual proportional to ϵ^{2n} if plugged into the evolution equation. A similar behavior is observed for the second initial condition, whereas the first one is exactly fulfilled.

Lemma 2.9. *The comparison function (2.104) satisfies the following IVP:*

$$\left.\begin{array}{ll} \mathrm{EQ}: & \epsilon^2 \tau \ddot{\hat{\sigma}}_\epsilon^{[n]} + \dot{\hat{\sigma}}_\epsilon^{[n]} + \lambda \hat{\sigma}_\epsilon^{[n]} = \epsilon^{2n+2} \tau \big(\ddot{\zeta}^{(2n)} + \mathrm{e}^{-\omega t/\epsilon^2} \ddot{\phi}^{(2n)} \big) \\ \mathrm{IC}: & \hat{\sigma}_\epsilon^{[n]}(0) = \alpha \quad \wedge \quad \dot{\hat{\sigma}}_\epsilon^{[n]}(0) = \begin{cases} \dot{\zeta}^{(0)}(0) & \text{if } n = 0 \\ \beta + \epsilon^{2n} \big(\dot{\zeta}^{(2n)}(0) + \dot{\phi}^{(2n)}(0) \big) & \text{if } n \geq 1 \end{cases} \end{array}\right\}$$

Proof: The proof is a straightforward computation exploiting the equations (2.94), (2.95) and (2.96) satisfied by $\zeta^{(0)}, ..., \zeta^{(2n)}$ and by $\phi^{(2)}, ..., \phi^{(2n)}$ respectively. ■

Let us now consider the difference

$$D_\epsilon^{[n]} := \hat{\sigma}_\epsilon^{[n]} - \sigma_\epsilon \,.$$

By linearity $D_\epsilon^{[n]}$ is the solution of the IVP

$$\left.\begin{array}{l} \mathrm{EQ}: \ \epsilon^2 \tau \ddot{D}_\epsilon^{[n]} + \dot{D}_\epsilon^{[n]} + \lambda D_\epsilon^{[n]} = \epsilon^{2n+2} \tau \big(\ddot{\zeta}^{(2n)} + \mathrm{e}^{-\omega t/\epsilon^2} \ddot{\phi}^{(2n)} \big) =: \epsilon^{2n+2} \tau Q_\epsilon^{[n]} \\ \mathrm{IC}: \ D_\epsilon^{[n]}(0) = 0 \ \wedge \ \dot{D}_\epsilon^{[n]}(0) =: \epsilon^{2n} A^{[n]} = \begin{cases} \dot{\zeta}^{(0)}(0) - \beta = -\alpha\lambda - \beta, \ n = 0 \\ \epsilon^{2n} \big(\dot{\zeta}^{(2n)}(0) + \dot{\phi}^{(2n)}(0) \big), n \geq 1 \end{cases} \end{array}\right\} \tag{2.105}$$

that we write as a first order system in matrix form:

$$
\begin{aligned}
\text{EQ:}\quad & \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} D_\epsilon^{[n]} \\ \dot D_\epsilon^{[n]} \end{pmatrix} + \omega/\epsilon^2 \overbrace{\begin{pmatrix} 0 & -\epsilon^2\tau \\ \lambda & 1 \end{pmatrix}}^{=:M_\epsilon} \overbrace{\begin{pmatrix} D_\epsilon^{[n]} \\ \dot D_\epsilon^{[n]} \end{pmatrix}}^{=:\mathbf{D}_\epsilon^{[n]}} &=& \ \epsilon^{2n} \overbrace{\begin{pmatrix} 0 \\ Q_\epsilon^{[n]} \end{pmatrix}}^{=:\mathbf{Q}_\epsilon^{[n]}} \\
\text{IC:}\quad & \begin{pmatrix} D_\epsilon^{[n]}(0) \\ \dot D_\epsilon^{[n]}(0) \end{pmatrix} &=& \ \epsilon^{2n} \underbrace{\begin{pmatrix} 0 \\ A^{[n]} \end{pmatrix}}_{=:\mathbf{A}^{[n]}}
\end{aligned}
\tag{2.106}
$$

As the source term on the right hand side of the evolution equation and of the initial condition scales with ϵ^{2n}, the solution should reflect this scaling due to the linearity of the problem. Nevertheless, it is necessary to estimate $D_\epsilon^{[n]}$ because of further dependencies on ϵ^2 appearing in the source $Q_\epsilon^{[n]}$ and the system matrix M_ϵ.

Once again *Duhamel's formula* (cf. page 112) renders a good service. Using the shortcuts from above, we obtain:

$$
\mathbf{D}_\epsilon^{[n]}(t) = \epsilon^{2n}\left[\mathrm{e}^{-\omega t/\epsilon^2} \mathrm{e}^{tM_\epsilon} \mathbf{A}^{[n]} + \int_0^t \mathrm{e}^{-\omega(t-\vartheta)/\epsilon^2} \mathrm{e}^{(t-\vartheta)M_\epsilon} \mathbf{Q}_\epsilon^{[n]}(\vartheta)\,\mathrm{d}\vartheta \right] .
$$

Let us now apply the maximum-norm $\|\cdot\|$ in $\mathbb{R}^2$. The time t should be contained in some compact interval $[0,T], T>0$. Using the triangle inequality and the standard norm inequality for matrices and linear operators, we get:

$$
\begin{aligned}
\left\|\mathbf{D}_\epsilon^{[n]}(t)\right\| &\equiv \max\left\{ \left|D_\epsilon^{[n]}(t)\right|,\ \left|\dot D_\epsilon^{[n]}(t)\right| \right\} \\
&\le \epsilon^{2n}\left[\mathrm{e}^{-\omega t/\epsilon^2} \left\|\mathrm{e}^{tM_\epsilon}\mathbf{A}^{[n]}\right\| + \int_0^t \mathrm{e}^{-\omega(t-\vartheta)/\epsilon^2} \left\|\mathrm{e}^{(t-\vartheta)M_\epsilon}\mathbf{Q}_\epsilon^{[n]}(\vartheta)\right\| \mathrm{d}\vartheta \right] \\
&\le \epsilon^{2n}\left[\mathrm{e}^{-\omega t/\epsilon^2} \left\|\mathrm{e}^{tM_\epsilon}\right\| \left\|\mathbf{A}^{[n]}\right\| + \int_0^t \mathrm{e}^{-\omega(t-\vartheta)/\epsilon^2} \left\|\mathrm{e}^{(t-\vartheta)M_\epsilon}\right\| \left\|\mathbf{Q}_\epsilon^{[n]}(\vartheta)\right\| \mathrm{d}\vartheta \right]
\end{aligned}
$$

Since the norm of an exponential matrix is dominated by the exponential function of the matrix norm, we draw the norm into the exponential function. Furthermore we replace the time argument in the exponential functions containing $\|M_\epsilon\|$ by the maximal time T and estimate the integrand, so that there remains only the factor $\mathrm{e}^{-\omega(t-\vartheta)/\epsilon^2}$ to be integrated:

$$
\begin{aligned}
\left\|\mathbf{D}_\epsilon^{[n]}(t)\right\| &\le \epsilon^{2n}\left[\mathrm{e}^{-\omega t/\epsilon^2} \mathrm{e}^{t\|M_\epsilon\|} \left|A^{[n]}\right| + \int_0^t \mathrm{e}^{-\omega(t-\vartheta)/\epsilon^2} \mathrm{e}^{(t-\vartheta)\|M_\epsilon\|} \left|Q_\epsilon^{[n]}(\vartheta)\right| \mathrm{d}\vartheta \right] \\
&\le \epsilon^{2n}\mathrm{e}^{T\|M_\epsilon\|}\left[\mathrm{e}^{-\omega t/\epsilon^2} |A^{[n]}| + \sup_{t\in[0,T]} |Q_\epsilon^{[n]}(t)| \int_0^t \mathrm{e}^{-\omega(t-\vartheta)/\epsilon^2}\,\mathrm{d}\vartheta \right] \\
&\le \epsilon^{2n}\mathrm{e}^{T\|M_\epsilon\|}\left[\mathrm{e}^{-\omega t/\epsilon^2} |A^{[n]}| + \sup_{t\in[0,T]} |Q_\epsilon^{[n]}(t)| \tau\epsilon^2 \left(1-\mathrm{e}^{-\omega t/\epsilon^2}\right) \right]
\end{aligned}
\tag{2.107}
$$

Now, the right hand side is much easier to control and we can formulate our central result.

Theorem 2.6. *For any $n \in \mathbb{N}_0$ there is a constant $C > 0$ such that for any $\epsilon \in (0,1]$ and $t \in [0,T]$ the following estimate holds true:*

$$\left|\hat{\sigma}_\epsilon^{[n]}(t) - \sigma_\epsilon(t)\right| \leq C\epsilon^{2n}.$$

The constant C depends on n, T and λ, τ as well as on the initial conditions of (2.69).

Proof: The main part of the proof has already been done above. Let us set

$$\begin{aligned}
m &:= \max\{\tau, 1+\lambda\} \geq \left\| \begin{pmatrix} 0 & -\epsilon^2\tau \\ \lambda & 1 \end{pmatrix} \right\| = \|M_\epsilon\|, \\
q &:= \sup_{t\in[0,T]} \tau\Big(|\ddot{\zeta}^{(2n)}(t)| + |\ddot{\phi}^{(2n)}(t)|\Big) \\
&\geq \sup_{t\in[0,T]} \tau\Big|\ddot{\zeta}^{(2n)}(t) + \mathrm{e}^{-\omega t/\epsilon^2}\ddot{\phi}^{(2n)}(t)\Big| \equiv \sup_{t\in[0,T]} |Q_\epsilon^{[n]}(t)|
\end{aligned}$$

m and q estimate $\|M_\epsilon\|$ and $\sup_{t\in[0,T]}|Q_\epsilon^{[n]}(t)|$ independently of ϵ: In particular, q is finite as supremum of a continuous functions over the compact time interval $[0,T]$. Furthermore q is independent of ϵ, because $\zeta^{(2n)}$ and $\phi^{(2n)}$ are computed by solving successively the IVPs (2.94),(2.95) and (2.96). Therefore, apart from the time t, $\ddot{\zeta}^{(2n)}(t)$ and $\ddot{\phi}^{(2n)}(t)$ depend ultimately only on the external initial data α, β of (2.94) and (2.95) as well as on the constants λ and $\tau = \omega^{-1}$. Similarly, $a := |A^{[n]}| \equiv |\dot{\zeta}^{(2n)}(0) + \dot{\phi}^{(2n)}(0)|$ is fully determined by these parameters.
Consider now the inequality in (2.107). We recall that $\left|\hat{\sigma}_\epsilon^{[n]}(t) - \sigma_\epsilon(t)\right| \leq \left\|\mathbf{D}_\epsilon^{[n]}(t)\right\|$. Evidently, the claim works with any constant C satisfying

$$C \;\geq\; \mathrm{e}^{mT}(a + \tau q) \;\geq\; \mathrm{e}^{T\|M_\epsilon\|}\Big(|A^{[n]}| + \tau \sup_{t\in[0,T]} |Q_\epsilon^{[n]}(t)|\Big),$$

Observe that $\mathrm{e}^{-\omega t/\epsilon^2}$ and $\epsilon^2(1-\mathrm{e}^{-\omega t/\epsilon^2})$ are crudely[36] estimated by 1. ∎

Remark 2.10. Theorem 2.6 gives an estimate being by a factor of ϵ^2 worse than one might have expected. In particular, it can not be concluded immediately, that σ_ϵ converges to the solution $\sigma = \zeta^{(0)}$ of the limit problem (2.70). To obtain this result we argue in the following way. Let C be the constant mentioned in the theorem and let $n = 2$. Then the subsequent estimate is possible

$$\begin{aligned}
\Big(C + \sup_{t\in[0,T]} |\sigma^{(2)}(t/\epsilon^2, t)|\Big)\epsilon^2 &\geq \left|\sigma_\epsilon(t) - \hat{\sigma}_\epsilon^{[2]}(t)\right| \;+\; \epsilon^2|\sigma^{(2)}(t/\epsilon^2, t)| \\
&\geq \left|\sigma_\epsilon(t) - \sigma^{(0)}(\tfrac{t}{\epsilon^2}, t) - \epsilon^2\sigma^{(2)}(\tfrac{t}{\epsilon^2}, t) + \epsilon^2\sigma^{(2)}(\tfrac{t}{\epsilon^2}, t)\right| \\
&\geq \left|\sigma_\epsilon(t) - \zeta^{(0)}(t)\right| = \left|\sigma_\epsilon(t) - \sigma(t)\right|,
\end{aligned}$$

which gives the desired result, as

$$|\sigma^{(2)}(t/\epsilon^2, t)| \leq |\zeta^{(2)}(t)| + |\mathrm{e}^{-\omega t/\epsilon^2}\phi^{(2)}(t)| \leq |\zeta^{(2)}(t)| + |\phi^{(2)}(t)|$$

can be estimated independently of ϵ over any compact time interval.

[36]Note that the theorem targets on the rate of asymptotic similarity in $\mathcal{C}^1([0,T])$. For this purpose we accept the rough estimate of the decaying exponential functions.

Remark 2.11. The detour to consider the expansion up to one order higher than intuitively necessary is met again in the subsequent chapter. Insofar it is typical for our approach to singularly perturbed problems.
In the present case, the reason for this circumstance is quite obvious: The estimate in theorem 2.6 refers to $\mathbf{D}_\epsilon^{[n]}$. So it is simultaneously done for $|\hat{\sigma}_\epsilon(t) - \sigma_\epsilon(t)|$ and the time derivative $|\dot{\hat{\sigma}}_\epsilon(t) - \dot{\sigma}_\epsilon(t)|$. But we already know from the previous section and from the heuristic analysis on page 117, that the time derivatives cannot be asymptotically similar of order 2 in $\mathcal{C}^0\big([0,T]\big)$. Only if $t > \theta > 0$, a finer estimate of $\mathrm{e}^{-\omega t/\epsilon^2}$ in (2.107) extends the second order asymptotic similarity in $\mathcal{C}^0\big([\theta,T]\big)$ to the first derivatives $\dot{\hat{\sigma}}_\epsilon(t), \dot{\sigma}_\epsilon(t)$. (Compare proposition 2.3 and 2.4.) In the current ODE setting (without sources) we do not need any regularity requirements as in later chapters.

Remark 2.12. Compatible initialization $\beta = -\alpha\lambda = \dot{\zeta}^{(0)}$ entails $A^{(0)} = 0$, such that we may infer from (2.107) as well, that σ_ϵ and $\hat{\sigma}_\epsilon^{[0]} = \sigma$ are asymptotically similar of second order in $\mathcal{C}^1\big([0,T]\big)$ under this initial condition.

Summary

The main feature of the singular IVP (2.69) is that the derivative of its solution converges generally in a weaker topology to its limit function than the solution itself (pointwise versus uniform convergence on $[0,T]$). The reason for this behavior was attributed to the appearance of an initial layer. It was the objective of this section to demonstrate that we can deal successfully with such a problem by a multiscale expansion, which is well-suited to understand the dynamics of the initial layer. The result of the formal two-scale expansion leads to a set of regular IVPs, that provide information about the time evolution and the composition of the deviation.

Chapter 3

Analysis of a D1P3 lattice-Boltzmann equation

This chapter is devoted to a study of a lattice-Boltzmann equation based on the D1P3 model. Even if the practical relevance of this model is rather limited, as it refers only to one space dimension, it represents a good point of departure for theoretical investigations. This is owed to the fact, that many features of lattice-Boltzmann equations already turn up analyzing a three velocity model equation. In this sense the D1P3 model becomes a *paradigm*, which is not only studied for the sake of its own. Some observations can be carried over in a more or less straightforward manner to more complicated lattice-Boltzmann equations, others might be more intricate. Nevertheless, a clear understanding of a simplified case provides a good feeling for the *how* and *why* of more complex situations.

We remind that a D1P3 lattice-Boltzmann equation has already been presented in subsection 1.1.2. In contrast to the example given there, we consider the *parabolic scaling* here. The analysis follows the guideline of section 1.3 using a *comparison function*, which is constructed in the form of an asymptotic expansion.

The energy estimate in section 3.1 is inspired by [41]. This paper provides a pioneering mathematical discussion of a D2P9 lattice-Boltzmann equation with the Navier-Stokes equilibrium (1.31). [41] as well as [1] emphasize the decisive role of certain algebraic properties that the collision operator should satisfy to guarantee stability. These properties[1] are reflected by lemma 3.1 in the context of the D1P3 BGK lattice-Boltzmann equation with equilibrium (3.3).

A paradigm

Let $L, T > 0$ such that $\mathcal{X} := [0, L]$ and $\mathcal{T} := [0, T]$ abbreviate a spatial domain and a time interval respectively. Furthermore let us set $\mathcal{X}_T := \mathcal{T} \times \mathcal{X}$. We consider the following initial value problem for the lattice-Boltzmann equation:

$$\begin{aligned} \mathsf{f}_\epsilon(0,x) &= v_0(x)\mathsf{w}, & x \in \mathcal{X} \qquad (3.1)\\ \partial_t \mathsf{f}_\epsilon(t,x) - \epsilon^{-1} S \partial_x \mathsf{f}_\epsilon(t,x) &= \epsilon^{-2} J_\epsilon \mathsf{f}_\epsilon(t,x) + g(t,x)\mathsf{w}, & (t,x) \in \mathcal{X}_T\ . \qquad (3.2) \end{aligned}$$

[1]The author thanks Prof. Wen-An Yong for the opportunity of a short discussion.

f_ϵ represents the $\mathcal{F}$-valued population function, where $\mathcal{F}$ stands for the set of real functions[2] over the D1P3 velocity space $\mathcal{S} = \{-1, 0, 1\}$ (confer table 1.1).
As far as boundary conditions are concerned, it is required that $\mathsf{f}_\epsilon \in \mathcal{C}^m_{\mathrm{per}}(\mathcal{X}_T, \mathcal{F})$ for some $m \in \mathbb{N},\ m \geq 1$. This means that f_ϵ and all its (mixed) derivatives up to the m^{th} order are continuous and L-periodic[3], i.e. $\partial_t^i \partial_x^j \mathsf{f}_\epsilon(t, 0) = \partial_t^i \partial_x^j \mathsf{f}_\epsilon(t, L)$ for $0 \leq i, j \leq m$ and $i + j \leq m$. We suppose that the initial condition $v_0 : \mathcal{X} \to \mathbb{R}$ and the source $g : \mathcal{X}_T \to \mathbb{R}$ are L-periodic and sufficiently smooth such that they induce the required regularity properties of f_ϵ.

The collision operator $J_\epsilon : \mathcal{F} \to \mathcal{F}$ with $J_\epsilon := \omega(E_\epsilon - I)$ is of BGK-type, where I denotes the identity on $\mathcal{F}$. The accompanying equilibrium operator is given by

$$E_\epsilon \mathsf{f} := (1 - \epsilon^2 \tau c)\, \langle \mathsf{f}, 1 \rangle\, \mathsf{w} \qquad \text{for } \mathsf{f} \in \mathcal{F}. \tag{3.3}$$

So E_ϵ and J_ϵ are linear operators. The reactivity $c \geq 0$ and the collision time $\tau := \omega^{-1} > 0$, which is the inverse of the collision frequency ω, are prescribed constants.

In the course of this chapter the following claim will be established: Let v_0, g, c, τ and the weight parameter $\theta > 1$ (occurring in w) be the same as above. Then the solution v of the initial value problem for the diffusion-reaction equation with periodic boundary conditions and $\nu := \tau/\theta$

$$\begin{array}{rcll} v(0,x) & = & v_0(x) & x \in \mathcal{X} \\ \partial_t v(t,x) - \nu \partial_x^2 v(t,x) + c v(t,x) & = & g(t,x) & (t,x) \in \mathcal{X}_T \end{array} \tag{3.4}$$

is approximated by the mass moment of the population function f_ϵ:

$$v(t,x) \;=\; \lim_{\epsilon \downarrow 0} \langle \mathsf{f}_\epsilon(t,x), 1 \rangle\,.$$

Therefore we refer to the above IVP as the (macroscopic) limit or target problem. We will refine this assertion further below providing a convergence rate *inter alia.*

Before we enter into details, let us motivate what makes the limit of the scaled lattice-Boltzmann equation (3.2) interesting and even surprising. A striking feature of equation (3.2) is the *singular* behavior if the scaling parameter ϵ tends to zero. This raises the question whether there is a reasonable limit and wether $\epsilon \to 0$ entails convergence of f_ϵ and of the entire equation (i.e. all occurring derivatives).
A priori, several scenarios are imaginable: the formally 'exploding' terms may cancel each other or are possibly canceled by an 'exploding' time derivative $\partial_t \mathsf{f}_\epsilon$. Alternatively, the singularities $\frac{1}{\epsilon}$ and $\frac{1}{\epsilon^2}$ may be compensated by f_ϵ such that the critical terms remain bounded as a whole.

---

[2]This space is isomorphic to $\mathbb{R}^3$; therefore f_ϵ can be considered as a tuple of three components. For instance, the homogeneous equilibrium (weight function) $\mathsf{w} \in \mathcal{F}$ is given by $(\frac{1}{2\theta}, \frac{\theta-1}{\theta}, \frac{1}{2\theta})$ where $\theta > 1$ is the weight parameter.

[3]Equivalently the problem could be formulated as a *full space* problem with the constraint of L-periodicity.

Multiplying equation (3.2) by ϵ^2 and sending it to zero seems to avoid the trouble with infinite or undefined terms. However, under the hypothesis that f_ϵ behaves decently (e.g. staying bounded), the differential equation becomes in the limit

$$0 = J_0\mathsf{f}(t,x) = \langle \mathsf{f}, 1\rangle \mathsf{w} - \mathsf{f},$$

which is dramatically different from (3.2). In contrast to the initial value problem having a unique solution, this purely algebraic limit equation permits an infinite multitude of solutions. Any function of the form $\mathsf{f}(t,x) = w(t,x)\mathsf{w}$, where $w : \mathcal{X}_T \to \mathbb{R}$ is some L-periodic function respecting the initial condition $w(0,x) = v_0(x)$, solves the limit problem.
What may not be grasped at first glance is the fact that w is not at all arbitrary but the solution of (3.4). Clearly, this reveals the singular nature of the equation.

In order to disentangle the confusing situation, we will make the ansatz, that f_ϵ depends in a *regular* manner on ϵ. This means, that we try to approximate the solution by a truncated power series in ϵ. Amazingly, this procedure provides a lot of insight, although only a special class of initial values can be captured[4].

3.1 Energy estimate and stability

Prior to the asymptotic consistency analysis, we establish a stability result. This states *grosso modo*, that the solution of the lattice-Boltzmann equation depends continuously on its initial value and a possible source term. The result is formulated in a way directly adapted to our purpose, guided by the following idea:
Consider an exact solution f_ϵ and an approximate solution $\hat{\mathsf{f}}_\epsilon$ of the lattice-Boltzmann equation. We can interpret the approximate solution as an exact solution with an additional source term corresponding to the residual. If stability holds, the two functions will stay close together during their evolution, provided they are initially almost equal and the residual is small. So asymptotic similarity can be inferred from stability, as soon as consistency is ensured, which means that the residual (i.e. the additional source driving $\hat{\mathsf{f}}_\epsilon$) vanishes with ϵ.

A crucial role in the stability proof comes up to the following lemma which relies on the special structure of the equilibrium operator. In the sequel let $W : \mathcal{F} \to \mathcal{F}$ be the linear operator[5] corresponding to a multiplication with the weight function w and let $B = W^{-1}$ its inverse.

Lemma 3.1. *BJ_ϵ is a* symmetric *(self-adjoint) operator with respect to the standard scalar product in $\mathcal{F}$. There exist orthogonal projectors Π_k, $k \in \{1,2,3\}$, with mutually disjoint ranges such that BJ_ϵ permits the following representation:*

$$BJ_\epsilon = -\sum_{k=1}^{3} \lambda_k \Pi_k = -\sum_{k=1}^{3} \lambda_k \Pi_k^\top \Pi_k \tag{3.5}$$

[4]It turns out that the zeroth order term of the expansion must be proportional to w. Therefore the initialization (3.1) has been chosen so restrictive. Since $\mathsf{f}_\epsilon(0,x) \in \mathcal{F}$, we have $\mathsf{f}_\epsilon(0,x) \in \operatorname{span}\{\mathsf{w}, \mathsf{s}, \frac{1}{\theta} - \mathsf{s}^2\}$. That is why the general initialization should be given by $\mathsf{f}_\epsilon(0,x) = a(x)\mathsf{w} + b(x)\mathsf{s} + c(x)\big(\frac{1}{\theta} - \mathsf{s}^2\big)$ with some functions $a, b, c : \mathcal{X} \to \mathbb{R}$. Moreover ϵ-dependent initializations are imaginable, e.g. in the form of truncated power series.

[5]If $\mathsf{f} = (\mathsf{f}_-, \mathsf{f}_0, \mathsf{f}_+)^\top$ then $W = \operatorname{diag}(\frac{1}{2\theta}, \frac{\theta-1}{\theta}, \frac{1}{2\theta})$.

Here the $-\lambda_k$*'s (with* $\lambda_k \geq 0$*) denote the* non-positive *eigenvalues.*

Proof: Since $BJ_\epsilon = \omega(BE_\epsilon - B)$ and due to the symmetry of B, the first part of the assertion is proved, if we verify, that BE_ϵ is symmetric too. Having $\mathsf{w} = \mathsf{w}\cdot\mathsf{1} = W\cdot\mathsf{1}$, we find

$$BE_\epsilon = (1-\epsilon^2\tau c)\langle\cdot,\mathsf{1}\rangle\ \underbrace{B\,W}_{=I}\cdot\mathsf{1} = (1-\epsilon^2\tau c)\langle\cdot,\mathsf{1}\rangle\mathsf{1}\ .$$

So, BE_ϵ is just a multiple of the rank-one orthogonal projector $\frac{1}{\langle \mathsf{1},\mathsf{1}\rangle}\langle\cdot,\mathsf{1}\rangle\,\mathsf{1}$, that is symmetric according to proposition 3.1. This implies that BE_ϵ has only a single non-zero eigenvalue given by $1-\epsilon^2\tau c$ with the associated eigenvector is $\mathsf{1}\in\mathcal{F}$.
By the *spectral theorem* for *symmetric* operators in Euclidean[6] vector spaces BJ_ϵ has an orthonormal basis of eigenvectors. Let Π_k for $k\in\{1,2,3\}$ denote the *orthogonal* projector[7] on the k'th eigenvector. Since Π_j,Π_k refer to different eigenvectors if $j\neq k$ we get $\Pi_j\Pi_k=0$. Using again proposition 3.1, the orthogonality of the projector Π_k entails its symmetry, i.e. $\Pi_k^\top = \Pi_k$, and thus $\Pi_k = \Pi_k^2 = \Pi_k^\top\Pi_k$. Designating with $-\lambda_k$ the eigenvalue of BJ_ϵ belonging to Π_k, the spectral decomposition of BJ_ϵ in (3.5) follows.
Let us finally check that all eigenvalues of BJ_ϵ are non-positive. Considering first the case $c=0$, we find

$$BJ_\epsilon = BJ_0 = \omega B(\langle\cdot,\mathsf{1}\rangle\mathsf{w} - I) = \omega\big(\langle\cdot,\mathsf{1}\rangle\mathsf{1} - B\big) = \omega\ \left[\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix} - \begin{pmatrix}2\theta&\cdot&\cdot\\ \cdot&\frac{\theta}{\theta-1}&\cdot\\ \cdot&\cdot&2\theta\end{pmatrix}\right]$$

Taking recourse to the explicit matrix representation of BJ_0, it can be verified that BJ_0 has the following non-positive eigenvalues:

$$0,\qquad -2\omega\theta,\qquad -\omega(2\theta+\tfrac{\theta}{\theta-1}-3)\qquad \text{with } \theta>1.$$

Hence BJ_0 is negative semi-definite. As the projector $-\epsilon^2\tau c\langle\cdot,\mathsf{1}\rangle\mathsf{1}$ is negative semi-definite as well (recall $\tau>0$ and $c\geq 0$) we conclude that $BJ_\epsilon = BJ_0 - \epsilon^2\tau c\langle\cdot,\mathsf{1}\rangle\mathsf{1}$ is also negative semi-definite. Hence positive eigenvalues are excluded. ■

Proposition 3.1. *Let P be a projector in a vector space V, i.e. $P=P^2$. Then the following two statements are equivalent:*

i) P is orthogonal, i.e. $\forall v,w\in V\quad \langle Pv,(I-P)w\rangle = \langle (I-P)v,Pw\rangle = 0$

ii) P is symmetric, i.e. $P=P^\top$.

Proof:

i) $\Rightarrow$ ii): Expanding the assumption yields: $\langle Pv,w\rangle = \langle Pv,Pw\rangle = \langle v,Pw\rangle$.
This implies: $0 = \langle Pv,w\rangle - \langle v,Pw\rangle = \langle (P-P^\top)v,w\rangle$.
Since this holds for all $v,w\in V$, we conclude: $P-P^\top = 0 \;\Leftrightarrow\; P=P^\top$.

ii) $\Rightarrow$ i): From the hypothesis we infer: $0 = P-P^2 = P^\top - P^\top P = P^\top(I-P)$.
Hence we find for all $v,w\in V$: $0 = \langle v,P^\top(I-P)w\rangle = \langle Pv,(I-P)w\rangle$. ■

[6]By an *Euclidean vector space* we understand a finite-dimensional, real vector space equipped with a scalar product.

[7]If $v\neq 0$ is a vector in an Euclidean vector space, then $\frac{1}{\langle v,v\rangle}\langle\cdot,v\rangle v$ is the *orthogonal* projection on v.

Let us make the notion of an approximate solution more precise by recalling the definition of *consistency.* An ϵ-dependent function $\hat{\mathsf{f}}_\epsilon \in \mathcal{C}^1_{\mathrm{per}}(\mathcal{X}_T, \mathcal{F})$ is said to be *consistent of order* α to the lattice-Boltzmann equation, if there exists an ϵ-dependent function $\mathsf{r}_\epsilon \in \mathcal{C}_{\mathrm{per}}(\mathcal{X}_T, \mathcal{F})$ – called the *residue function* – such that:

i) $\hat{\mathsf{f}}_\epsilon$ is an (exact) solution of the lattice-Boltzmann equation with the additional source term $\epsilon^\alpha \mathsf{r}_\epsilon$:

$$\partial_t \hat{\mathsf{f}}_\epsilon \;-\; \epsilon^{-1} S \partial_x \hat{\mathsf{f}}_\epsilon \;=\; \epsilon^{-2} J_\epsilon \hat{\mathsf{f}}_\epsilon \;+\; g\mathsf{w} \;+\; \epsilon^\alpha \mathsf{r}_\epsilon. \tag{3.6}$$

ii) r_ϵ is uniformly bounded with respect to ϵ:

$$\exists C > 0,\ \forall \epsilon \in (0,1] : \quad \int_0^T \int_{\mathcal{X}} \|\mathsf{r}_\epsilon(t,x)\|^2_{\mathcal{F}} \,\mathrm{d}x\,\mathrm{d}t \;<\; C. \tag{3.7}$$

We refer to $\tilde{\mathsf{r}}_\epsilon := \epsilon^\alpha \mathsf{r}_\epsilon$ as the *residual.* The residue function is, so to say, the *unscaled* residual of order $\mathrm{O}(1)$.

Now we can formulate the central stability result of this chapter:

Theorem 3.1. *Let $\mathsf{f} \in \mathcal{C}^1_{\mathrm{per}}(\mathcal{X}_T; \mathcal{F})$ be a solution of the lattice-Boltzmann equation. Suppose that $\hat{\mathsf{f}}_\epsilon \in \mathcal{C}^1_{\mathrm{per}}(\mathcal{X}_T; \mathcal{F})$ is consistent to the lattice-Boltzmann equation of order $\alpha > 0$ and*

$$\|\mathsf{f}_\epsilon(0,\cdot) - \hat{\mathsf{f}}_\epsilon(0,\cdot)\|_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} < K_0 \epsilon^\alpha \tag{3.8}$$

for some constant K_0. Then there exists a constant $K > 0$ independent of $\epsilon \in (0,1]$ such that

$$\sup_{t\in[0,T]} \|\mathsf{f}_\epsilon(t,\cdot) - \hat{\mathsf{f}}_\epsilon(t,\cdot)\|_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} < K\epsilon^\alpha \,.$$

Proof: For convenience let us introduce the *deviation* $\mathsf{d}_\epsilon := \mathsf{f}_\epsilon - \hat{\mathsf{f}}_\epsilon$. As our equilibrium operator is linear, the evolution equation for d_ϵ is simply derived by subtracting (3.6) from (3.2).

$$\partial_t \mathsf{d}_\epsilon + \tfrac{1}{\epsilon} S \partial_x \mathsf{d}_\epsilon = \tfrac{1}{\epsilon^2}\, J_\epsilon \mathsf{d}_\epsilon - \epsilon^\alpha \mathsf{r}_\epsilon \tag{3.9}$$

Moreover note, that B is a positive operator in $\mathcal{F}$, whose operator norm satisfies $1 \le \min\{2\theta, \frac{\theta}{\theta-1}\} \le \|B\| \le \max\{2\theta, \frac{\theta}{\theta-1}\}$ if $\theta > 1$ and $\|B\| = 2$ for $\theta = 1$. Therefore the norm defined by $|\cdot|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} := \int_{\mathcal{X}} \langle B\cdot,\cdot\rangle$ is equivalent to the standard norm in $\mathcal{L}^2(\mathcal{X};\mathcal{F})$ and we have $\|\cdot\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \le |\cdot|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \le \|B\|\,\|\cdot\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})}$.

Now the proof is done by an *energy estimate* of the deviation. This is obtained by taking the scalar product of equation (3.9) with $B\mathsf{d}_\epsilon$ and integrating over the spatial domain $\mathcal{X}$.

$$\int_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, \partial_t \mathsf{d}_\epsilon\rangle \;+\; \frac{1}{\epsilon}\int_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, S\partial_x \mathsf{d}_\epsilon\rangle = \frac{1}{\epsilon^2}\int_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, J_\epsilon \mathsf{d}_\epsilon\rangle \;-\; \epsilon^\alpha \int_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, \mathsf{r}_\epsilon\rangle \tag{3.10}$$

Let us consider the second integral on the left hand side, which is given in more detail by:

$$\int_0^L \big\langle B\mathsf{d}_\epsilon(t,x), S\partial_x \mathsf{d}_\epsilon(t,x)\big\rangle \mathrm{d}x = \frac{1}{2}\int_0^L \frac{\mathrm{d}}{\mathrm{d}x}\big\langle B\mathsf{d}_\epsilon(t,x), S\mathsf{d}_\epsilon(t,x)\big\rangle \mathrm{d}x = 0 \tag{3.11}$$

Since $\mathsf{f}_\epsilon, \hat{\mathsf{f}}_\epsilon$ and consequently d_ϵ are L-periodic, i.e. $\mathsf{d}_\epsilon(t,0) = \mathsf{d}_\epsilon(t,L)$, the integral vanishes. Remarking that $\langle B\mathsf{d}_\epsilon, J_\epsilon \mathsf{d}_\epsilon\rangle = \langle \mathsf{d}_\epsilon, BJ_\epsilon \mathsf{d}_\epsilon\rangle$, because B is diagonal as a multiplication operator and *a fortiori* symmetric, we can exploit lemma 3.1. So equation (3.10) becomes

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int\limits_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, \mathsf{d}_\epsilon\rangle = -\frac{1}{\epsilon^2}\sum_{k=1}^{3}\lambda_k \int\limits_{\mathcal{X}} \langle \Pi_k \mathsf{d}_\epsilon, \Pi_k \mathsf{d}_\epsilon\rangle \; - \; \epsilon^\alpha \int\limits_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, \mathsf{r}_\epsilon\rangle$$

which corresponds to

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\,|\mathsf{d}_\epsilon|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;=\; -\frac{1}{\epsilon^2}\sum_{k=1}^{3}\lambda_k \|\Pi_k \mathsf{d}_\epsilon\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;-\; \epsilon^\alpha \int\limits_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, \mathsf{r}_\epsilon\rangle \,.$$

The expression resulting from the integral over the collision term (the first one of the right hand side) is obviously negative. Thus we can estimate applying additionally the Cauchy-Schwarz inequality:

$$\begin{aligned}
\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\,|\mathsf{d}_\epsilon|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;&\le\; -\,\epsilon^\alpha \int\limits_{\mathcal{X}} \langle B\mathsf{d}_\epsilon, \mathsf{r}_\epsilon\rangle \;\le\; \epsilon^\alpha\,|\mathsf{d}_\epsilon|_{\mathcal{L}^2(\mathcal{X};\mathcal{F})}\,|\mathsf{r}_\epsilon|_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \\
&\le\; \epsilon^\alpha\Big(\epsilon^{-\alpha}|\mathsf{d}_\epsilon|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} + \epsilon^\alpha |\mathsf{r}_\epsilon|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})}\Big) \\
&=\; |\mathsf{d}_\epsilon|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;+\; \epsilon^{2\alpha}\,|\mathsf{r}_\epsilon|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})}
\end{aligned}$$

In the second line we made use of the inequality $ab = \epsilon^{-\alpha/2}a\epsilon^{\alpha/2}b \le \epsilon^{-\alpha}a^2 + \epsilon^\alpha b^2$ for arbitrary reals a, b, ϵ. Finally, we apply Gronwall's inequality[8]:

$$|\mathsf{d}_\epsilon(t,\cdot)|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;\le\; \exp(t)\,|\mathsf{d}_\epsilon(0,.)|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} + \epsilon^{2\alpha}\exp(t)\|B\| \int_0^t \|\mathsf{r}_\epsilon(\sigma,\cdot)\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})}\mathrm{d}\sigma$$

By the initial condition (3.8) we have $|\mathsf{d}_\epsilon(0,\cdot)|_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \le \|B\|K_0\epsilon^\alpha$. From the definition of consistency we know, that the integral is dominated by some constant $C > 0$ for $0 \le t \le T$. So we end up with

$$\|\mathsf{d}_\epsilon(t,\cdot)\|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \le |\mathsf{d}_\epsilon(t,\cdot)|^2_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \le \exp(T)\|B\|K_0^2\epsilon^{2\alpha} + \epsilon^{2\alpha}\exp(T)\|B\|C$$

for $t \in \mathcal{T}$, which yields the assertion with $K := \sqrt{\exp(T)\|B\|\big(K_0^2 + C\big)}$. ■

Suppose that f_ϵ and $\hat{\mathsf{f}}_\epsilon$ are $\mathcal{C}^m_{\text{per}}$-functions with $m > 1$. Then the residual is in $\mathcal{C}^{m-1}_{\text{per}}(\mathcal{X}_T, \mathcal{F})$ and we are allowed to differentiate equation (3.9) up to $m-1$ times. Due to its linearity and the constancy of the parameters ω and c, its form will

[8] Gronwall's lemma in differential form claims, that the non-negative function $\eta \in \mathcal{C}^1([0,T])$ satisfies the estimate

$$\eta(t) \le \exp\Big(\int_0^t \phi(s)\mathrm{d}s\Big)\;\Big[\eta(0) + \int_0^t \psi(s)\mathrm{d}s\Big] \qquad \text{for } \; t \in [0,T],$$

if the differential inequality $\frac{\mathrm{d}}{\mathrm{d}t}\eta(t) \le \phi(t)\eta(t) + \psi(t)$ holds true with the nonnegative functions $\phi, \psi \in \mathcal{L}^1([0,T])$. For further reference see [17] page 624.

not be altered. Therefore we get also the asymptotic similarity of the derivatives, simply by multiplying the equation

$$\partial_t(\partial_x^k \mathsf{d}_\epsilon) + \epsilon^{-1} S\partial_x(\partial_x^k \mathsf{d}_\epsilon) = \epsilon^{-2} J\, \partial_x^k \mathsf{d}_\epsilon - \epsilon^\alpha \partial_x^k \mathsf{r}_\epsilon, \qquad 1 \le k \le m-1$$

with $B\partial_x^k \mathsf{d}_\epsilon$ and performing exactly the same steps as in the proof of the previous theorem.

Corollary 3.1. *Suppose the assumptions of theorem 3.1 are satisfied. If additionally* $\mathsf{f}_\epsilon, \hat{\mathsf{f}}_\epsilon \in \mathcal{C}^m_{\text{per}}(\mathcal{X}_T, \mathcal{F})$ *then the two functions are for* $\epsilon \downarrow 0$ *asymptotically similar in* $\mathcal{C}\big(\mathcal{T}, \mathcal{H}^{m-1}(\mathcal{X}, \mathcal{F})\big)$ *with a residual of order* α*, i.e. there is a constant* $K > 0$ *such that*

$$\sup_{t\in\mathcal{T}} \|\partial_x^k \mathsf{f}_\epsilon(t,\cdot) - \partial_x^k \hat{\mathsf{f}}_\epsilon(t,\cdot)\|_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} < K\epsilon^\alpha \qquad 0 \le k \le m-1 \,.$$

Sobolev's embedding theorem[9] tells us, that any weakly differentiable function on an interval of the real axis is continuous (strictly speaking its equivalence class with respect to *almost everywhere equality* contains a continuous representative). Moreover its supremum-norm can be estimated by the $\mathcal{L}^2$-norm of the function and its weak derivative. Especially this entails in our case for $m \ge 2$

$$\begin{aligned} &\sup_{t\in\mathcal{T}} \|\partial_x^k \mathsf{f}_\epsilon(t,\cdot) - \partial_x^k \hat{\mathsf{f}}_\epsilon(t,\cdot)\|_{\mathcal{L}^2(\mathcal{X};\mathcal{F})} \;=\; \mathrm{O}(\epsilon^\alpha)\,, \qquad 0 \le k \le m-1 \\ &\qquad \Longrightarrow \qquad \sup_{t\in\mathcal{T}} \sup_{x\in\mathcal{X}} \|\partial_x^k \mathsf{f}_\epsilon(t,x) - \partial_x^k \hat{\mathsf{f}}_\epsilon(t,x)\|_{\mathcal{F}} \;=\; \mathrm{O}(\epsilon^\alpha)\,, \qquad 0 \le k \le m-2 \end{aligned}$$

and allows us to formulate a second corollary.

Corollary 3.2. *Under the hypothesis of corollary 3.1,* f_ϵ *and* $\hat{\mathsf{f}}_\epsilon$ *are asymptotically similar in* $\mathcal{C}^{m-2}_{\text{per}}(\mathcal{X}_T, \mathcal{F})$*, i.e. with respect to the supremum-norm over* $\mathcal{X}_T = \mathcal{T} \times \mathcal{X}$.

3.2 Regular expansion and consistency

This section deals with the explicit construction of a comparison (prediction) function in the form of a truncated *regular expansion* for some arbitrary but fixed $n \in \mathbb{N}_0$:

$$\hat{\mathsf{f}}^{[n]}_\epsilon := \mathsf{f}^{(0)} + \epsilon\, \mathsf{f}^{(1)} + \ldots + \epsilon^n\, \mathsf{f}^{(n)} \,. \tag{3.12}$$

In general it is not possible to choose the *asymptotic order functions* $\mathsf{f}^{(k)}$, $0 \le k \le n$, in such a way that the comparison function becomes an exact solution of the lattice-Boltzmann equation (3.2). Hence we try to determine the $\mathsf{f}^{(k)}$'s so that the *residual*[10]

$$\partial_t \hat{\mathsf{f}}^{[n]}_\epsilon \;-\; \epsilon^{-1} S\partial_x \hat{\mathsf{f}}^{[n]}_\epsilon \;-\; \epsilon^{-2} J_0 \hat{\mathsf{f}}^{[n]}_\epsilon \;+\; c\langle \hat{\mathsf{f}}^{[n]}_\epsilon, \mathsf{1}\rangle \mathsf{w} \;-\; g\mathsf{w} \;=:\; \tilde{\mathsf{r}}^{[n]}_\epsilon$$

is asymptotically minimized for $\epsilon \downarrow 0$. $\tilde{\mathsf{r}}^{[n]}_\epsilon$ inherits the structure of a finite power series starting however with ϵ^{-2}. In order to make the residual vanish as fast as possible for $\epsilon \to 0$, we equate terms of equal power in ϵ with 0. The $\mathsf{f}^{(k)}$'s must be

[9] [4] page 129

[10] This expression is obtained by plugging the comparison function $\hat{\mathsf{f}}^{[n]}_\epsilon$ into the lattice-Boltzmann equation and putting all terms on the left hand side.

determined in such a manner that the obtained equations are satisfied. *A priori*, the residual comprises $n+3$ different orders from ϵ^{-2} to ϵ^n which yields $n+3$ different equations. As we have to determine $n+1$ asymptotic order functions, we expect that all equations but two can be fulfilled by an appropriate choice of the $\mathsf{f}^{(k)}$'s. We pick the equations referring to ϵ^{-2} up to ϵ^{n-2} which ensures the residual to be of magnitude $\mathrm{O}(\epsilon^{n-1})$. So the subsequent equations result

$$\left.\begin{aligned} \epsilon^{-2} &: & 0 &= J_0\mathsf{f}^{(0)} \\ \epsilon^{-1} &: & S\partial_x\mathsf{f}^{(0)} &= J_0\mathsf{f}^{(1)} \\ \epsilon^{k} &: & \partial_t\mathsf{f}^{(k)} + S\partial_x\mathsf{f}^{(k+1)} &= J_0\mathsf{f}^{(k+2)} - c\langle\mathsf{f}^{(k)},1\rangle\mathsf{w} + g\delta_{k0}\mathsf{w} \end{aligned}\right\} \tag{3.13}$$

where the index k runs through the set $\{0,...,n-2\}$. Apart from the two first equations, the obtained equations share the same algebraic structure.

If we achieve to satisfy the above equations by a special selection of $\mathsf{f}^{(0)},\ldots,\mathsf{f}^{(n)}$, then only terms of the order $n-1$ and n will survive.

$$\epsilon^{n-1}\partial_t\mathsf{f}^{(n-1)} + \epsilon^{n-1}S\partial_x\mathsf{f}^{(n)} + \epsilon^n\partial_t\mathsf{f}^{(n)} = -\epsilon^{n-1}c\langle\mathsf{f}^{(n-1)},1\rangle\mathsf{w} - \epsilon^n c\langle\mathsf{f}^{(n)},1\rangle\mathsf{w} + \tilde{\mathsf{r}}_\epsilon^{[n]}$$

These remaining terms have to be compensated by the actual residual, if equality shall hold. Resolving thus for $\tilde{\mathsf{r}}_\epsilon^{[n]}$ yields

$$\tilde{\mathsf{r}}_\epsilon^{[n]} = \epsilon^{n-1}\Big\{\underbrace{\partial_t\mathsf{f}^{(n-1)} + S\partial_x\mathsf{f}^{(n)} + \epsilon\partial_t\mathsf{f}^{(n)} + c\langle\mathsf{f}^{(n-1)},1\rangle\mathsf{w} + \epsilon c\langle\mathsf{f}^{(n)},1\rangle\mathsf{w}}_{\text{residue function } \mathsf{r}_\epsilon^{[n]}}\Big\}\,. \tag{3.14}$$

If $\mathsf{f}^{(n-1)}$ and $\mathsf{f}^{(n)}$ are sufficiently smooth, the residue function is uniformly bounded for $\epsilon \in (0,1]$. Hence we find the ansatz (3.12) consistent of order $n-1$ provided that we manage to solve system (3.13).

For ease of notation we define two differential operators

$$D_t := -\tau\partial_t\,, \qquad D_x := -\tau S\partial_x$$

mapping $\mathcal{C}^m_{\mathrm{per}}(\mathcal{X}_T;\mathcal{F})$ to $\mathcal{C}^{m-1}_{\mathrm{per}}(\mathcal{X}_T;\mathcal{F})$. Let us collect those terms in (3.13) on the left hand side, which contain the highest occurring asymptotic order function. By introducing the linear map

$$M:\mathcal{F}\to\mathcal{F}, \qquad M := I - \langle\cdot,1\rangle\mathsf{w} = -\tau J_0 \tag{3.15}$$

the equations (3.13) are recast in the form:

$$M\,\mathsf{f}^{(0)} = 0 \tag{3.16}$$

$$M\,\mathsf{f}^{(1)} = D_x\mathsf{f}^{(0)} \tag{3.17}$$

$$M\,\mathsf{f}^{(k+2)} = D_t\mathsf{f}^{(k)} + D_x\mathsf{f}^{(k+1)} - \tau c\langle\mathsf{f}^{(k)},1\rangle\mathsf{w} + \tau g\delta_{k0}\mathsf{w} \tag{3.18}$$

Before we continue, a quick study of the properties of M seems to be recommended.

Lemma 3.2. *The linear operator M, as defined above, is a projector. For given $\mathsf{y}\in\mathcal{F}$ the equation $M\mathsf{x}=\mathsf{y}$ is only solvable if $\langle\mathsf{y},1\rangle = 0$. Its general solution is given by* $\mathsf{x} = \ker M + \mathsf{y} = \mathbb{R}\mathsf{w} + \mathsf{y}$.

Proof: Since $\langle \mathsf{w}, 1\rangle = 1$ we get:

$$\begin{aligned} M^2\mathsf{x} &= M\mathsf{x} - \langle M\mathsf{x}, 1\rangle\mathsf{w} \;=\; \mathsf{x} - \langle \mathsf{x}, 1\rangle\mathsf{w} - \big\langle \mathsf{x} - \langle \mathsf{x}, 1\rangle\mathsf{w}, 1\big\rangle\mathsf{w} \\ &= \mathsf{x} - 2\langle \mathsf{x}, 1\rangle\mathsf{w} + \langle \mathsf{x}, 1\rangle\langle \mathsf{w}, 1\rangle\mathsf{w} \;=\; \mathsf{x} - \langle \mathsf{x}, 1\rangle\mathsf{w} \\ &= M\mathsf{x} \end{aligned}$$

So, we see that M is a projector, indeed. Its nullspace is obviously given by $\ker M = \operatorname{span}\{\mathsf{w}\}$. Hence M is not invertible and the equation admits only a solution if $\mathsf{y} \in \operatorname{range} M$. By *Fredholm's alternative* for Euclidean vector spaces (see the proposition below) we have $\operatorname{range} M = (\ker M^\top)_\perp$. Let us therefore compute the adjoint operator $M^\top$:

$$\begin{aligned} \langle M\mathsf{x}, \mathsf{z}\rangle &= \big\langle \mathsf{x} - \langle \mathsf{x}, 1\rangle\mathsf{w},\ \mathsf{z}\big\rangle \;=\; \langle \mathsf{x}, \mathsf{z}\rangle - \langle \mathsf{x}, 1\rangle\langle \mathsf{w}, \mathsf{z}\rangle \;=\; \langle \mathsf{x}, \mathsf{z}\rangle - \langle \mathsf{x}, \langle \mathsf{z}, \mathsf{w}\rangle 1\rangle \\ &= \big\langle \mathsf{x},\ \mathsf{z} - \langle \mathsf{z}, \mathsf{w}\rangle 1\big\rangle \;=\; \langle \mathsf{x}, M^\top\mathsf{z}\rangle \qquad \Rightarrow \qquad M^\top = I - \langle \cdot, \mathsf{w}\rangle 1 \end{aligned}$$

Evidently, we have $\ker M^\top = \operatorname{span}\{1\}$, which yields the necessary solvability condition $\langle y, 1\rangle = 0 \;\Leftrightarrow\; \mathsf{y} \in (\ker M^\top)_\perp$. Furthermore being a projector, M acts like the identity[11] on its range. So $\mathsf{y} \in \operatorname{range} M$ is a specific solution. As the general solution is given as a sum of a particular solution plus an arbitrary solution of the homogeneous equation being an element of $\ker M = \operatorname{span}\{\mathsf{w}\}$, we find the asserted formula verified. ■

Proposition 3.2. *Let $A : V \to V$ be an endomorphism on an Euclidean vector space V. Then the range (image) of A is the orthogonal complement to the kernel (nullspace) of the adjoint operator $A^\top$, i.e.:* $\operatorname{range} A = (\ker A^\top)_\perp$.

Proof: Let $w \in \ker A^\top$ and $u \in \operatorname{range} A$, i.e. there exists a $v \in V$ with $u = Av$. Then we have $u \perp v$, since

$$\langle u, w\rangle \;=\; \langle Av, w\rangle \;=\; \langle v, A^\top w\rangle \;=\; \langle v, 0\rangle \;=\; 0\ .$$

This shows $\operatorname{range} A \subset (\ker A^\top)_\perp$. In order to prove the equality we must take recourse to the finite dimensionality of V. Therefore it is possible to consider a matrix representation of A and $A^\top$. We note that a matrix and its transpose have the same rank, since *row rank* and *column rank* are equal for any matrix. Thus we conclude: $\dim(\operatorname{range} A) = \dim(\operatorname{range} A^\top)$. Moreover the following relations hold true:

$$\begin{array}{lrcl} \text{i)} & V \;=\; \ker A^\top \oplus (\ker A^\top)_\perp & \Rightarrow & \dim\big((\ker A^\top)_\perp\big) \;=\; \dim V - \dim(\ker A^\top) \\ \text{ii)} & \dim V = \dim(\ker A^\top) + \dim(\operatorname{range} A^\top) & \Leftrightarrow & \dim(\operatorname{range} A^\top) \;=\; \dim V - \dim(\ker A^\top) \end{array}$$

Hence we get: $\dim(\operatorname{range} A) = \dim\big((\ker A^\top)_\perp\big)$. So $\operatorname{range} A$ is a subspace of $(\ker A^\top)_\perp$ having the same dimension. This enforces, however, the equality of the two spaces. ■

By means of the preceding lemma, equations (3.16),(3.17) and (3.18) are easily solved. As $\mathsf{f}^{(0)}$ is in the kernel of M according to equation (3.16), we conclude

$$\mathsf{f}^{(0)} = u^{(0)}\mathsf{w}\ , \tag{3.19}$$

with some $u^{(0)} : \mathcal{X}_T \to \mathbb{R}$ to be specified more precisely later. If we take this expression for $\mathsf{f}^{(0)}$ and remember $\langle S\mathsf{w}, 1\rangle$=0, we notice immediately, that the right hand

[11] Let $\mathsf{y} \in \operatorname{range} M$ with $\mathsf{y} = M\mathsf{x}$. The defining property of projectors, $M^2 = M$, implies $\mathsf{y} - M\mathsf{y} = M\mathsf{x} - M^2\mathsf{x} = M\mathsf{x} - M\mathsf{x} = 0$. Therefore M acts like the identity on $\operatorname{range} M$.

side of equation (3.17) fulfills the solvability condition $\langle D_x \mathsf{f}^{(0)}, 1\rangle = 0$ independently of $u^{(0)}$. Therefore we obtain

$$\mathsf{f}^{(1)} = u^{(1)}\mathsf{w} + D_x \mathsf{f}^{(0)} \tag{3.20}$$

with another unknown function $u^{(1)}$ remaining also undetermined so far. Similarly, we proceed with the higher asymptotic orders. This yields for $k \in \{0, ..., l-2\}$ the following recursion relation

$$\mathsf{f}^{(k+2)} = u^{(k+2)}\mathsf{w} + D_t \mathsf{f}^{(k)} + D_x \mathsf{f}^{(k+1)} - \tau c \langle \mathsf{f}^{(k)}, 1\rangle \mathsf{w} + \tau g \delta_{k0} \mathsf{w} \ , \tag{3.21}$$

provided the solvability condition

$$\left\langle D_t \mathsf{f}^{(k)} + D_x \mathsf{f}^{(k+1)} - \tau c \langle \mathsf{f}^{(k)}, 1\rangle \mathsf{w} + \tau g \delta_{k0} \mathsf{w},\ 1 \right\rangle \ = \ 0 \tag{3.22}$$

is fulfilled. Obviously, for any order a new scalar function comes into play. To all appearances, the undetermined $\mathcal{F}$-valued asymptotic order $\mathsf{f}^{(k)}$ is replaced by another unknown but scalar-valued function $u^{(k)}$, that seems to be quite arbitrary at first glance. Therefore the impression imposes, that we are not winning much, since the complete meaning of these scalar-valued functions remains a little bit vague.
So far it is only known that $u^{(k)} = \langle \mathsf{f}^{(k)}, 1\rangle$ due to the solvability condition; thus $u^{(k)}$ is the mass moment of $\mathsf{f}^{(k)}$. However, we already anticipate, that the solvability condition inflicts certain constraints on the $u^{(k)}$'s, which make them not at all arbitrary. In order to demonstrate this, it is necessary, that we substitute in equation (3.21) the asymptotic orders recursively by their corresponding mass moments.

To do so, it is convenient to define the following family of differential operators by recurrence, which spares the cumbersome writing of lengthy terms.

$$P_k := 0 \quad \text{for } k < 0, \quad P_0 := I, \quad P_k := D_x P_{k-1} + D_t P_{k-2} \quad \text{for } k \geq 1 \tag{3.23}$$

We consider P_k as a differential operator from $\mathcal{C}^{m+k}_{\text{per}}(\mathcal{X}_T; \mathcal{F})$ to $\mathcal{C}^{m}_{\text{per}}(\mathcal{X}_T; \mathcal{F})$ for $m \in \mathbb{N}_0$. In the sequel of this section we tacitly assume enough smoothness, so that all occurring differential operators can be applied in a meaningful sense.

Remark 3.1. (Parity) An essential property of the P_k's is, that they exhibit a well defined parity, if we interpret them as polynomials with respect to the operator S. By a simple induction it is verified for $j \in \mathbb{N}_0$ that even powers of S including $S^0 = I$ occur only in P_{2j}, whereas P_{2j+1} contains odd ones. Hence the relations $S^{2j+1} = S$ and $\langle S\mathsf{w}, 1\rangle = 0$ entail

$$\langle P_{2j+1} f \mathsf{w}, 1\rangle = 0 \tag{3.24}$$

for any function $f \in \mathcal{C}^{2j+1}_{\text{per}}(\mathcal{X}_T; \mathbb{R})$.

Remark 3.2. (Homogeneity) Another characteristic feature should not stay unmentioned, for it will become of much use further down. From a formal point of view, we can consider the differential operator P_k as a polynomial in D_t and D_x. Let us therefore write $P_k = P_k(D_t, D_x)$. Moreover, let us replace the differential operators D_t, D_x by the formal variables χ, ξ, so that we can define the polynomial $p_k(\chi,\xi) := P_k(\chi^2,\xi)$. Evidently, these polynomials obey the recursion relation

$$p_0(\chi,\xi) = 1, \qquad p_1(\chi,\xi) = \xi, \qquad p_k(\chi,\xi) := \xi\, p_{k-1}(\chi,\xi) + \chi^2\, p_{k-2}(\chi,\xi)\ .$$

By an induction argument one can check, that these polynomials are *homogeneous* and their degree of homogeneity is given by their index. Thus $p_k(\lambda\chi,\lambda\xi) = \lambda^k p_k(\chi,\xi)$. Hence, the following consequence ensues: If we assign to each monomial of $P_k(D_t,D_x)$ a degree given by *twice the order of the time derivative plus the order of the spatial derivative*, it turns out, that we obtain for every monomial of P_k the same degree. Introducing the function spaces

$$\mathscr{C}^n_{\mathrm{per}}(\mathcal{X}_T;\mathcal{F}) := \left\{ f \in \mathcal{C}_{\mathrm{per}}(\mathcal{X}_T;\mathcal{F}) :\ \partial_t^i \partial_x^j f \in \mathcal{C}_{\mathrm{per}}(\mathcal{X}_T;\mathcal{F})\ \text{ if }\ 2i + j \le n \right\}\ ,$$

it is now straightforward to see, that P_k maps $\mathscr{C}^{n+k}_{\mathrm{per}}(\mathcal{X}_T;\mathcal{F})$ to $\mathscr{C}^n_{\mathrm{per}}(\mathcal{X}_T;\mathcal{F})$.

The objective of the next lemma is to resolve the recursion relation (3.21) for the asymptotic order functions and to express them in terms of their mass moments.

Lemma 3.3. $\mathsf{f}^{(k+2)}, 0 \le k \le n-2$, *admits the following explicit representation*

$$\mathsf{f}^{(k+2)} = \sum_{j=0}^{k+2} P_{k+2-j} u^{(j)} \mathsf{w}\ -\ \tau c \sum_{j=0}^{k} P_{k-j} u^{(j)} \mathsf{w}\ +\ \tau P_k g \mathsf{w}. \tag{3.25}$$

Proof: The proof is done by induction.
Start: It is easily checked that the formula holds even true for $\mathsf{f}^{(0)}$ and $\mathsf{f}^{(1)}$, i.e. for $k \in \{-2,-1\}$, if we set $u^{(j)} = 0$ for $j < 0$. Using $\delta_{00} = 1$ we get for $k = 0$:

$$\begin{aligned}
\mathsf{f}^{(2)} &= u^{(2)}\mathsf{w} + D_t \mathsf{f}^{(0)} + D_x \mathsf{f}^{(1)} - \tau c \langle \mathsf{f}^{(0)}, 1\rangle \mathsf{w} + \tau g \mathsf{w} \\
&= u^{(2)}\mathsf{w} + D_t u^{(0)}\mathsf{w} + D_x u^{(1)}\mathsf{w} + D_x^2 u^{(0)}\mathsf{w} - \tau c u^{(0)} \langle \mathsf{w}, 1\rangle \mathsf{w} + \tau g \mathsf{w} \\
&= P_0 u^{(2)}\mathsf{w} + P_1 u^{(1)}\mathsf{w} + P_2 u^{(0)}\mathsf{w} - \tau c P_0 u^{(0)}\mathsf{w} + \tau P_0 g \mathsf{w}
\end{aligned}$$

Thus we obtain the claimed formula also in this case.
Step: Assuming formula (3.25) to be true for $k \ge 0$, we show that it is also retrieved for $k+1 > 0$. Note that the Kroncker delta vanishes then. Setting in equation (3.21) k to $k+1$ yields:

$$\mathsf{f}^{(k+3)} \stackrel{!}{=} u^{(k+3)}\mathsf{w} + D_x \mathsf{f}^{(k+2)} + D_t \mathsf{f}^{(k+1)} - \tau c u^{(k+1)}\mathsf{w}.$$

Now the induction hypothesis is used to substitute $\mathsf{f}^{(k+1)}$ and $\mathsf{f}^{(k+2)}$ by analogous formulas:

$$\begin{aligned}
\mathsf{f}^{(k+3)} = {} & u^{(k+3)}\mathsf{w} + D_x \left[\sum_{j=0}^{k+2} P_{k+2-j} u^{(j)}\mathsf{w} - \tau c \sum_{j=0}^{k} P_{k+1-j} u^{(j)}\mathsf{w} + \tau P_k g \mathsf{w} \right] \\
& + D_t \left[\sum_{j=0}^{k+1} P_{k+2-j} u^{(j)}\mathsf{w} - \tau c \sum_{j=0}^{k-1} P_{k+1-j} u^{(j)}\mathsf{w} + \tau P_{k-1} g \mathsf{w} \right] - \tau c u^{(k+1)}\mathsf{w}.
\end{aligned}$$

Next, corresponding sums inside each of the square brackets are rewritten such that the index j runs everywhere over the same range. So the sums can be condensed:

$$\begin{aligned}\mathsf{f}^{(k+3)} = \;& P_0 u^{(k+3)}\mathsf{w} + P_1 u^{(k+2)}\mathsf{w} + \sum_{j=0}^{k+1}\big(D_x P_{k+2-j} + D_t P_{k-1-j}\big)u^{(j)}\mathsf{w} \\ & - \tau c P_0 u^{(k+1)}\mathsf{w} - \tau c P_1 u^k \mathsf{w} - \tau c \sum_{j=0}^{k-1}\big(D_x P_{k-j} + D_t P_{k-1-j}\big)u^{(j)}\mathsf{w} \\ & + \tau\big(D_x P_k + D_t P_{k-1}\big)g\mathsf{w}.\end{aligned}$$

Finally the recursion relation (3.23) is applied

$$\mathsf{f}^{(k+3)} = \sum_{j=0}^{k+3} P_{k+3-j}u^{(j)}\mathsf{w} - \tau c \sum_{j=0}^{k+1} P_{k+1-j}u^{(j)}\mathsf{w} + \tau P_{k+1} g\mathsf{w},$$

to recover formula (3.25) *mutatis mutandis* where k is replaced by $k+1$. ■

At last, we are prepared to analyze the solvability condition. Equation (3.25) is equivalent to

$$\mathsf{f}^{(k+2)} = u^{(k+2)}\mathsf{w} + \sum_{j=0}^{k+1} P_{k+2-j}u^{(j)}\mathsf{w} \;-\; \tau c \sum_{j=0}^{k} P_{k-j}u^{(j)}\mathsf{w} \;+\; \tau P_k g\mathsf{w}\;.$$

Therefore the solvability condition becomes

$$\Big\langle \sum_{j=0}^{k+1} P_{k+2-j}u^{(j)}\mathsf{w} \;-\; \tau c \sum_{j=0}^{k} P_{k-j}u^{(j)}\mathsf{w} \;+\; P_k g\mathsf{w},\; 1\Big\rangle \;=\; 0\;.$$

Taking all addends containing $u^{(k+1)}$ and $u^{(k)}$ out of the sums, we obtain

$$\begin{aligned}&\langle P_1 u^{(k+1)}\mathsf{w}, 1\rangle \;+\; \langle P_2 u^{(k)}\mathsf{w}, 1\rangle \;-\; \tau c\langle P_0 u^{(k)}\mathsf{w}, 1\rangle \;= \\ &\qquad = \;-\Big\langle \sum_{j=0}^{k-1} P_{k+2-j}u^{(j)}\mathsf{w} - \tau c\sum_{j=0}^{k-1} P_{k-j}u^{(j)}\mathsf{w} + P_k g\mathsf{w},\; 1\Big\rangle.\end{aligned}$$

Using the relations

i) $\langle P_1 u^{(k+1)}\mathsf{w}, 1\rangle \;=\; \langle D_x u^{(k+1)}\mathsf{w}, 1\rangle \;=\; -\tau\partial_x u^{(k+1)}\langle S\mathsf{w}, 1\rangle \;=\; 0$

ii) $\langle P_2 u^{(0)}\mathsf{w}, 1\rangle \;=\; \langle (D_t + D_x^2)u^{(0)}\mathsf{w}, 1\rangle \;=\; -\tau\partial_t u^{(0)}\underbrace{\langle \mathsf{w}, 1\rangle}_{=1} - \tau^2\partial_x^2 u^{(0)}\underbrace{\langle S^2\mathsf{w}, 1\rangle}_{=1/\theta}$

we arrive at a diffusion-reaction equation for $u^{(k)}$ with a source $\mathfrak{R}_k$:

$$\partial_t u^{(k)} - \tfrac{\tau}{\theta}\partial_x^2 u^{(k)} + cu^{(k)} \;=\; \underbrace{\omega\Big\langle \sum_{j=0}^{k-1} P_{k+2-j}u^{(j)}\mathsf{w} - \tau c\sum_{j=0}^{k-1} P_{k-j}u^{(j)}\mathsf{w} + P_k g\mathsf{w},\; 1\Big\rangle}_{=:\mathfrak{R}_k}.$$

Let k be even. Due to equation (3.24) only those terms survive the scalar product, where $k+2-j$ is even too. Thus we obtain for $k=2n$ at the right hand side an expression containing only $u^{(0)}, u^{(2)}, ..., u^{(2n-2)}$ and g as well as certain (mostly mixed) partial derivatives of these quantities. For odd k the situation is completely analogous except that g does not occur. This results in a *decoupling* of even and odd orders, that emerges as the responsible feature for the second order convergence mentioned in section 3. Setting either $k = 2l$ or $k = 2l + 1$ we obtain the following evolution equations from the above one:

$$\partial_t u^{(2l)} - \tfrac{\tau}{\theta}\partial_x^2 u^{(2l)} + cu^{(2l)} = \mathfrak{R}_{2n}\big(u^{(0)}, u^{(2)}, ..., u^{(2l-2)}, g\big), \tag{3.26}$$

$$\partial_t u^{(2l+1)} - \tfrac{\tau}{\theta}\partial_x^2 u^{(2l+1)} + cu^{(2l+1)} = \mathfrak{R}_{2l+1}\big(u^{(1)}, u^{(3)}, ..., u^{(2l-1)}\big). \tag{3.27}$$

In the case of the two lowest orders we get in particular:

$$\partial_t u^{(0)} - \tfrac{\tau}{\theta}\partial_x^2 u^{(0)} + cu^{(0)} \;=\; g \tag{3.28}$$

$$\partial_t u^{(1)} - \tfrac{\tau}{\theta}\partial_x^2 u^{(1)} + cu^{(1)} \;=\; 0 \tag{3.29}$$

These evolution equations respresent the desired conditions, we have to impose on the mass moments, in order to make the ansatz (3.12) “as consistent as possible”. Together with appropriate initial conditions, this hierarchy of diffusion-reaction equations determines the mass moments $u^{(0)}, ..., u^{(n-2)}$ uniquely. In contrast, no constraints arise from the solvability condition for the mass moments $u^{(n-1)}, u^{(n)}$. Besides sufficient regularity requirements, they can be freely chosen. That is why we are allowed to set them to zero.

Remark 3.3. (Vanishing of odd orders) The homogeneity of equation (3.29) has an important consequence, as $u^{(1)}$ remains zero once and for all, if it is initialized so. By equation (3.27) higher odd orders depend entirely on lower ones. Therefore we conclude, that all subsequent odd orders vanish too, provided they are initialized by zero. This fact emanates from the circumstance that the inhomogeneity $g\mathsf{w}$ of the lattice-Boltzmann equation (3.2) appears only in the zeroth order as $\mathrm{O}(1)$-term. Of course, things would change, if some $\mathrm{O}(\epsilon)$-term were admitted as additional source.

Remark 3.4. (Construction of the comparison function) How can we construct a truncated regular expansion, that reaches on the one hand a prescribed order of consistency but that is on the other hand as simple as possible? The previous considerations give the following suggestion, that will serve as guideline for the next subsection. In order to distinguish easily between even and odd order functions we consider $\mathsf{f}^{[2n+3]}$ (instead of $\mathsf{f}^{[n]}$) from now on.

a) Find sufficiently smooth functions $u^{(0)}, u^{(2)}, ..., u^{(2n)}$ satisfying (3.28) and (3.27) respectively.

b) Set $u^{(1)}, u^{(3)}, ..., u^{(2n+1)}, u^{(2n+3)}$ and $u^{(2n+2)}$ equal to 0 and compute by equation (3.25) the $2m + 4$ asymptotic orders $\mathsf{f}^{(0)}, \mathsf{f}^{(1)}, ..., \mathsf{f}^{(2n+3)}$.

c) Set $\hat{\mathsf{f}}^{[2n+3]} := \sum_{k=0}^{2n+3} \epsilon^k \mathsf{f}^{(k)}$, which is then consistent of order $2n + 2$.

d) As $u^{(2n+2)}, u^{(2n+3)} = 0$ the mass moments of $\mathsf{f}^{(2n+2)}, \mathsf{f}^{(2n+3)}$ vanish. According to equation (3.14) the residual is therefore simplified to

$$\tilde{\mathsf{r}}_\epsilon^{[2n+3]} = \epsilon^{2n+2}\Big\{\partial_t \mathsf{f}^{(2n+2)} + S\partial_x \mathsf{f}^{(2n+3)} + \epsilon\partial_t \mathsf{f}^{(2n+3)}\Big\}\,. \tag{3.30}$$

The proposed construction of the comparison function has a nice impact on the internal structure of the asymptotic orders. It ensures that even asymptotic orders are exclusively composed of terms containing an even power of s (i.e. $1 \equiv \mathsf{s}^0, \mathsf{s}^2$). In contrast, the addends pertaining to odd asymptotic order functions are always proportional to $\mathsf{s} \equiv \mathsf{s}^1$, which is the only odd power. We prove this property by exploiting the orthogonality of $\mathsf{s} \in \mathcal{F}$ with respect to the subspace span $\{1, \mathsf{s}^2\} \subset \mathcal{F}$.

Lemma 3.4. *Suppose that the asymptotic orders are constructed as described above, such that the mass moments of the odd orders vanish all. Then the following relations hold true for $j \in \{0, 1, ..., 2n+3\}$.*

$$\langle \mathsf{f}^{(2j)}, \mathsf{s}\rangle = 0, \qquad \langle \mathsf{f}^{(2j+1)}, \mathsf{s}^2\rangle = 0, \qquad \langle \mathsf{f}^{(2j+1)}, \mathsf{s}^2 - \tfrac{1}{\theta}\rangle = 0 \tag{3.31}$$

Proof: The proof is done by induction.
Start: Using $u^{(1)} = 0$ we get from (3.19) and (3.20)

$$\mathsf{f}^{(0)} = u^{(0)}\mathsf{w}\,, \qquad \mathsf{f}^{(1)} = D_x \mathsf{f}^{(0)} = -\tau\partial_x u^{(0)}\mathsf{sw}\,.$$

Recalling the orthogonality of $\mathsf{w}, \mathsf{s} \in \mathcal{F}$, (3.31) becomes directly clear for $j = 0$.
Step: Let us now suppose, that the relations are true for the asymptotic orders $\mathsf{f}^{(2j)}$ and $\mathsf{f}^{(2j+1)}$ for some $j \in \mathbb{N}_0$. Applying the recursive representation (3.21) for $k = 2j$ we find

$$\begin{aligned}\langle \mathsf{f}^{(2j+2)}, \mathsf{s}\rangle &= u^{(2j+2)}\langle \mathsf{w}, \mathsf{s}\rangle + \langle D_t \mathsf{f}^{(2j)}, \mathsf{s}\rangle + \langle D_x \mathsf{f}^{(2j+1)}, \mathsf{s}\rangle - \tau c\langle \mathsf{f}^{(2j)}, 1\rangle\,\langle \mathsf{w}, \mathsf{s}\rangle \\ &= 0 \;-\; \tau\partial_t\langle \mathsf{f}^{(2j)}, \mathsf{s}\rangle \;-\; \tau\partial_x\langle \mathsf{s}\mathsf{f}^{(2j+1)}, \mathsf{s}\rangle \;-\; 0 \;=\; 0\end{aligned}$$

Using $\langle \mathsf{s}\mathsf{f}^{(2j+1)}, \mathsf{s}\rangle = \langle \mathsf{f}^{(2j+1)}, \mathsf{s}^2\rangle$ we see, that both scalar products vanish due to the induction hypothesis. Analogously we consider equation (3.21) for $k = 2j+1$ using the assumption of the lemma ($u^{(2j+3)} = 0$) and the induction hypothesis $\langle \mathsf{f}^{(2j+1)}, 1\rangle = u^{(2j+1)} = 0$.

$$\begin{aligned}\langle \mathsf{f}^{(2j+3)}, \mathsf{s}^2\rangle &= \langle D_t \mathsf{f}^{(2j+1)}, \mathsf{s}^2\rangle \;+\; \langle D_x \mathsf{f}^{(2j+2)}, \mathsf{s}^2\rangle \\ &= -\tau\partial_t\langle \mathsf{f}^{(2j+1)}, \mathsf{s}^2\rangle \;-\; \tau\partial_x\langle \mathsf{s}\mathsf{f}^{(2j+2)}, \mathsf{s}^2\rangle \;=\; 0\end{aligned}$$

The first scalar product vanishes by the induction hypothesis. Shifting s from the left to the right hand side of the second scalar product and using $\mathsf{s}^3 = \mathsf{s}$, we obtain $\langle \mathsf{f}^{(2j+2)}, \mathsf{s}\rangle$. But this vanishes by what we have just proved above.
The third asserted relation is an immediate consequence of the second one and the assumption that the odd mass moments are zero.

$$\langle \mathsf{f}^{(2j+1)}, \mathsf{s}^2 - \tfrac{1}{\theta}\rangle = \langle \mathsf{f}^{(2j+1)}, \mathsf{s}^2\rangle - \tfrac{1}{\theta}\langle \mathsf{f}^{(2j+1)}, 1\rangle = 0 - 0 = 0 \qquad \blacksquare$$

These relations are exploited in the next section, when we show that the moments with respect to s and $\mathsf{s}^2 - \frac{1}{\theta}$ exhibit also a privileged convergence behavior besides the mass moment.

3.3 Smoothness conditions and convergence

After these rather formal computations, we shall formulate results about *asymptotic similarity* and *convergence* concerning the population function f_ϵ and the comparison function $\hat{\mathsf{f}}_\epsilon^{[2n+3]}$. Above all, this should include the answer to the following questions:

> Under what regularity conditions can equations (3.26) and (3.27) be solved successively up to a certain order? What regularity is necessary for the mass moments in order to bestow a classical meaning upon the recursive formula (3.21) for constructing the asymptotic order functions and thus $\hat{\mathsf{f}}_\epsilon^{[2n+3]}$?

The issue of smoothness is of utmost importance, since lower order mass moments occur as source terms in the evolution equations of higher order ones. So the regularity of every order function influences the regularity of the following ones and decides thus whether the expansion ansatz (3.12) can be extended by further orders.

Due to these reasons, it would be imperative to study regularity properties concerning the IVPs on an interval with periodic boundary conditions for the diffusion-reaction equation and the lattice-Boltzmann equation. Since our main focus is on lattice-Boltzmann specific arguments, we choose a pragmatic way and postulate the required smoothness statements in the form of two assumptions.

Assumption 3.1. For given $k \in \mathbb{N}_0$ let $v_0 \in \mathcal{C}^{k+2}_{\text{per}}(\mathcal{X})$ and $g \in \mathscr{C}^{k}_{\text{per}}(\mathcal{X}_T)$. Furthermore we assume the constants $\nu > 0$ and $c \geq 0$. Then the initial value problem for the diffusion-reaction equation

$$v(0,\cdot) = v_0, \qquad \partial_t v - \nu\partial_x^2 v + cv = g \quad \text{with } \nu = \tfrac{\tau}{\theta} \tag{3.32}$$

has a unique solution in $\mathscr{C}^{k+2}_{\text{per}}(\mathcal{X}_T)$.

Assumption 3.2. For given $k \in \mathbb{N}_0$ let $\mathsf{f}_{\text{init}} \in \mathcal{C}^{k+1}_{\text{per}}(\mathcal{X};\mathcal{F})$ and $g \in \mathcal{C}^{k}_{\text{per}}(\mathcal{X}_T;\mathcal{F})$. Then the initial value problem for the lattice-Boltzmann equation

$$\mathsf{f}_\epsilon(0,\cdot) = \mathsf{f}_\epsilon^{\text{init}}, \qquad \partial_t \mathsf{f}_\epsilon + \epsilon^{-1} S \partial_x \mathsf{f}_\epsilon \;=\; \epsilon^{-2} J_\epsilon \mathsf{f}_\epsilon + g\mathsf{w} \tag{3.33}$$

has a unique solution in $\mathcal{C}^{k+1}_{\text{per}}(\mathcal{X}_T;\mathcal{F})$.

Lemma 3.5. *For given $n, \mu \in \mathbb{N}_0$ let $v_0 \in \mathcal{C}^{2n+2+\mu}_{\text{per}}(\mathcal{X})$ and $g \in \mathscr{C}^{2n+\mu}_{\text{per}}(\mathcal{X}_T)$. The subsequent initial value problems*

$$\begin{aligned}
&u^{(k)}(0,\cdot) = v_0 \delta_{k0} \\
&\partial_t u^{(2k)} - \tfrac{\tau}{\theta}\partial_x^2 u^{(2k)} + c u^{(2k)} = \mathfrak{R}_{2k}\big(u^{(0)}, u^{(2)}, ..., u^{(2k-2)}, g\big)
\end{aligned}$$

can be solved successively for $0 \leq k \leq n$ and we have $u^{(2k)} \in \mathscr{C}^{2n-2k+2+\mu}_{\text{per}}(\mathcal{X}_T)$.

Proof: Starting with $k = 0$ the initial value problems are solved one after the other. Due to the required smoothness of v_0 and g, assumption 3.1 guarantees the existence of a solution $u^{(0)} \in \mathscr{C}^{2n+2+\mu}_{\text{per}}(\mathcal{X}_T)$.
It remains to check, whether the source term $\mathfrak{R}_{2k}$ on the right hand side is smooth enough, such that $u^{(2k)}$ possesses the claimed regularity. This is done by induction.
Start: The case $k = 0$ was just settled.
Hypothesis: The assertion of the lemma is supposed to be correct for $j \in \{0, ..., k-1\}$, i.e. $u^{(2j)} \in \mathscr{C}^{2n-2k+2+\mu}(\mathcal{X}_T)$.

Step: In order to find $u^{(2k)} \in \mathscr{C}^{2n-2k+2+2m}$, we have to verify the premises of assumption 3.1. So we need $\mathfrak{R}_{2k} \in \mathscr{C}^{2n-2k+2m}$. From

$$\mathfrak{R}_{2k} = \omega\Big\langle \sum_{j=0}^{k-1} P_{2k+2-2j} u^{(2j)} \mathsf{w} - \tau c \sum_{j=0}^{k-1} P_{2k-2j} u^{(2j)} \mathsf{w} + P_{2k} g \mathsf{w},\ 1\Big\rangle$$

the regularity of the individual addends can be directly read off if we use the induction hypothesis and take remark 3.2 into account.

$$\begin{array}{rcll} P_{2k+2-2j}u^{(2j)}\mathsf{w} & \in & \mathscr{C}_{\text{per}}^{(2n-2j+2+2m)-(2k+2-2j)}(\mathcal{X}_T;\mathcal{F}) & \subseteq \mathscr{C}_{\text{per}}^{2n-2k+2m}(\mathcal{X}_T;\mathcal{F}) \\ P_{2k-2j}u^{(2j)}\mathsf{w} & \in & \mathscr{C}_{\text{per}}^{(2n-2j+2+2m)-(2k-2j)}(\mathcal{X}_T;\mathcal{F}) & \subseteq \mathscr{C}_{\text{per}}^{2n-2k+2m}(\mathcal{X}_T;\mathcal{F}) \\ P_{2k}g\mathsf{w} & \in & \mathscr{C}_{\text{per}}^{(2n+2m)-2k}(\mathcal{X}_T;\mathcal{F}) & \subseteq \mathscr{C}_{\text{per}}^{2n-2k+2m}(\mathcal{X}_T;\mathcal{F}) \end{array}$$

Thus, $\mathfrak{R}_{2k}$ has the required smoothness and assumption 3.1 yields the assertion $u^{(2k)} \in \mathscr{C}_{\text{per}}^{2n-2k+2+\mu}(\mathcal{X}_T)$. ∎

Let us now consider the case, where $\mu = 1 + 2m$ with some $m \in \mathbb{N}$. So the given data is $v_0 \in \mathcal{C}_{\text{per}}^{2n+3+2m}(\mathcal{X})$ and $g \in \mathscr{C}_{\text{per}}^{2n+1+2m}(\mathcal{X}_T)$. The preceding lemma provides $u^{(0)}, u^{(2)}, ..., u^{(2n)}$, such that we can proceed as described in remark 3.4. Introducing additionally $u^{(2n+2)} \equiv 0$, this yields the $2n + 3$ asymptotic orders

$$\mathsf{f}^{(2l)} = \sum_{k=0}^{l} P_{2l-2k} u^{(2k)} - \tau c \sum_{k=0}^{l-2} P_{2l-2-2k} u^{(2k)} + \tau P_{2l-2} g \mathsf{w} \tag{3.34}$$

$$\mathsf{f}^{(2l+1)} = \sum_{k=0}^{l} P_{2l+1-2k} u^{(2k)} - \tau c \sum_{k=0}^{l-2} P_{2l-1-2k} u^{(2k)} + \tau P_{2l-1} g \mathsf{w} \tag{3.35}$$

for $0 \leq l \leq n + 1$.

Lemma 3.6. $\mathsf{f}^{(2l)}, \mathsf{f}^{(2l+1)}$ *are well defined functions in* $\mathcal{C}_{\text{per}}^{m}(\mathcal{X}_T, \mathcal{F})$ *for the specified range of* l*.*

Proof: As we have

$$g \in \mathscr{C}_{\text{per}}^{2n+1+2m}(\mathcal{X}_T) \quad \text{and} \quad u^{(2k)} \in \mathscr{C}_{\text{per}}^{2n-2k+3+2m}(\mathcal{X}_T) \quad \text{by lemma 3.5,}$$

we verify the assertion by counting the degree of regularity which each addend in (3.34) and (3.35) has. For instance:

$$\begin{array}{rcll} P_{2l+1-2k}u^{(2k)}\mathsf{w} & \in & \mathscr{C}_{\text{per}}^{(2n-2k+3+2m)-(2l+1-2k)}(\mathcal{X}_T;\mathcal{F}) & \subset \mathscr{C}_{\text{per}}^{2n-2l+2+2m}(\mathcal{X}_T;\mathcal{F}) \\ P_{2l-1}g\mathsf{w} & \in & \mathscr{C}_{\text{per}}^{(2n+1+2m)-(2l-1)}(\mathcal{X}_T;\mathcal{F}) & \subset \mathscr{C}_{\text{per}}^{2n-2l+2+2m}(\mathcal{X}_T;\mathcal{F}) \end{array}$$

Due to remark 3.23, the application of $P_{2l+1-2k}$ or P_{2l-1} reduces the degree of smoothness by $2l+1-2k$ or $2l-1$ respectively. Since l is at most $n+1$, all addends are at least in $\mathscr{C}_{\text{per}}^{2m}(\mathcal{X}_T;\mathcal{F}) \subset \mathcal{C}_{\text{per}}^{m}(\mathcal{X}_T;\mathcal{F})$. ∎

Theorem 3.2. *(Main result, asymptotic similarity) Let* $v_0 \in \mathcal{C}_{\text{per}}^{2n+3+2m}(\mathcal{X})$ *and* $g \in \mathcal{C}_{\text{per}}^{2n+1+2m}(\mathcal{X}_T)$ *and suppose that assumption 3.1 holds true such that* $\hat{\mathsf{f}}_\epsilon^{[2n+3]} \in$

$\mathcal{C}^m_{\text{per}}(\mathcal{X}_T)$ can be constructed as described above. Let f_ϵ be the solution of the lattice-Boltzmann equation 3.33 initialized by $\hat{\mathsf{f}}^{[2n+3]}_\epsilon$, i.e.

$$\mathsf{f}_\epsilon(0,\cdot) = \hat{\mathsf{f}}^{[2n+3]}_\epsilon(0,\cdot) \qquad \textit{with} \quad \hat{\mathsf{f}}^{[2n+3]}_\epsilon := \sum_{j=0}^{2n+3} \epsilon^j \mathsf{f}^{(j)}.$$

If assumption 3.2 holds true, such that $\mathsf{f}_\epsilon \in \mathcal{C}^m_{\text{per}}(\mathcal{X}_T)$, then there exists a constant $K > 0$ independent of ϵ such that

$$\left\|\mathsf{f}_\epsilon(t,\cdot) - \hat{\mathsf{f}}^{[2n+3]}_\epsilon(t,\cdot)\right\|_{\mathcal{C}\left(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X};\mathcal{F})\right)} \;<\; K\epsilon^{2n+2}\,.$$

Proof: The claimed estimate comes out as an almost automatic consequence: The comparison function has been defined in such a way, that the residual is given by equation (3.30). Since $\mathsf{f}^{(2n+2)}, \mathsf{f}^{(2n+3)} \in \mathcal{C}^m(\mathcal{X}_T,\mathcal{F})$ according to lemma 3.6, we get $\mathsf{r}^{[2n+3]}_\epsilon \in \mathcal{C}^{m-1}(\mathcal{X}_T,\mathcal{F})$. Hence there is a constant dominating the residue function $\mathsf{r}^{[2n+3]}_\epsilon$ and its derivatives (up to order $m-1$) on the whole compact space-time domain $\mathcal{X}_T$ for arbitrary $\epsilon \in (0,1]$. *A fortiori*, condition (3.7) is satisfied by $\mathsf{r}^{[2n+3]}_\epsilon$. Furthermore f_ϵ and $\hat{\mathsf{f}}^{[2n+3]}_\epsilon$ agree for $t = 0$. So the hypothesis of theorem 3.1 is granted for $\alpha = 2n+2$ and we can allege corollary 3.1 to verify the assertion. ■

Let us recall, that for any $\mathsf{x} \in \mathcal{F}$ the quantity $\langle \mathsf{f}_\epsilon, \mathsf{x}\rangle$ is called the *moment* associated with x. We have already seen in the last subsection that the mass moment u_ϵ, which is the moment with respect to $\mathsf{1} \in \mathcal{F}$, plays a major role in the analysis of the lattice-Boltzmann equation. Beside u_ϵ, two other moments turn out to be relevant:

$$u_\epsilon := \langle \mathsf{f}_\epsilon, \mathsf{1}\rangle, \qquad \phi_\epsilon := \langle \mathsf{f}_\epsilon, \mathsf{s}\rangle, \qquad \psi_\epsilon := \langle \mathsf{f}_\epsilon, \mathsf{s}^2 - \tfrac{1}{\theta}\rangle\,. \tag{3.36}$$

ϕ_ϵ and ψ_ϵ are referred to as the *first* and *second* moment[12] respectively. Like the comparison function $\hat{\mathsf{f}}^{[2n+3]}_\epsilon$ itself, all its moments are polynomials in ϵ. In particular, $\phi^{(k)}$ and $\psi^{(k)}$ figure as coefficient functions of the first and second moment respectively, being thus the first and second moment of $\mathsf{f}^{(k)}$. However, unlike $\hat{\mathsf{f}}^{[2n+3]}_\epsilon$, $\hat{\phi}^{[2n+3]}_\epsilon$ and $\hat{\psi}^{[2n+3]}_\epsilon$ expose a feature that $\hat{u}^{[2n+3]}_\epsilon$ has by construction: only every second power of ϵ has a non-zero coefficient:

$$\begin{array}{lclcl}
\hat{u}^{[2n+3]}_\epsilon & := & \langle \hat{\mathsf{f}}^{[2n+3]}_\epsilon, \mathsf{1}\rangle & = & u^{(0)} + \epsilon^2 u^{(2)} + \ldots + \epsilon^{2n} u^{(2n)},\\
\hat{\phi}^{[2n+3]}_\epsilon & := & \langle \hat{\mathsf{f}}^{[2n+3]}_\epsilon, \mathsf{s}\rangle & = & \epsilon\phi^{(1)} + \epsilon^3\phi^{(3)} + \ldots + \epsilon^{2n+3}\phi^{(2n+3)},\\
\hat{\psi}^{[2n+3]}_\epsilon & := & \langle \hat{\mathsf{f}}^{[2n+3]}_\epsilon, \mathsf{s}^2 - \frac{1}{\theta}\rangle & = & \epsilon^2\psi^{(2)} + \epsilon^4\psi^{(4)} + \ldots + \epsilon^{2n+2}\psi^{(2n+2)}.
\end{array}$$

This property is easily understood with the help of lemma 3.4. There we have proved, that certain moments of the asymptotic order functions vanish if it is constructed as described in remark 3.4. The expansion of $\hat{\psi}_\epsilon$ starts with a second order term, since $\mathsf{s}^2 - \frac{1}{\theta}$ is orthogonal to w and hence to $\mathsf{f}^{(0)} = u^{(0)}\mathsf{w}$.

[12] The linear independence of the three moment generating polynomials in $\mathcal{F}$ should be noticed. Shortly we will see, the first moment is related to the (diffusive) flux $-\nu\partial_x v$ wherefore it is also referred to as the *flux moment*. The second moment contains the the second derivative of v which has no intuitive physical meaning.

Corollary 3.3. *(Asymptotic similarity of the moments) Under the conditions of the previous theorem, the moments of the comparison function* $\hat{\mathsf{f}}_\epsilon^{[2n+3]}$ *and of the population function* f_ϵ *are asymptotically similar with the same order as* f_ϵ *and* $\hat{\mathsf{f}}_\epsilon^{[2n+3]}$*. The following estimates hold true where* $K > 0$ *is the constant from the theorem:*

$$\begin{aligned} \|u_\epsilon \ - \ \hat{u}_\epsilon^{[2n+3]}\|_{\mathcal{C}\left(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X})\right)} \ &\leq \ \|1\|_{\mathcal{F}} \ K\,\epsilon^{2n+2}, \\ \|\phi_\epsilon \ - \ \hat{\phi}_\epsilon^{[2n+3]}\|_{\mathcal{C}\left(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X})\right)} \ &\leq \ \|\mathsf{s}\|_{\mathcal{F}} \ K\,\epsilon^{2n+2}, \\ \|\psi_\epsilon \ - \ \hat{\psi}_\epsilon^{[2n+3]}\|_{\mathcal{C}\left(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X})\right)} \ &\leq \ \|\mathsf{s}^2 - \tfrac{1}{\theta}\|_{\mathcal{F}} \ K\,\epsilon^{2n+2}. \end{aligned} \tag{3.37}$$

Proof: It is sufficient to consider the case of the mass moment, since the other two cases can be treated analogically. Recalling the definition of the norm in $\mathcal{C}\big(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X})\big)$ and using the Cauchy-Schwarz inequality from the second to the third line yields the assertion:

$$\begin{aligned} \big\|u_\epsilon - \hat{u}_\epsilon^{[2n+3]}\big\|_{\mathcal{C}\left(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X})\right)} &= \sup_{t\in\mathcal{T}} \left[\sum_{i=0}^{m-1} \int_{\mathcal{X}} \big|\partial_x^i u_\epsilon(t,x) - \partial_x^i \hat{u}_\epsilon^{[2n+3]}(t,x)\big|^2 \mathrm{d}x\right]^{\frac{1}{2}} \\ &= \sup_{t\in\mathcal{T}} \left[\sum_{i=0}^{m-1} \int_{\mathcal{X}} \big|\langle\partial_x^i \mathsf{f}_\epsilon(t,x) - \partial_x^i \hat{\mathsf{f}}_\epsilon^{[2n+3]}(t,x), 1\rangle\big|^2 \mathrm{d}x\right]^{\frac{1}{2}} \\ &\leq \sup_{t\in\mathcal{T}} \left[\sum_{i=0}^{m-1} \int_{\mathcal{X}} \big\|\partial_x^i \mathsf{f}_\epsilon(t,x) - \partial_x^i \hat{\mathsf{f}}_\epsilon^{[2n+3]}(t,x)\big\|_{\mathcal{F}}^2 \ \|1\|_{\mathcal{F}}^2 \ \mathrm{d}x\right]^{\frac{1}{2}} \\ &\leq \|1\|_{\mathcal{F}} \ \big\|\mathsf{f}_\epsilon - \hat{\mathsf{f}}_\epsilon^{[2n+3]}\big\|_{\mathcal{C}\left(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X};\mathcal{F})\right)} \ \leq \ \|1\|_{\mathcal{F}} \ K\epsilon^{2n+2}. \end{aligned}$$

■

The construction of $\hat{\mathsf{f}}_\epsilon^{[3]}$ needs only the solution v of (3.32) that is identified with $u^{(0)}$. The asymptotic order functions of $\hat{\mathsf{f}}_\epsilon^{[3]}$ become explicitly

$$\left.\begin{aligned} \mathsf{f}^{(0)}(t,x) \ &= \ v(t,x)\mathsf{w} \\ \mathsf{f}^{(1)}(t,x) \ &= \ -\tau\partial_x v(t,x)\mathsf{sw} \\ \mathsf{f}^{(2)}(t,x) \ &= \ \tau^2\partial_x^2 v(t,x)\Big(\mathsf{s}^2\mathsf{w} - \tfrac{1}{\theta}\mathsf{w}\Big) \\ \mathsf{f}^{(3)}(t,x) \ &= \ -\tau^3\partial_x^3 v(t,x)\Big(\mathsf{s}^3\mathsf{w} - \tfrac{1}{\theta}\mathsf{sw}\Big) + \tau^2\partial_t\partial_x v(t,x)\mathsf{sw} \end{aligned}\right\} \tag{3.38}$$

with the residue function

$$\begin{aligned} \mathsf{r}_\epsilon^{[3]} \ = \ &\tau^2\partial_t\partial_x^2 v \ \left(\mathsf{s}^2\mathsf{w} - \tfrac{1}{\theta}\mathsf{w}\right) \ - \ \tau^3\partial_x^4 v \ \left(\mathsf{s}^4\mathsf{w} - \tfrac{1}{\theta}\mathsf{s}^2\right) \\ &- \epsilon\left[\tau^3\partial_t\partial_x^3 v \ \left(\mathsf{s}^3\mathsf{w} - \tfrac{1}{\theta}\mathsf{sw}\right) \ - \ \tau^2\partial_t^2\partial_x v \ \mathsf{sw}\right]. \end{aligned}$$

Using the orthogonality of $\mathsf{w}, \mathsf{s}, \mathsf{s}^2 - \frac{1}{\theta} \in \mathcal{F}$ we obtain the following moments of $\hat{\mathsf{f}}_\epsilon^{[3]}$:

$$\langle\hat{\mathsf{f}}_\epsilon^{[3]}, 1\rangle = v, \qquad \langle\hat{\mathsf{f}}_\epsilon^{[3]}, \mathsf{s}\rangle = -\epsilon\tfrac{\tau}{\theta}\partial_x v, \qquad \langle\hat{\mathsf{f}}_\epsilon^{[3]}, \mathsf{s}^2 - \tfrac{1}{\theta}\rangle = -\epsilon^2\tfrac{\tau^2}{2\theta}\partial_x^2 v.$$

Apart from higher order corrections these formulas hold also true with repsect to other comparison functions than $\hat{\mathsf{f}}_\epsilon^{[3]}$ taking more asymptotic order functions into account. This results into the final corollary.

Corollary 3.4. *(Convergence of Moments) Under the conditions of the preceding theorem the mass moment* $u_\epsilon = \langle \mathsf{f}_\epsilon, 1\rangle$ *converges quadratically in* ϵ *to the solution* v *of* (3.32), *while the first and second moment* $\phi_\epsilon, \psi_\epsilon$ *vanish in the limit. However, the the limit of the* rescaled *moments* $\epsilon^{-1}\phi_\epsilon$ *and* $\epsilon^{-2}\psi_\epsilon$ *is related to the first and second derivative of* v*:*

$$\begin{aligned} u_\epsilon &= v + \mathrm{O}(\epsilon^2)\,, \\ \epsilon^{-1}\phi_\epsilon &= -\tfrac{\tau}{\theta}\partial_x v + \mathrm{O}(\epsilon^2)\,, \\ \epsilon^{-2}\psi_\epsilon &= -\tfrac{\tau^2}{2\theta^2}\partial_x^2 v + \mathrm{O}(\epsilon^2)\,. \end{aligned}$$

All relations are understood with respect to the $\mathcal{C}\big(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X})\big)$*-norm.*

Proof: The statement is a straightforward consequence of the previous corollary and the triangle inequality. By adding the same term on both sides of inequality (3.37), we get

$$\|u_\epsilon - \sum_{j=0}^{n} u^{(2j)}\epsilon^{2j}\| + \|\sum_{j=1}^{n} u^{(2j)}\epsilon^{2j}\| \leq \|1\|_{\mathcal{F}} K\epsilon^{2n+2} + \|\sum_{j=1}^{n} u^{(2j)}\epsilon^{2j}\|,$$

where the norms refer to $\mathcal{C}\big(\mathcal{T};\mathcal{H}^{m-1}(\mathcal{X})\big)$. Since arbitrary reals a, b satisfy $|a-b| \leq |a| + |b|$, the left hand side is shrunk to $\|u_\epsilon - u^{(0)}\|$. The right hand side contains only terms being at least of order ϵ^2. As the sum extends only over finitely many terms, which are continuous due to the regularity assumptions and thence bounded on the compact time-space domain $\mathcal{X}_T$, the right hand side can be estimated for $\epsilon \in (0,1]$ by a constant times ϵ^2. So it is an $\mathrm{O}(\epsilon^2)$-term.
Repeating the same arguments for the first and second moment, leads to:

$$\begin{aligned} \|\phi_\epsilon - \epsilon\,\phi^{(1)}\| &\leq \|\mathsf{s}\|_{\mathcal{F}}\,K\epsilon^{2n+2} + \mathrm{O}(\epsilon^3) \\ \|\psi_\epsilon - \epsilon^2\psi^{(2)}\| &\leq \|\mathsf{s}^2 - \tfrac{1}{\theta}\|_{\mathcal{F}}\,K\epsilon^{2n+2} + \mathrm{O}(\epsilon^4) \end{aligned}$$

Division by ϵ^1 and ϵ^2 respectively transforms the right hand side into an $\mathrm{O}(\epsilon^2)$-term if $n \geq 1$. ■

Observe that we need at least $\hat{\mathsf{f}}_\epsilon^{[5]}$ to conclude the assertion for ϕ_ϵ and ψ_ϵ.

3.4 Initial conditions and irregular expansions

As the regular expansion considered so far starts with a term of the form $u^{(0)}\mathsf{w}$, it cannot approximate solutions belonging to initial conditions whose leading order does not lie in the span of $\mathsf{w} \in \mathcal{F}$. This invites the question how arbitrary initial values evolve.

To answer this, we try to find other, non-regular expansions to construct approximate solutions of the lattice-Boltzmann equation[13]. Before we present the working expansion, we discuss an ansatz which turns out to fail.

[13]For simplicity the source term is set to 0.

We track the presumption that initial values being in span$\{\mathsf{s}, \mathsf{s}^2 - \frac{1}{\theta}\}$ are subject to a fast evolution in comparison with those in span$\{\mathsf{w}\}$. In this context *fast* means that they become marginally small after an ϵ-depending time that is short compared with 1 (e.g. $\epsilon, \epsilon^2, ...$).

This motivates the introduction of a scaled time variable t/ϵ which distinguishes the following ansatz

$$\mathsf{j}_\epsilon(t/\epsilon, x) = \mathsf{j}^{(0)}(t/\epsilon, x) + \epsilon \mathsf{j}^{(1)}(t/\epsilon, x) + ... \tag{3.39}$$

from the regular expansion. The role of j_ϵ is to approximate the population function f_ϵ just like the comparison function $\hat{\mathsf{f}}_\epsilon$. In order to ensure that (3.39) makes sense we require that the coefficient functions are bounded for $\epsilon \downarrow 0$. This entails in particular that $\epsilon^k \mathsf{j}^{(k)}(t/\epsilon, x) = \mathrm{O}(\epsilon^k)$. Plugging (3.39) into the lattice-Boltzmann equation, the standard comparison of coefficients[14] yields

$$\left.\begin{array}{lrcl} \epsilon^{-2}: & 0 & = & J_0\,\mathsf{j}^{(0)} \\ \epsilon^{-1}: & \partial_1 \mathsf{j}^{(0)} + S\partial_x \mathsf{j}^{(0)} & = & J_0\,\mathsf{j}^{(1)} \\ \epsilon^{k-1}: & \partial_1 \mathsf{j}^{(k)} + S\partial_x \mathsf{j}^{(k)} & = & J_0\,\mathsf{j}^{(k+1)} - c\langle \mathsf{j}^{(k-1)}, \mathsf{1}\rangle \mathsf{w} \end{array}\right\} \tag{3.40}$$

where ∂_1 denotes the derivative with respect to the first variable, i.e. t/ϵ (observe: $\partial_t = \epsilon^{-1}\partial_1$). The first equation pertaining to the order ϵ^{-2} agrees with the corresponding result in (3.13) for the regular expansion. So we obtain

$$\mathsf{j}^{(0)} \in \ker J_0 \quad \Rightarrow \quad \mathsf{j}^{(0)}(t/\epsilon, x) = w^{(0)}(t/\epsilon, x)\mathsf{w} \tag{3.41}$$

with some unknown function $w^{(0)}$. Already now this indicates clearly that (3.39) is not pertinent for our goal because it produces structurally the same term in zeroth order as the regular expansion. Nevertheless let us continue to check whether the ansatz works in principle.
Using (3.41) to compute the mass moment of the second equation in (3.40)

$$\begin{array}{rcl} \langle \partial_1 \mathsf{j}^{(0)}, \mathsf{1}\rangle \;+\; \langle S\partial_x \mathsf{j}^{(0)}, \mathsf{1}\rangle & = & \langle J_0\,\mathsf{j}^{(1)}, \mathsf{1}\rangle \\ \partial_1 w^{(0)} \underbrace{\langle \mathsf{w}, \mathsf{1}\rangle}_{=1} + \partial_x w^{(0)} \underbrace{\langle \mathsf{sw}, \mathsf{1}\rangle}_{=0} & = & \omega \underbrace{\big\langle \langle \mathsf{j}^{(1)}, \mathsf{1}\rangle \mathsf{w} - \mathsf{j}^{(1)}, \mathsf{1}\big\rangle}_{=0} \end{array}$$

we deduce $\partial_1 w^{(0)} = 0$, which shows that $w^{(0)}$ is constant in time. Hence $\mathsf{j}^{(1)}$ is solution of the equation

$$J_0 \mathsf{j}^{(1)} = \partial_x w^{(0)} \mathsf{sw}.$$

According to lemma 3.2 this equation is solvable since $\langle \mathsf{sw}, \mathsf{1}\rangle = 0$. Furthermore this yields

$$\mathsf{j}^{(1)} = w^{(1)}\mathsf{w} - \tau \partial_x w^{(0)} \mathsf{sw}. \tag{3.42}$$

Let us now turn to the zero order equation being the third one in (3.40) with $k = 1$:

$$\partial_1 w^{(1)}\mathsf{w} + \partial_x w^{(1)}\mathsf{sw} - \tau \partial_x w^{(0)} \mathsf{s}^2\mathsf{w} + c w^{(0)}\mathsf{w} = J_0 \mathsf{j}^{(2)} \tag{3.43}$$

[14] Strictly speaking, the comparison of coefficients applies for polynomials and power series if the coefficients are independent of the expansion variable. This situation is not given here. However the boundedness of the coefficients is enough to justify the procedure.

Again, this equation is only solvable if the mass moment of the left hand side vanishes which provides the equation to determine $w^{(1)}$ more precisely:

$$\partial_1 w^{(1)} - \tfrac{\tau}{\theta}\partial_x w^{(0)} + c w^{(0)} = 0.$$

As $w^{(0)}$ does not depend on time we conclude

$$w^{(1)}(t/\epsilon, x) = \frac{t}{\epsilon}\big(\tfrac{\tau}{\theta}\partial_x w^{(0)}(x) - c w^{(0)}(x)\big) + \text{const}.$$

However, this result violates the assumption that the order functions $\mathsf{j}^{(0)}, \mathsf{j}^{(1)}, ...$ remain bounded for fixed t and x as ϵ goes to 0. To overcome this inconsistency, we are obliged to equate the coefficient behind $\frac{t}{\epsilon}$ with 0, which implies $w^{(1)}$ to be constant in time as well. Thus we are led to a differential equation determining the spatial dependence of $w^{(0)}$:

$$c w^{(0)} = \tfrac{\tau}{\theta}\partial_x w^{(0)}.$$

For $c \neq 0$ the solution is given by an exponential function. This, however, cannot comply with the periodic boundary conditions. Therefore we must infer that $w^{(0)}$ vanishes if $c \neq 0$. In any case, $w^{(0)}$ must be constant in space and time. So (3.43) becomes

$$J_0 \mathsf{j}^{(2)} = \partial_x w^{(1)} \mathsf{sw}.$$

Since this representation is analogous to (3.42) and since the equation for $\mathsf{j}^{(3)}$ is structurally the same as (3.43) for $\mathsf{j}^{(2)}$, we see that we can repeat the arguments to infer that $w^{(1)}$ is constant too. This shows finally that only trivial or at most constant solutions of the lattice-Boltzmann equation can be expanded by (3.39). Due to this reason the ansatz must be rejected.

In order to obtain a more enlightening expansion we have to choose the time scaling t/ϵ^α such that the equation in the lowest order is not of the type $J_0 \mathsf{x} = 0$. It is quickly seen that this requirement is fulfilled for $\alpha = 2$. Therefore, let us consider the expansion

$$\mathsf{j}_\epsilon(t/\epsilon^2, x) = \mathsf{j}^{(0)}(t/\epsilon^2, x) + \epsilon \mathsf{j}^{(1)}(t/\epsilon^2, x) + \dots\,, \tag{3.44}$$

resulting in the subsequent equations for the order functions:

$$\left.\begin{array}{lrcl} \epsilon^{-2}: & \partial_1 \mathsf{j}^{(0)} & = & J_0\, \mathsf{j}^{(0)} \\ \epsilon^{-1}: & \partial_1 \mathsf{j}^{(1)} + S\partial_x \mathsf{j}^{(0)} & = & J_0\, \mathsf{j}^{(1)} \\ \epsilon^{k-2}: & \partial_1 \mathsf{j}^{(k)} + S\partial_x \mathsf{j}^{(k-1)} & = & J_0\, \mathsf{j}^{(k)} - c\langle \mathsf{j}^{(k-2)}, \mathsf{1}\rangle \mathsf{w} \end{array}\right\} \tag{3.45}$$

In contrast to the order functions of the regular expansion, we obtain for $\mathsf{j}^{(0)}, \mathsf{j}^{(1)}, ...$ *no PDEs but ODEs*, since the equations do not involve the spatial derivative of the highest occurring order function. So, the $\mathsf{j}^{(k)}$'s come out as solutions of ODEs with respect to the scaled time variable $r = t/\epsilon^2$ having the structure

$$\partial_1 \mathsf{j}^{(k)} - J_0 \mathsf{j}^{(k)} = -S\partial_x \mathsf{j}^{(k-1)} - c\langle \mathsf{j}^{(k-1)}, \mathsf{1}\rangle \mathsf{w}.$$

The right hand side is a true inhomogeneity as it depends only on $\mathsf{j}^{(k-1)}$ being determined by the previous equation. Thus the solution is formally given by means of the *variation of constants formula*:

$$\begin{aligned}\mathsf{j}^{(k)}(t/\epsilon^2, x) &= \mathrm{e}^{tJ_0/\epsilon^2}\mathsf{j}^{(k)}(0,x)\\ &\quad - \mathrm{e}^{tJ_0/\epsilon^2}\int_0^{t/\epsilon^2} \mathrm{e}^{-rJ_0}\big[S\partial_x \mathsf{j}^{(k-1)}(r,x) + c\langle \mathsf{j}^{(k-1)}(r,x), 1\rangle \mathsf{w}\big]\,\mathrm{d}r\end{aligned}$$

The leading order function $\mathsf{j}^{(0)}$ satisfies the homogeneous equation $\partial_1 \mathsf{j}^{(0)} = J_0 \mathsf{j}^{(0)}$.

Let us now consider each of the three basis vectors $\mathsf{w}, \mathsf{s}, \mathsf{s}^2 - \frac{1}{\theta} \in \mathcal{F}$ as initial value for $\mathsf{j}^{(0)}$. The following computation shows that these vectors constitute an eigenbasis of the collision operator J_0:

$$\begin{aligned} J_0\mathsf{w} &= \omega\Big[\langle \mathsf{w}, 1\rangle \mathsf{w} - \mathsf{w}\Big] &&= 0,\\ J_0\mathsf{s} &= \omega\Big[\langle \mathsf{s}, 1\rangle \mathsf{w} - \mathsf{s}\Big] &&= -\omega\mathsf{s},\\ J_0(\mathsf{s}^2 - \tfrac{1}{\theta}) &= \omega\Big[\langle \mathsf{s}^2 - \tfrac{1}{\theta}, 1\rangle \mathsf{w} - \mathsf{s}^2 - \tfrac{1}{\theta}\Big] &&= -\omega(\mathsf{s}^2 - \tfrac{1}{\theta}).\end{aligned}$$

Since it is assumed that ω is strictly positive, the eigenvalues of s and $\mathsf{s}^2 - \frac{1}{\theta}$ are negative which has a strong impact on their dynamics. So the solution associated with s and $\mathsf{s}^2 - \frac{1}{\theta}$ as initial values decreases exponentially[15] in the fast time t/ϵ^2. Using the variation of constants formula, it becomes clear that the exponential decay devolves upon the subsequent order functions $\mathsf{j}^{(1)}, \mathsf{j}^{(2)}, \ldots$ too. A fortiori, they remain bounded for fixed t, x as ϵ tends to 0, which justifies the ansatz (3.45).

As far as the first eigenvector w is concerned, the expansion displays a quite different behavior. Due to $J_0\mathsf{w} = 0$, the initial value $\mathsf{j}_\epsilon(0,x) = w^{(0)}(x)\mathsf{w}$ implies that $\mathsf{j}^{(0)}$ is constant in time, i.e. $\mathsf{j}^{(0)}(t/\epsilon^2, x) = w^{(0)}(x)\mathsf{w}$. For $\mathsf{j}^{(1)}$ we obtain then the initial value problem

$$\partial_1 \mathsf{j}^{(1)} + J_0 \mathsf{j}^{(1)} = \partial_x w^{(0)}(x)\mathsf{ws} \qquad \text{with} \quad \mathsf{j}^{(1)}(0,x) = 0.$$

Since ws is also an eigenvector of J_0 pertaining to the eigenvalue $-\omega$, this yields

$$\mathsf{j}^{(1)}(t/\epsilon^2, x) = \tau \mathrm{e}^{-\omega t/\epsilon^2}\partial_x w^{(0)}(x)\mathsf{ws} - \tau\partial_x w^{(0)}(x)\mathsf{ws}.$$

The exponentially decaying part is to adjust the initial condition while the second part (solution of the inhomogeneous equation) is constant in time and resembles the expression for $\mathsf{f}^{(1)}$ in equation (3.38). Without going into details of further calculations to compute higher order functions, we may note that the expansion appears also reasonable in the case of a w-proportional initial condition. In opposition to the other cases it contains constant terms. Once the decaying terms have vanished practically, a regular expansion is left[16] whose mass moment is given by the time

[15] For the initial value $\mathsf{j}^{(0)}(0,x) = w^{(0)}\mathsf{s}$ we obtain the solution $\mathsf{j}^{(0)}(t,x) = w^{(0)}(x)\mathrm{e}^{tJ_0/\epsilon^2}\mathsf{s} = w^{(0)}(x)\mathrm{e}^{-\omega t/\epsilon^2}\mathsf{s}$. If $\mathsf{s}^2 - \frac{1}{\theta}$ or sw is chosen instead of s, a completely analogous result is found.

[16] In the short time scale the "diffusive" decay of $w^{(0)}\mathsf{w}$ is not observable.

independent function $w^{(0)}$.

In summary, the following observations can be stated:

- The irregular expansion (3.44) contributes together with the regular expansion to a complete comprehension of the dynamics generated by the lattice-Boltzmann equation. Thanks to the linearity of the system it is easily possible to combine both expansions to understand the evolution associated to arbitrary initial values.

- Expansion (3.44) gives evidence for the existence of *initial layers*. This is a particular class of solutions decaying towards zero over a time interval of magnitude $O(\epsilon^2)$.

  Missing higher order terms in the regular expansion of the initial condition can be compensated by initial layers up to a maximal order depending on the regularity of the initial data.

- Finally, the parallel to section 2.3 should be noticed.

With regard to the first point, it remains to check whether any ansatz using another scaled time t/ϵ^α, $\alpha \neq 2$ can only approximate constant or trivial solutions of the lattice-Boltzmann equation (non-existence of further time scales).

3.5 A glimpse of boundary conditions

The conversion of macroscopic boundary conditions[17] into kinetic ones for the lattice-Boltzmann equation is a non-trivial task and still subject to current research [67, 6]. This difficulty is not of purely discrete nature faced in the context of lattice-Boltzmann schemes. In this section we want to substantiate that the problem is of intrinsic kinetic origin and may arise already at the level of kinetic differential equations. The reason for this lies in the fact that kinetic equations involve other quantities as primary variables than the macroscopic ones. This impedes a straight imposition of macroscopic boundary conditions since those quantities are not directly accessible which shall be kept on a certain value. Moreover, kinetic boundary conditions must mediate between those for the target equation and the requirements of the kinetic equation, that might be quite different from a mathematical point of view.

The force-driven Poiseuille flow through a canal provides a simple but illustrative example. Along the walls no-slip conditions apply while inflow and outflow are modeled by periodic boundary conditions. In section 1.2 it was pointed out, that the flow profile of such a quasi 1D flow is a solution of the diffusion equation.
Concretely, the Poiseuille flow is assumed to occur in y-direction. A symmetric interval $[-w, w]$, $w > 0$, on the x-axis may correspond to the cross-section of the canal. The stationary flow profile $[-w, w] \ni x \mapsto v(x)$ is then solution of a 1D Poisson

[17] Referring to the target equation.

equation (stationary diffusion equation), where homogeneous Dirichlet boundary conditions reflect the vanishing flow velocity at the walls.

$$\left.\begin{array}{r} -\nu v''(x) = g \\ v(-w) = 0 = v(w) \end{array}\right\} \quad \Rightarrow \quad v(x) = \frac{g}{2\nu}(w^2 - x^2). \tag{3.46}$$

On the right hand side, a constant g appears representing the strength of the force field which is directed parallel to the y-axis and induces the Poiseuille flow. The solution of (3.46) represents a parabola opened from below with roots at $-w$ and w. Its maximum is located in the center of the canal ($x = 0$) where the flow reaches its highest speed depending on the ratio of the force g and the viscosity ν.

Let us investigate how (3.46) is approximated in the framework of the kinetic approach using the D1P3 lattice-Boltzmann equation. Thanks to the stationarity of the target problem (3.46) we consider the time-independent lattice-Boltzmann equation[18]

$$\tfrac{1}{\epsilon} S \partial_x \mathsf{f} = \tfrac{1}{\epsilon^2} J_0 \mathsf{f} + g\mathsf{w}. \tag{3.47}$$

In order to avoid a reactive term appearing in (3.2), the reactivity c is set to 0, which results into the collision operator J_0. Equation (3.47) must be complemented by boundary conditions at w and $-w$.

It is well known that the bounce-back rule realizes a kinetic approximation of the macroscopic no-slip condition for the Stokes and Navier-Stokes equation particularly with regard to the D2P9 model. In section 1.2 it has been shown (see page 49) that the reduction from the D2P9 to the D1P3 model yields a bounce-back-like condition with sign-flipping. So we arrive at the following differential algebraic boundary value problem, which is the componentwise version of (3.47) multiplied by $\tau\epsilon^2$.

$$\left.\begin{array}{lcl lcl} \tau\epsilon\partial_x \mathsf{f}_1(x) &=& \mathsf{f}_1(x) - \frac{1}{2\theta}u(x) - \frac{1}{2\theta}\epsilon^2\tau g & \mathsf{f}_1(w) &=& -\mathsf{f}_2(w) \\ 0 &=& \mathsf{f}_0(x) - \frac{\theta-1}{\theta}u(x) - \frac{\theta-1}{\theta}\epsilon^2\tau g & & & \\ -\tau\epsilon\partial_x \mathsf{f}_2(x) &=& \mathsf{f}_2(x) - \frac{1}{2\theta}u(x) - \frac{1}{2\theta}\epsilon^2\tau g & \mathsf{f}_2(-w) &=& -\mathsf{f}_1(-w) \end{array}\right\} \tag{3.48}$$

Let us derive a solution of this boundary value problem: Motivated by the solution of the target problem, we assume that the populations and their mass moment u are described by quadratic polynomials. To reduce the number of unknown variables, we exploit the symmetry of the problem.

So we require that u and the rest population f_0 are invariant under reflection at the origin: $u(x) = u(-x)$ and $\mathsf{f}_0(x) = \mathsf{f}_0(-x)$. This means that the zeros of the associated quadratic polynomials are symmetrically situated with respect to 0. In particular linear terms are excluded. In opposition, left and right moving populations should obey the relation $\mathsf{f}_1(-x) = \mathsf{f}_2(x)$. All assumptions are taken into account by the following ansatz

$$\begin{array}{rcl cl} u(x) &=& a(n-x)(n+x) &=& -ax^2 + an^2 \\ \mathsf{f}_0(x) &=& b(p-x)(p+x) &=& -bx^2 + bp^2 \\ \mathsf{f}_1(x) &=& d(q-x)(r+x) &=& -dx^2 + d(q-r)x + dqr \\ \mathsf{f}_2(x) &=& d(r-x)(q+x) &=& -dx^2 + d(r-q)x + dqr \end{array}$$

[18]In this section we do not stress the ϵ-dependence by index notation. So we write $\mathsf{f}, \hat{\mathsf{f}}$ instead of $\mathsf{f}_\epsilon, \hat{\mathsf{f}}_\epsilon$.

where $a, b, d \geq 0$ and $n, p, q, r \geq 0$ are respectively the amplitudes and roots of the unknown polynomials. It turns out that these seven coefficients are uniquely determined by (3.48).

- First, the boundary conditions for f_1 and f_2 yield: $\quad rq = w^2$.
- Performing a comparison of coefficients in front of the x^2-term and the constant term occurring in the algebraic equation for $\mathsf{f}^{(0)}$ ends up with the relations:
$$b = \tfrac{\theta-1}{\theta} a \qquad \text{and} \qquad a(z^2 - p^2) = \epsilon^2 \tau g. \tag{3.49}$$
- Similarly, we extract from the equation for f_1
$$d = \tfrac{1}{2\theta a} \qquad r - q = 2\tau\epsilon \qquad a(p^2 - w^2) = 2a\tau^2\epsilon^2 - g\tau^2\epsilon^2 \tag{3.50}$$
- The relation $u = \mathsf{f}_1 + \mathsf{f}_0 + \mathsf{f}_2$ imposes the equation
$$(\theta - 1)n^2 + w^2 = \theta p^2. \tag{3.51}$$

Adding the right equations of (3.49) and (3.50), and solving for p gives $p = \sqrt{w^2 + 2\tau\epsilon^2}$.
Combining this with (3.51) produces $n = \sqrt{w^2 + 2\frac{\theta-1}{\theta}\tau^2\epsilon^2}$.
Utilizing $a = \frac{1}{2}\frac{\theta}{\tau} g$ leads to the final result for u:

$$\boxed{u(x) = \tfrac{g}{2\nu}\left(w^2 - x^2 + 2\tfrac{\theta-1}{\theta}\epsilon^2\tau^2\right) = v(x) + \tfrac{g}{\nu}\tfrac{\theta-1}{\theta}\tau^2\epsilon^2} \tag{3.52}$$

Comparing the mass moment u with the solution v of the target problem (3.46) confirms the statement of corollary 3.4 i.e.:

$$\langle \mathsf{f}, 1 \rangle = u \xrightarrow{\epsilon \downarrow 0} v \qquad \text{more precisely} \qquad v = u + \mathrm{O}(\epsilon^2).$$

Formula (3.52) corroborates that homogeneous Dirichlet boundary conditions are not exactly satisfied. This is symptomatic for other macroscopic boundary conditions, whose translation into kinetic boundary conditions is possibly less evident.

Returning to (3.52), the error term vanishes remarkably if θ attains its limit value 1. Hence, for the example of the Poiseuille flow, the solution of the target equation is produced exactly if the D1P2 model is used as a special case of the D1P3 model.

In general, kinetic boundary conditions are obtained as conditions on the the moments of the population function f.

$$\begin{array}{lllll} \text{bounce-back with} \quad \text{sign-flipping} & \mathsf{f}_1 = -\mathsf{f}_2 & \langle \mathsf{s}^2, \mathsf{f} \rangle = 0 & \rightarrow & \text{Dirichlet b.c.} \end{array} \tag{3.53}$$
$$\begin{array}{lllll} \text{bounce-back without sign-flipping} & \mathsf{f}_1 = \mathsf{f}_2 & \langle \mathsf{s}, \mathsf{f} \rangle = 0 & \rightarrow & \text{Neumann b.c.} \end{array} \tag{3.54}$$

In order to see what kind of macroscopic boundary condition ensues from (3.53), we insert the comparison function. More precisely the boundary conditions for the mass moments $u^{(0)}, u^{(2)}, \dots$ of the asymptotic order functions are chosen such that the comparison function $\hat{\mathsf{f}}^{[2n+3]}$ satisfies the boundary condition with a residual as small as possible. Explicitly let us consider $\hat{\mathsf{f}}^{[3]}$ which is constructed alone by the solution $v = u^{(0)}$ of the target equation (3.32). Using the expressions for $\mathsf{f}^{(0)}, \mathsf{f}^{(1)}, \mathsf{f}^{(2)}$ and $\mathsf{f}^{(3)}$ in (3.38) we obtain

$$\langle \hat{\mathsf{f}}^{[3]}, \mathsf{s}^2 \rangle = \tfrac{1}{\theta} u^{(0)} \;+\; 0 \cdot \epsilon \;+\; \tfrac{\theta-1}{\theta^2} \tau^2 \partial_x^2 u^{(0)} \, \epsilon^2 \;+\; 0 \cdot \epsilon^3$$

Obviously, the right hand side is minimized if we require $u^{(0)}$ to vanish at the boundaries, which yields Dirichlet conditions for $u^{(0)}$. Hence (3.53) may be used to approximate the target problem (3.32) endowed with homogeneous Dirichlet instead of periodic boundary conditions, i.e. $v(t,0) = v(t,L) = 0$. $\hat{\mathsf{f}}^{[3]}$ leaves a residual of magnitude $\mathrm{O}(\epsilon^2)$ since it is not possible to additionally require $\partial_x^2 u^{(0)} = 0$ which would result in an ill-posed problem for $u^{(0)}$. To obtain a higher order of consistency, we need another quantity that can compensate the $\partial_x^2 u^{(0)}$-term. This is achieved by taking $\hat{\mathsf{f}}^{[5]}$ where $u^{(2)}$ occurs in the associated asymptotic order function $\mathsf{f}^{(2)}$:

$$\mathsf{f}^{(2)} = u^{(2)} \mathsf{w} + \tau^2 \partial_x^2 u^{(2)} (\mathsf{s}^2 \mathsf{w} - \tfrac{1}{\theta} \mathsf{w})$$

$$\Rightarrow \quad \langle \hat{\mathsf{f}}^{[5]}, 1 \rangle = \tfrac{1}{\theta} u^{(0)} \;+\; 0 \cdot \epsilon \;+\; \big(\tfrac{1}{\theta} u^{(2)} + \tfrac{\theta-1}{\theta^2} \tau^2 \partial_x^2 u^{(0)} \big) \epsilon^2 \;+\; 0 \cdot \epsilon^3$$

So we may demand inhomogeneous Dirichlet conditions for $u^{(2)}$ by setting

$$u^{(2)} = -\tfrac{\theta-1}{\theta} \tau^2 \partial_x^2 u^{(0)} \quad \text{at the boundary.}$$

In this way, $\hat{\mathsf{f}}^{[5]}$ reaches a consistency order of at least 4, because $\langle \mathsf{f}^{(3)}, \mathsf{s}^2 \rangle$ must vanish due to lemma 3.4 ($\mathsf{f}^{(3)}$ has only terms containing s which is orthogonal to s^2). Let us finally note that the mass moment (3.52) of the example is recovered as sum of $u^{(0)} = v$ and $u^{(2)} = -\frac{\theta-1}{\theta} \tau^2 \partial_x^2 v = \frac{\theta-1}{\theta} \tau^2 \frac{g}{\nu}$. Further terms do not appear in the expansion of u because they involve derivatives that vanish for the time-independent quadratic polynomial v.

The second kinetic boundary condition (3.54) is analyzed analogously. Here we obtain

$$\langle \hat{\mathsf{f}}^{[3]}, \mathsf{s} \rangle = 0 \;-\; \tfrac{\tau}{\theta} \partial_x u^{(0)} \, \epsilon \;+\; 0 \cdot \epsilon^2 \;-\; \big(\tau^3 \tfrac{\theta-1}{\theta^2} \partial_x^3 u^{(0)} - \tau^2 \tfrac{1}{\theta} \partial_t \partial_x v \big) \, \epsilon^3 \,.$$

The residual becomes at most of magnitude $\mathrm{O}(\epsilon^3)$ if we set $\partial_x u^{(0)} = 0$ at the boundary. Hence we can employ (3.54) to approximate target problems with Neumann[19] boundary conditions. By using $\hat{\mathsf{f}}^{[5]}$, a higher order of consistency can be attained which yields an inhomogeneous Neumann condition for $u^{(2)}$.

Finally, an objection with regard to condition (3.53) may be thrown in: It is true that bounce-back type conditions are well adapted to the hyperbolic character of lattice-Boltzmann equations, since they couple populations referring to out-going and incoming characteristics at the boundary points. But from the perspective of

[19] Observe that $-\nu \partial_x v$ is interpreted as the diffusive flux which is inversely proportional to the gradient of the density to level out spatial density fluctuations.

the target equation, (3.53) is not a natural approach to realize Dirichlet boundary conditions upon the mass moment u. Instead, it would appear much more intuitive to require directly

$$\langle \mathsf{f}, 1\rangle = 0. \tag{3.55}$$

Since $1 \in \mathcal{F}$ is orthogonal to $\mathsf{s}^2\mathsf{w} - \frac{1}{\theta}\mathsf{w}$ in opposition to s^2, we infer $\langle \hat{\mathsf{f}}^{[3]}, 1\rangle = 0$ if $u^{(0)} = 0$. Similarly we get $\langle \hat{\mathsf{f}}^{[5]}, 1\rangle = 0$ if $u^{(0)} = 0$ and $u^{(2)} = 0$ at the boundary. This time, homogeneous Dirichlet conditions are also found for the second and even for higher order functions of the mass moment.

Against this backdrop, the exceptional role of the D1P2 ($\theta = 1$) model with respect to the accuracy of (3.52) becomes immediately comprehensible, because condition (3.53) and (3.55) coincide in this case as $\mathsf{s}^2 = 1$.

Remark 3.5. Both condition (3.53) and (3.55) are suitable for the D1P3 lattice-Boltzmann algorithm because there is only one population at each boundary node which is not updated by the standard evolution step. This population can be determined either by the bounce-back prescription or by solving a linear equation such that the mass moment remains zero. In two dimensions there are usually more than one population (even along a rectilinear boundary) where boundary conditions must be applied. In this case condition (3.55) yields only a single equation which is not enough to determine several unknown populations. In this situation bounce-back offers a reasonable possibility to cope with the under-determined problem.

Boundary conditions and energy estimate

Let us finally check whether the energy estimate in section 3.1 also works if we assume bounce-back type boundary conditions like (3.53) or (3.54) instead of periodic boundary conditions. The essential step in the proof of theorem 3.1, where the periodicity of f and $\hat{\mathsf{f}}$ has been exploited, is the elimination of the integral

$$\begin{aligned} I := \int_0^L \Big\langle B\big(\mathsf{f}(t,x) - \hat{\mathsf{f}}(t,x)\big),\ S\partial_x\big(\mathsf{f}(t,x) - \hat{\mathsf{f}}(t,x)\big)\Big\rangle\,\mathrm{d}x \\ = \int_0^L \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\Big\langle B\big(\mathsf{f}(t,x) - \hat{\mathsf{f}}(t,x)\big),\ S\big(\mathsf{f}(t,x) - \hat{\mathsf{f}}(t,x)\big)\Big\rangle\,\mathrm{d}x. \end{aligned}$$

in (3.11). We consider now the boundary conditions

$$\mathsf{f}_2(t,0) = \pm\mathsf{f}_1(t,0) \qquad \text{and} \qquad \mathsf{f}_2(t,L) = \pm\mathsf{f}_1(t,L),$$

which are supposed to be satisfied likewise by the comparison function $\hat{\mathsf{f}}$ up to a residual of magnitude $\mathrm{O}(\epsilon^\alpha)$. Then

$$\hat{\mathsf{f}}_2(t,0) = \pm\hat{\mathsf{f}}_1(t,0) + \mathrm{O}(\epsilon^\alpha) \qquad \text{and} \qquad \hat{\mathsf{f}}_2(t,L) = \pm\hat{\mathsf{f}}_1(t,L) + \mathrm{O}(\epsilon^\alpha)$$

Evaluating the integral by the fundamental theorem of calculus we obtain

$$I = \frac{1}{2}\Big\langle B\big(\mathsf{f} - \hat{\mathsf{f}}\big),\ S\big(\mathsf{f} - \hat{\mathsf{f}}\big)\Big\rangle\Big|_{(t,0)}^{(t,L)}$$

which becomes more explicitly

$$
\begin{aligned}
I &= \frac{1}{2}\left\langle \begin{pmatrix} 2\theta \ \big(\mathsf{f}_1 - \hat{\mathsf{f}}_1\big) \\ \frac{\theta}{\theta-1}\big(\mathsf{f}_0 - \hat{\mathsf{f}}_0\big) \\ 2\theta \ \big(\mathsf{f}_2 - \hat{\mathsf{f}}_2\big) \end{pmatrix}, \begin{pmatrix} -1\cdot\big(\mathsf{f}_1 - \hat{\mathsf{f}}_1\big) \\ 0\cdot\big(\mathsf{f}_0 - \hat{\mathsf{f}}_0\big) \\ +1\cdot\big(\mathsf{f}_2 - \hat{\mathsf{f}}_2\big) \end{pmatrix} \right\rangle \Bigg|_{(t,0)}^{(t,L)} \\
&= \theta\Big[(\mathsf{f}_2 - \hat{\mathsf{f}}_2)^2 - (\mathsf{f}_1 - \hat{\mathsf{f}}_1)^2\Big]_{(t,0)}^{(t,L)} \\
&= \theta\Big[(\pm 1)^2\big(\mathsf{f}_1 - \hat{\mathsf{f}}_1 + \mathrm{O}(\epsilon^2)\big)^2 - \big(\mathsf{f}_1 - \hat{\mathsf{f}}_1\big)^2\Big]_{(t,0)}^{(t,L)} \\
&= \mathrm{O}(\epsilon^\alpha).
\end{aligned}
$$

In contrast to the periodic case the integral usually does not become zero since it is not possible to construct comparison functions, which satisfy the non-periodic boundary conditions exactly. However, the integral vanishes in the limit of $\epsilon \downarrow 0$. Therefore it acts like a small additional source term, which shows that the an energy estimate similar to theorem 3.1 is basically possible.

Outlook – What else?

The following aspects of the presented D1P3 lattice-Boltzmann equation might be interesting for further research.

- Lemma 3.1 does not refer to an equilibrium operator which includes advection, e.g. $E_\epsilon \mathsf{f} = (1-\epsilon^2\tau c)\langle \mathsf{f}, 1\rangle \mathsf{w} + \epsilon a\theta\langle \mathsf{f}, 1\rangle \mathsf{ws}$, because then BE_ϵ is not symmetric. Here another approach might be necessary to close this gap.

- The energy estimate should be performed carefully for non-periodic boundary conditions particularly with respect to the resulting order of asymptotic similarity. Furthermore norms stronger than $\mathcal{L}^2$ should be taken into account to get also asymptotic similarity for the derivatives of f and $\hat{\mathsf{f}}$. What are the boundary conditions for $\partial_x \mathsf{f}, \partial_x^2 \mathsf{f}, \ldots$?

- Instead of (3.53) and (3.54) one might consider the boundary condition $\langle \mathsf{f}, \mathsf{s}^2 - \frac{1}{\theta}\rangle = 0$ which does not appear suspicious as it is analogue to the other moment conditions. However, the asymptotic analysis reveals quickly that it does not result in a macroscopic boundary conditions for $u^{(0)}$ or $\partial_x u^{(0)}$ but for $\partial_x^2 u^{(0)}$. This leads to an ill-posed problem for $u^{(0)}$, whose second spatial derivative cannot be prescribed as being a solution of a second order parabolic equation. If the boundary condition $\langle \mathsf{f}, \mathsf{s}^2 - \frac{1}{\theta}\rangle = 0$ is well-posed for the lattice-Boltzmann equation, how does the solution look like? Certainly, it cannot be approximated by means of a regular expansion alone (arising of boundary layers).

- Another question concerns the natural translation of Neumann boundary conditions. So $\partial_x v = 0$ suggests $0 = \partial_x u = \partial_x\langle \mathsf{f}, 1\rangle = \langle \partial_x \mathsf{f}, 1\rangle$. Where is the snag?

- Classical lattice-Boltzmann algorithms work on uniform (mostly quadratic or cubic) grids. Local grid refinement requires the coupling of grids with different mesh spacings (coarse grid $\leftrightarrow$ fine grid). The grid spacing of the discrete algorithm is related to the sacling parameter ϵ of the lattice-Boltzmann equation (see end of subsection 1.1.2, page 41). Therefore it might be interesting from the point of grid coupling to study the model problem for a locally constant but discontinuous parameter function which replaces ϵ.

- In the course of the computations particularly in section 3.2 it became noticeable, that the full information needed to determine an order function univocally was not directly found in the equations pertaining to this order but two orders later. This obliges to continue the expansion, which makes stronger regularity assumptions necessary concerning initial data and sources, as already observed at the end of section 2.3. What can be said, if these regularity assumptions are not available? We will encounter related problems in analyzing numerical schemes throughout the next chapters.

Chapter 4

Consistency of a D1P3 lattice-Boltzmann algorithm

In this chapter we analyze the D1P3 lattice-Boltzmann algorithm. Analogously to the D1P3 lattice-Boltzmann equation discussed in the previous chapter, this paradigmatic algorithm is suitable to pick up some general understanding of BGK lattice-Boltzmann schemes applying also to schemes in two or three dimensions.
Before introducing the D1P3 lattice-Boltzmann algorithm, we shortly present the target equation to be solved numerically, which is sometimes denoted synonymously as *limit* or *macroscopic* equation.

Target problem. Let $L \in (0, \infty)$ and $T \in (0, \infty]$ be arbitrary but fixed throughout this chapter. Furthermore the functions $w \in \mathcal{F}(\mathbb{R}/L\mathbb{R}, \mathbb{R})$ and $g \in \mathcal{F}([0, T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ are be given. w is called the *initial condition* while g plays the role of an *external source*. In the sequel the following *initial value problem* for the advection-diffusion-reaction equation

$$\begin{aligned} u(0, \cdot) &= w \\ \partial_t u + a\partial_x u - \nu\partial_x^2 u + cu &= g \end{aligned} \tag{4.1}$$

serves as target problem for our D1P3 lattice-Boltzmann algorithm. Its solution $u \in \mathcal{F}([0, T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ as well as w and g are supposed to be periodic functions on $[0, L]$ wherefore they are introduced as functions over $\mathbb{R}/\mathbb{R}$.
As far as the three constants are concerned, the *reactivity* c and the *diffusivity* ν satisfy the condition $c \geq 0$ and $\nu > 0$ respectively while the *advection velocity* a may be a signed.
Depending on the concrete physical problem, that might be hidden behind the target problem, u could correspond to a concentration or density, to a temperature or to the velocity profile of a shear flow as shown in section 1.2.

Ingredients of the lattice-Boltzmann algorithm. The *finite velocity space* of the D1P3 model is given by the set

$$\mathcal{S} := \{-1, 1, 0\} \equiv \{\ominus, \oplus, 0\} \qquad .$$

Obviously $\mathcal{S}$ satisfies the reflection property, i.e. $\mathcal{S} = -\mathcal{S}$. With exception of the equilibrium (see below), we denote by $\mathsf{sans\text{-}serif}$ letters those quantities, which can

0^{th} moment	odd moments	even moments
$\sum_{s\in\mathcal{S}} \mathcal{D}_s(u) = u$	$\sum_{s\in\mathcal{S}} \mathcal{D}_s(u)\, s^{2m-1} = 0$	$\sum_{s\in\mathcal{S}} \mathcal{D}_s(u)\, s^{2m} = \frac{1}{\theta}\, u$
$\sum_{s\in\mathcal{S}} \mathcal{A}_s(u) = 0$	$\sum_{s\in\mathcal{S}} \mathcal{A}_s(u)\, s^{2m-1} = a\, u$	$\sum_{s\in\mathcal{S}} \mathcal{A}_s(u)\, s^{2m} = 0$
$\sum_{s\in\mathcal{S}} \mathcal{R}_s(u) = -\tau c u$	$\sum_{s\in\mathcal{S}} \mathcal{R}_s(u)\, s^{2m-1} = 0$	$\sum_{s\in\mathcal{S}} \mathcal{R}_s(u)\, s^{2m} = -\frac{1}{\theta}\, \tau\, c\, u$

Table 4.1: Moments of the equilibrium.

be considered as real-valued functions over $\mathcal{S}$. The argument is written in index notation. Let us also introduce the *multiplication-operator* $S : \mathcal{F}(\mathcal{S}, \mathbb{R}) \to \mathcal{F}(\mathcal{S}, \mathbb{R})$, defined by $(S\,\mathsf{f})_s := s\mathsf{f}_s$.

Furthermore, the *weights* or *rest populations* $\mathsf{w} \in \mathcal{F}(\mathcal{S}, \mathbb{R})$ are given by

$$\mathsf{w}_{\ominus} := \frac{1}{2\theta} \qquad \mathsf{w}_0 := \frac{\theta - 1}{\theta} \qquad \mathsf{w}_{\oplus} := \frac{1}{2\theta} \qquad \theta \geq 1$$

which is summarized into a single formula by $\mathsf{w}_s = (1 - s^2)\frac{\theta-1}{\theta} + s^2 \frac{1}{2\theta}$. The weights must add up to 1; furthermore $\theta \geq 1$ is required to ensure their non-negativity. For $\theta = 1$ the D1P3 algorithm reduces to the D1P2 algorithm. The following *structure relations* are found:

$$\sum_{s\in\mathcal{S}} \mathsf{w}_s = 1 \qquad \sum_{s\in\mathcal{S}} \mathsf{w}_s\, s^{2m-1} = 0 \qquad \sum_{s\in\mathcal{S}} \mathsf{w}_s\, s^{2m} = \frac{1}{\theta} \qquad m \in \mathbb{N}\,.$$

It should be recalled that specific terms and parameters of the target equation can be tuned by the equilibrium[1]. Here, the equilibrium $\mathcal{E}_\epsilon : \mathcal{S} \times \mathbb{R} \to \mathbb{R}$ is composed of three terms; a mandatory *diffusive* part $\mathcal{D}$ as well as an optional *advective* and *reactive* part $\mathcal{A}, \mathcal{R}$ vanishing for $a = 0$ and $c = 0$ respectively.

$$\begin{aligned} \mathcal{E}_\epsilon(u) &:= \mathcal{D}(u) + \epsilon\, \mathcal{A}(u) + \epsilon^2\, \mathcal{R}(u) && (4.2) \\ &= u\,\mathsf{w} + \epsilon\, \theta a u\, S\,\mathsf{w} - \epsilon^2\, \tau c u\, \mathsf{w} && (4.3) \end{aligned}$$

Later, the scaling parameter ϵ is substituted by grid spacing while τ represents the *collision time*, that turns out to be related to the diffusivity of the target equation (cf. next paragraph). To be available as reference for computations, the *moments*[2] of the equilibrium are listed in table 4.1.

[1]We consider here an equilibrium, generalizing the D1P3 equilibrium, that we have originally derived in section 1.2 from the D2P9 Stokes equilibrium and that leads only to the diffusion equation. By means of the formal asymptotic analysis it was found, how the equilibrium should be modified to produce an additional advective and reactive term in the limit equation. It is important to scale these additional terms $\mathcal{A}, \mathcal{R}$ correctly with respect to the grid spacing. Thus, the equilibrium becomes dependent on another parameter denoted here by ϵ.

[2]If we multiply an s-dependent quantity by s^k, with $k \in \mathbb{N}_0$, and sum over $s \in \mathcal{S}$, we obtain the k'th *moment* of this quantity. This name is chosen in analogy to the theory of the Boltzmann equation, where the three lowest moments correspond to physical quantities (density, momentum or velocity and kinetic energy or temperature).

The lattice-Boltzmann algorithm for the target problem. For the constant L specified by the target problem let us define N_L as the smallest integer greater or equal than L. The spatial domain $\mathbb{R}/L\mathbb{R}$ is covered by a uniform mesh

$$\mathcal{L}_h := \left\{x_i = ih : \ \ i \in \mathbb{Z}/N\mathbb{Z}\right\} .$$

with the mesh-size $h := \frac{L}{N}$ and $N \in \mathbb{N}_L := \{N_L, N_L + 1, ...\}$. So, N_L corresponds just to the minimal number of mesh nodes, for we want to admit only mesh spacings $h \leq 1$. Note, that h can attain only discrete values out of the set $\mathcal{H}_L := \{\frac{L}{N} : N \in \mathbb{N}_L\}$; for given h the number of spatial grid nodes is denoted by N_h. Because of its additive group structure it is advantageous to choose $\mathbb{Z}/N\mathbb{Z}$ as spatial index set[3].
It is important to notice, that the lattice-Boltzmann algorithm requires a strict connection between the temporal and spatial discretization parameter. The time increment corresponding to one iteration (time-step) is given by h^2, i.e. the square of the spatial mesh-size. If the maximal time T is finite, the number of time-steps, that can be performed on a spatial grid with N_h nodes, before T has been exceeded, is given by $N_T(h) := [T/h^2]$, with the square brackets denoting the greatest integer less or equal than T/h^2. The time-interval $[0, T]$ is thus discretized by

$$\mathcal{T}_h := \left\{t_n = nh^2 : \ \ n \in \mathbb{T}_h\right\} ,$$

where the index-set of the iterations is $\mathbb{T}_h := \{0, 1, ..., N_T(h)\}$ if T is finite and $\mathbb{T}_h := \mathbb{N}_0$ if $T = \infty$. So we obtain the time-space grid $\mathcal{G}_h := \mathcal{T}_h \times \mathcal{L}_h$.

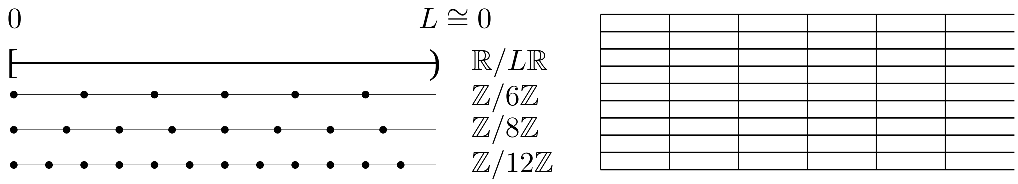


Figure 4.1: Left: Spatial domain. In $\mathbb{R}/L\mathbb{R}$, the numbers 0 and L belong to the same equivalence class. Therefore the point with coordinate L is identified with the origin 0, providing thus in a natural way the periodic boundary conditions. As for the discretization of $\mathbb{R}/L\mathbb{R}$, the left neighbor of the leftmost node corresponds to the rightmost node and vice versa. Right: Time-space grid with 6 nodes and 9 iterations. The time advances only with h^2 per iteration.

The lattice-Boltzmann algorithm is a prescription to compute the population or particle function $\mathsf{F}_h \in \mathcal{F}(\mathcal{S} \times \mathbb{T}_h \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R})$ being the primary variable. The subscript h refers both to the explicit h-dependence of the numerical solution as well as to the implicit one due to the h-dependence of its domain of definition.
There are several possibilities to initialize F_h. Classically[4], this is done by means of the equilibrium

$$\mathsf{F}_h(0, i) \ = \ \mathcal{E}_h\big(w(x_i)\big) , \tag{4.4}$$

[3]Thanks to the additive group structure of $\mathbb{Z}/N\mathbb{Z}$ we have not to distinguish between interior nodes and boundary nodes in equation (4.5) as it would be the case for the index set $\{1, ..., N\}$. The indexation is therefore quite suggestive, since periodic boundary conditions are – strictly speaking – no boundary conditions.

[4]We will see later, that the initialization by the equilibrium is not recommandable. In fact, it represents a rather crude initialization triggering more or less strong initial layers, that can be

where w is the initial condition of the target problem (4.1). The evolution is determined by the *discrete lattice-Boltzmann equation* coming along as a sort of difference equation.

$$\mathsf{F}_{h,s}(n+1,i+s) = (1-\omega)\mathsf{F}_{h,s}(n,i) + \omega\mathcal{E}_{h,s}\big(U_h(n,i)\big) + h^2\, g(t_n,x_i)\,\mathsf{w}_s \tag{4.5}$$

The parameter ω is called the *collision frequency*; its inverse $\tau := \omega^{-1}$, which will also occur quite frequently, is referred to as the *collision time*. The argument of the equilibrium

$$U_h := \sum_{s\in\mathcal{S}} \mathsf{F}_{h,s}$$

is the *zeroth moment* and represents the actual *quantity of interest* in contrast to the primary variable, because of the following connection with the solution of the target problem.

Assertion: *The zeroth moment approximates the solution of the target problem*

$$u(t_n,x_i) \;\approx\; U_h(n,i)\;, \tag{4.6}$$

i) if the diffusivity of the target equation is related to the collision[5] *time by*

$$\nu = \tfrac{1}{\theta}(\tau - \tfrac{1}{2})\;, \tag{4.7}$$

ii) if the parameters a, c of the equilibrium are the advection speed and reactivity of the target equation,

iii) if the population function is initialized by the initial condition of the target problem according to eqn.(4.4).

It is the intention of the following two sections to make this allegation more precise and justify it eventually.

Remark 4.1. (Notation) If the index s is omitted, we consider $\mathsf{w}, \mathcal{E}(u), \mathsf{F}(n,i)$ etc. as functions over the finite velocity space $\mathcal{S}$. In this case one may take these quantities alternatively as three-tuples (triples). $\mathsf{w}_s, \mathcal{E}_s(u), \mathsf{F}_s(n,i)$ signify the component associated with the velocity $s \in \mathcal{S}$. Consequently the multiplication-operator S should be thought as a diagonal matrix. In order to emphasize the functional dependence of F_h from the discrete velocities (analogous to the distribution function of the Boltzmann equation), we abstain from the standard index notation thus avoiding any artificial numbering.

4.1 Formal expansion

This section is devoted to a formal analysis of the lattice-Boltzmann algorithm based on a regular expansion with respect to the discretization parameter h. Formal means, that we do not care about questions of existence, convergence and

avoided by taking additional terms into account to achieve a higher order of consistency. However, these terms require the knowledge of spatial derivatives of w necessitating numeric differentiation or special initialization routines if they cannot be computed analytically.

[5] In particular, τ must be greater than $\frac{1}{2}$, in order to ensure a positive diffusivity.

smoothness. Even if we do not reach rigorous statements about the convergence, we gain valuable insight into the structure of the algorithm. The main result consists in deriving a limit equation, that is a priori not evident. Furthermore, we obtain also strong indications about the convergence rate of the scheme and its inherent error. The major objective consists in acquiring an idea and feeling of how the scheme behaves. Therefore the computations of this section serve to find a suitable asymptotic ansatz, which will be used in section 4.3 for a more rigorous treatment.

Working Hypothesis: Let us assume that there exists a function $\mathsf{f}_\epsilon \in \mathcal{F}(\mathcal{S} \times [0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ depending analytically on the parameter, ϵ which *interpolates* the population function of the lattice-Boltzmann algorithm on arbitrary grids specified by $h \in \mathcal{H}_L$:

$$\mathsf{F}_{h,s}(n,i) \;=\; \mathsf{f}_{h,s}\big(t_n(h), x_i(h)\big)\;.$$

This assumption is reasonable if the domain is equipped with periodic boundary conditions (or corresponds to the full space) but turns out to be generally wrong on bounded domains (see the study of numerical boundary layers in chapter 7). Moreover we assume, that f_ϵ depends on ϵ via a *regular* expansion (power series).

$$\mathsf{f}_\epsilon \;=\; \mathsf{f}^{(0)} \;+\; \epsilon\,\mathsf{f}^{(1)} \;+\; \epsilon^2\,\mathsf{f}^{(2)} \;+\; \ldots$$

The coefficient functions $\mathsf{f}^{(k)} \in \mathcal{F}(\mathcal{S} \times [0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ with $k \in \mathbb{N}_0$ are designated as the *asymptotic orders*. They are supposed to be smooth ϵ-independent functions of time and space. Let us also introduce the zeroth moment of the regular expansion

$$u^{(k)} := \sum_{s\in\mathcal{S}} \mathsf{f}^{(k)}_s \qquad u_\epsilon := \sum_{s\in\mathcal{S}} \mathsf{f}_{\epsilon,s} = \sum_k u^{(k)}\,\epsilon^k\,.$$

In order to avoid the tedious distinction of special cases, it is reasonable to introduce $u^{(-2)} \equiv 0$ and $u^{(-1)} \equiv 0$. Furthermore we will make use of the following shortcuts.

$$\left.\begin{array}{c}\mathcal{D}^{(k)}\\ \mathcal{A}^{(k)}\\ \mathcal{R}^{(k)}\end{array}\right\} := \left\{\begin{array}{l}\mathcal{D}(u^{(k)})\\ \mathcal{A}(u^{(k)})\\ \mathcal{R}(u^{(k)})\end{array}\right. \qquad \mathcal{E}^{(k)} := \mathcal{D}^{(k)} + \mathcal{A}^{(k-1)} + \mathcal{R}^{(k-2)} \tag{4.8}$$

Due to the linearity of $\mathcal{E}_\epsilon$ we obtain $\mathcal{E}_\epsilon\Big(\sum_{\alpha\geqslant 0}\epsilon^\alpha\,u^{(\alpha)}\Big) \;=\; \sum_{\alpha\geqslant 0}\epsilon^\alpha\,\mathcal{E}^{(\alpha)}\;.$

Comparison of Coefficients. Let us plug f_ϵ into the discrete lattice-Boltzmann equation (4.5). Setting $\epsilon = h$ and replacing the indices by the coordinates of the time-space grid nodes, they stand for, we get

$$\mathsf{f}_{h,s}(t_n + h^2, x_i + sh) = (1-\omega)\,\mathsf{f}_{h,s}(t_n, x_i) + \omega\,\mathcal{E}_{h,s}\big(u_h(t_n,x_i)\big) + h^2\,\mathsf{w}_s\,g(t_n,x_i)\,.$$

If we can construct coefficients $\mathsf{f}^{(k)}$ such that this equation holds for We assume, arbitrary $t \in [0,T]$, $x \in \mathbb{R}/L\mathbb{R}$ and $\epsilon \geq 0$

$$\mathsf{f}_{\epsilon,s}(t+\epsilon^2, x+s\epsilon) \;=\; (1-\omega)\,\mathsf{f}_{\epsilon,s}(t,x) + \omega\,\mathcal{E}_{\epsilon,s}\big(u_\epsilon(t,x)\big) + \epsilon^2\,\mathsf{w}_s\,g(t,x)\,, \tag{4.9}$$

it will also satisfy the lattice-Boltzmann equation in the grid. Substituting $\mathsf{f}_{\epsilon,s}(t+\epsilon^2, x+s\epsilon)$ by its formal Taylor series around t, x yields:

$$\begin{aligned}
\mathsf{f}_{\epsilon,s}(t+\epsilon^2, x+s\epsilon) &= \sum_{\beta,\gamma\geq 0} \epsilon^{2\beta+\gamma} \frac{1}{\beta!\,\gamma!}\,\partial_t^\beta\,(s\partial_x)^\gamma\,\mathsf{f}_{\epsilon,s}(t,x) \\
&= \sum_{\alpha\geq 0} \epsilon^\alpha \sum_{\substack{2\beta+\gamma=\alpha\\ \beta,\gamma\geq 0}} \frac{1}{\beta!\,\gamma!}\,\partial_t^\beta\,(s\partial_x)^\gamma\,\mathsf{f}_{\epsilon,s}(t,x) \\
&= \sum_{\alpha\geq 0} \epsilon^\alpha\, T_\alpha(\partial_t, s\partial_x)\mathsf{f}_{\epsilon,s}(t,x)
\end{aligned}$$

A rearrangement has been performed in the second line, such that equal powers of ϵ are collected. The coefficient behind ϵ^α is a sum of diverse derivatives of $\mathsf{f}_{\epsilon,s}$ with respect to $t$and x. To avoid these cumbersome expressions, we have implicitly introduced the following polynomials,

$$T_\alpha(\vartheta,\varsigma) := \sum_{\substack{2\beta+\gamma=\alpha\\ \beta,\gamma\geq 0}} \frac{1}{\beta!\,\gamma!}\,\vartheta^\beta\,\varsigma^\gamma \qquad \alpha\in\mathbb{N}_0. \tag{4.10}$$

which come into play by the <u>T</u>aylor expansion due to the coupling of the differently scaled time and space increment. Equation (4.9) can then be written in the form:

$$\sum_{\alpha\geq 0} \epsilon^\alpha\, T_\alpha(\partial_t, S\partial_x)\,\mathsf{f}_\epsilon \;=\; (1-\omega)\,\mathsf{f}_\epsilon \;+\; \omega\,\mathcal{E}_\epsilon(u_\epsilon) \;+\; \epsilon^2\,g\,\mathsf{w}$$

For brevity we will often write T_α instead of $T_\alpha(\partial_t, S\partial_x)$. Remarking that the first term of the sum ($\alpha=0$) is equal to f_ϵ, since $T_0=1$, this term can be canceled. Dividing by $\omega=\tau^{-1}$ we arrive at:

$$\mathsf{f}_\epsilon \;=\; \mathcal{E}_\epsilon(u_\epsilon) \;-\; \tau\sum_{\alpha\geq 1}\epsilon^\alpha\, T_\alpha(\partial_t, S\partial_x)\,\mathsf{f}_\epsilon \;+\; \epsilon^2\,\tau\,g\,\mathsf{w}$$

Let us now replace $\mathsf{f}_\epsilon, u_\epsilon$ by their expansions, in order to obtain an equation, that is *fully expanded* with respect to ϵ, such that we are allowed to perform the *comparison of coefficients*.[6]

$$\sum_{\beta\geq 0}\epsilon^\beta\,\mathsf{f}^{(\beta)} \;=\; \sum_{\beta\geq 0}\epsilon^\beta\,\mathcal{E}^{(\beta)} \;-\; \tau\sum_{\alpha\geq 1}\sum_{\beta\geq 0}\epsilon^{\alpha+\beta}\,T_\alpha\,\mathsf{f}^{(\beta)} \;+\; \epsilon^2\,\tau\,g\,\mathsf{w} \tag{4.11}$$

Equating terms containing equal power of ϵ, yields the following infinite system, connecting the asymptotic orders with each other.

$$\begin{aligned}
\mathsf{f}^{(n)} &= \mathcal{E}^{(n)} - \tau \sum_{\substack{\alpha\geq 1,\,\beta\geq 0\\ \alpha+\beta=n}} T_\alpha\,\mathsf{f}^{(\beta)} + \delta_{2n}\,\tau\,g\,\mathsf{w} \qquad n\geq 0 \\
\Leftrightarrow\quad \mathsf{f}^{(n)} &= \mathcal{E}^{(n)} - \tau\sum_{\alpha=1}^{n} T_\alpha\,\mathsf{f}^{(n-\alpha)} + \delta_{2n}\,\tau\,g\,\mathsf{w}
\end{aligned} \tag{4.12}$$

[6] Let $a_k, b_k\in\mathbb{C}$ for $k\in\mathbb{N}_0$. Two power series $A_\epsilon := \sum_k a_k\epsilon^k$ and $B_\epsilon := \sum_k b_k\epsilon^k$ are identical, i.e. equal for all $\epsilon\in\mathbb{C}$, if and only if the coefficients are equal, i.e. $a_k=b_k$ for all $k\in\mathbb{N}_0$. The direction "$\Leftarrow$" is trivial, while "$\Rightarrow$" is seen by the following iterative/repetitive process: The equality of the power series implies in the special case of $\epsilon=0$: $a_0=b_0$. If we know already $a_k=b_k$ for $k\leq n$ then consider $(A_\epsilon - \sum_{k\leq n} a_k\epsilon^k)/\epsilon^{n+1} = \sum_{k>n} a_k\epsilon^{k-n-1}$ and $(B_\epsilon - \sum_{k\leq n} b_k\epsilon^k)/\epsilon^{n+1} = \sum_{k>n} b_k\epsilon^{k-n-1}$. Since these expressions are equal by assumption for all ϵ, this yields by choosing $\epsilon=0$: $a_{n+1}=b_{n+1}$.

Since $\mathsf{f}^{(n)}$ depends, by this formula, only on preceding asymptotic order functions, the system is *triangular* and can successively be solved for the $\mathsf{f}^{(n)}$'s. Let us evaluate the first two orders explicitly:

$$\begin{aligned} \mathsf{f}^{(0)} &= \mathcal{E}^{(0)} = \mathcal{D}^{(0)} = u^{(0)}\,\mathsf{w} && (4.13)\\ \mathsf{f}^{(1)} &= \mathcal{E}^{(1)} - \tau\, T_1(\partial_t, s\partial_x)\,\mathsf{f}^{(0)} = \mathcal{D}^{(1)} + \mathcal{A}^{(0)} - \tau\, s\partial_x\,\mathcal{D}^{(0)} \\ &= u^{(1)}\mathsf{w} + \theta\, a u^{(0)}\, S\mathsf{w} - \tau\,\partial_x u^{(0)} S\,\mathsf{w} && (4.14)\end{aligned}$$

Inserting these expressions into (4.12) for $n = 2$, leads to a similar but longer expression, that gives $\mathsf{f}^{(2)}$ in terms of $u^{(2)}, u^{(1)}, u^{(0)}$ and the source g, which did not yet appear in the above formulae for $\mathsf{f}^{(0)}, \mathsf{f}^{(1)}$. So the asymptotic orders of the population function should be expressed in terms of the external source function g and of the asymptotic orders pertaining to the zeroth moment. This is surprising, since the asymptotic orders of the *three unknown* functions $\mathsf{f}_{h,\ominus}, \mathsf{f}_{h,\oplus}, \mathsf{f}_{h,0}$ are basically constructed from the asymptotic orders of a *single unknown* function, namely u_h. However, this gives rise to the *question*, what equations determine the evolution of the $u^{(k)}$'s.

Before we come to the answer of this problem, we want to establish a general representation formula of the $\mathsf{f}^{(n)}$'s. Since the expressions become rather long with increasing n, we introduce another class of polynomials $P_m = P_m(\tau, \vartheta, \varsigma)$ defined recursively by

$$P_0 := 1 \qquad \text{for } m \geqslant 1 \quad P_m := -\tau \sum_{j=1}^{m} T_j P_{m-j} \tag{4.15}$$

On the left hand side there are only polynomials of degree less than m; so P_m is well determined. For ease of notation, we define $P_{-1} \equiv 0$ and $P_{-2} \equiv 0$.

Proposition 4.1. *The asymptotic orders of* f_ϵ *have the following representation,*

$$\mathsf{f}^{(n)} = \sum_{k=0}^{n} P_{n-k}\,\mathcal{E}^{(k)} + \tau\, P_{n-2}\, g\,\mathsf{w} \qquad n \geq 0 \tag{4.16}$$

where P_{n-k}, P_{n-2} *abbreviate* $P_{n-k}(\tau, \partial_t, S\partial_x)$ *respectively* $P_{n-2}(\tau, \partial_t, S\partial_x)$.

Even if (4.16) looks like an explicit formula for the $\mathsf{f}^{(n)}$'s, contrasting the recursive representation (4.12), the recursion is only shifted to the definition of the polynomials (4.15). Nevertheless, (4.16) is advantageous, since, first of all, the recursion relation for polynomials is practically much easier to execute, for instance by computer algebra packages. Secondly, from the theoretical point of view, we can attempt to exploit algebraic properties of the P_m's to discover features of the recursion, which would remain mysterious otherwise.

Proof: In order to start a proof by induction, we check that (4.16) is correct for $n = 0$:

$$\mathsf{f}^{(0)} = \sum_{k=0}^{0} P_{-k}\,\mathcal{E}^{(k)} + \tau\, P_{-2}\, g = P_0\,\mathcal{E}^{(0)} + 0 = \mathcal{E}^{(0)}$$

Obviously, this result agrees with equation (4.12) setting $n = 0$. Let us assume that (4.16) is true for all $m \in \mathbb{N}_0$ with $m \leq n \in \mathbb{N}_0$. We have to verify that (4.16) holds then also for $n + 1$, i.e.:

$$\mathsf{f}^{(n+1)} \;=\; \sum_{k=0}^{n+1} P_{n+1-k}\, \mathcal{E}^{(k)} \;+\; \tau\, P_{n-1}\, g$$

Inserting in (4.12) the induction hypothesis for $\mathsf{f}^{(0)}, ..., \mathsf{f}^{(n)}$, we get:

$$\begin{aligned}
\mathsf{f}^{(n+1)} &= \mathcal{E}^{(n+1)} - \tau \sum_{k=1}^{n+1} T_k\, \mathsf{f}^{(n+1-k)} \;+\; \delta_{2(n+1)}\, \tau\, g\, \mathsf{w} \\
&= \mathcal{E}^{(n+1)} - \tau \sum_{k=1}^{n+1} T_k \Big\{ \sum_{j=0}^{n+1-k} P_{n+1-k-j}\, \mathcal{E}^j \;+\; \tau\, P_{n+1-k-2}\, g\, \mathsf{w} \Big\} + \tau \delta_{2(n+1)}\, g\, \mathsf{w}
\end{aligned}$$

Introducing the quantities

$$\begin{aligned}
A_{n+1} &:= \mathcal{E}^{(n+1)} \;-\; \tau \sum_{k=1}^{n+1} \sum_{j=0}^{n+1-k} T_k\, P_{n+1-k-j}\; \mathcal{E}^j \\
B_{n+1} &:= -\,\tau^2 \sum_{k=1}^{n+1} T_k\, P_{n+1-k-2}\; g\, \mathsf{w} \;+\; \delta_{(n+1)\,2}\, \tau\, g\, \mathsf{w}
\end{aligned}$$

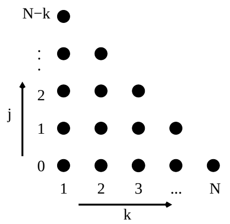
with $\mathsf{f}^{(n+1)} = A_{n+1} + B_{n+1}$, it remains to be shown for $n \geq 0$:

$$A_{n+1} \;=\; \sum_{k=0}^{n+1} P_{n+1-k}\, \mathcal{E}^k\,, \qquad\qquad B_{n+1} \;=\; \tau\, P_{n-1}\, g\, \mathsf{w}\,.$$

Regarding the first relation, we permute the summations (see figure 4.2) to reduce the double sum into a simple sum utilizing (4.15). Recall that $P_0 = 1$ while all other polynomials of degree ≥ 1 are recursively defined.

$$\begin{aligned}
A_{n+1} \;&=\; \mathcal{E}^{(n+1)} \;-\; \tau \sum_{k=1}^{n+1} \sum_{j=0}^{n+1-k} T_k\, P_{n+1-k-j}\, \mathcal{E}^{(j)} \\
&=\; \mathcal{E}^{(n+1)} \;-\; \tau \sum_{j=0}^{n} \sum_{k=1}^{n+1-j} T_k\, P_{n+1-k-j}\, \mathcal{E}^{(j)} \\
&=\; \mathcal{E}^{(n+1)} \;+\; \sum_{j=0}^{n} \Big[-\tau \sum_{k=1}^{n+1-j} T_k\, P_{n+1-j-k} \Big]\, \mathcal{E}^{(j)} \\
&=\; P_0\, \mathcal{E}^{(n+1)} \;+\; \sum_{j=0}^{n} P_{n+1-j}\, \mathcal{E}^{(j)} \;=\; \sum_{j=0}^{n+1} P_{n+1-j}\, \mathcal{E}^{(j)}
\end{aligned}$$

Let us now turn to the second term B_{n+1}. Due to the Kronecker-δ term and the



set $N := n+1$

$$\sum_{k=1}^{N}\sum_{j=0}^{N-k} = \sum_{j=0}^{N-1}\sum_{k=1}^{N-j}$$

Figure 4.2: Swapping the sums.

non-applicability of the recursion relation (4.15) for $n = 0$, we have to distinguish three situations. Recall that $P_{-1} = P_{-2} = 0$ by definition.

$$\begin{aligned}
n = 0: \qquad & B_{0+1} &=& -\tau^2 T_1 P_{-2}\, g\,\mathsf{w} \;+\; \delta_{21}\,\tau\, g\,\mathsf{w} \;=\; 0 \;=\; \tau P_{0-1}\, g\,\mathsf{w} \\
n = 1: \qquad & B_{1+1} &=& -\tau^2\big(T_1 P_{-1} + T_2 P_{-2}\big)\, g\,\mathsf{w} \;+\; \tau\,\delta_{2\,2}\, g\,\mathsf{w} \\
& &=& 0 + \tau\, g\,\mathsf{w} \;=\; \tau P_0\, g\,\mathsf{w} \;=\; \tau P_{1-1}\, g\,\mathsf{w} \\
n \geqslant 2: \qquad & B_{n+1} &=& -\,\tau^2 \sum_{k=1}^{n+1} T_k P_{n+1-k-2}\, g\,\mathsf{w} \;+\; \delta_{(n+1)2}\,\tau\, g\,\mathsf{w} \\
& &=& -\,\tau^2 \sum_{k=1}^{n-1} T_k P_{n-1-k}\, g\,\mathsf{w} \;+\; 0 \;\stackrel{(4.15)}{=}\; \tau P_{n-1}\, g\,\mathsf{w}
\end{aligned}$$

■

The polynomials T_α exhibit a well defined *parity*[7] with respect to ς, which they owe to the fact, that the temporal increment in equation (4.9) is the square of the spatial increment. This quite simple algebraic property is - besides the reflection property of the finite velocity set $\mathcal{S}$ - responsible for the so-called *odd-even decoupling* (cf. remark 4.2) of the evolution equations for the asymptotic orders of u_ϵ. So it ultimately gives rise to the second order convergence behavior of the lattice-Boltzmann algorithm.

Lemma 4.1. *If $\alpha \in \mathbb{N}_0$ is either* even *or* odd *respectively, then $\varsigma \mapsto T_\alpha(\vartheta, \varsigma)$ is an* even *or* odd *polynomial, i.e. $T_\alpha(\vartheta, -\varsigma) = (-1)^\alpha T_\alpha(\vartheta, \varsigma)$.*

Proof: Any even integer can only be written as sum of two integers, if both are either even or odd. Likewise, if an odd integer is represented as the sum of two integers, then one addend must be even, while the other is odd. Since T_α contains only monomials of the form $\vartheta^\beta \varsigma^\gamma$ with $2\beta + \gamma = \alpha$, we see that γ is even (odd) if and only if α is even (odd). Therefore all occurring monomials are either even or odd with respect to ς. ■

The symmetry behavior of the T_m's under the reflection of ς at the origin carries over to the P_m's. Thus, T_m and P_m are even (odd) polynomials, if m is even (odd).

[7] A function $f : \mathbb{R} \to \mathbb{R}$ is said to be of *even* parity, if it is symmetric with respect to reflection at the origin, i.e. $f(-x) = f(x)$. It is of *odd* parity if it is antisymmetric, i.e. $f(-x) = -f(x)$. A polynomial $p(x)$ is of even (odd) parity, iff it contains only even (odd) powers of x [even (odd) polynomial].

Lemma 4.2. *If $l \in \mathbb{N}_0$ is* even *(*odd*), then $\varsigma \mapsto P_m(\tau,\vartheta,\varsigma)$ is an* even *(*odd*) polynomial, i.e. $P_m(\tau,\vartheta,-\varsigma) = (-1)^m P_m(\tau,\vartheta,\varsigma)$.*

Proof: The verification is done by induction. Since $P_0(\tau,\vartheta,-\varsigma) = (-1)^0 \cdot 1 = 1 = P_0(\tau,\vartheta,\varsigma)$, the assertion is true for $m = 0$. Let $m \in \mathbb{N}$ and suppose that the assetion is true for all $j \in \mathbb{N}_0$ with $j < m$. Then:

$$\begin{aligned}
P_m(\tau,\vartheta,-\varsigma) &= -\tau \sum_{j=1}^{m} T_j(\vartheta,-\varsigma)\; P_{m-j}(\tau,\vartheta,-\varsigma) \\
&= -\tau \sum_{j=1}^{m} (-1)^j\, T_j(\vartheta,-\varsigma)\; (-1)^{m-j}\, P_{m-j}(\tau,\vartheta,-\varsigma) \\
&= (-1)^m\, P_m(\tau,\vartheta,\varsigma)
\end{aligned}$$

■

$T_0(\vartheta,\varsigma) = 1$	$P_0(\tau,\vartheta,\varsigma) = 1$
$T_1(\vartheta,\varsigma) = \varsigma$	$P_1(\tau,\vartheta,\varsigma) = -\tau\varsigma$
$T_2(\vartheta,\varsigma) = \vartheta + \frac{1}{2}\varsigma^2$	$P_2(\tau,\vartheta,\varsigma) = \tau(\tau-\frac{1}{2})\varsigma^2 - \tau\vartheta$
$T_3(\vartheta,\varsigma) = \vartheta\,\varsigma + \frac{1}{6}\varsigma^3$	$P_3(\tau,\vartheta,\varsigma) = -\tau(\tau^2-\tau+\frac{1}{6})\varsigma^3 - \tau(2\tau-1)\vartheta\varsigma$
$T_4(\vartheta,\varsigma) = \frac{1}{2}\vartheta^2 + \vartheta\varsigma^2 + \frac{1}{24}\varsigma^4$	$P_4(\tau,\vartheta,\varsigma) = \tau(\tau^3-\frac{3}{2}\tau^2+\frac{7}{12}\tau-\frac{1}{24})\varsigma^4 + \tau(\tau-\frac{1}{2})\varsigma^2 - \tau(3\tau^2-3\tau+\frac{1}{2})\vartheta\varsigma^2$

Table 4.2: The first five representatives of the two polynomial families.

Fortunately, for our further arguments (cf. theorem 4.2) the explicit knowledge of the polynomials, with exception of P_0, P_1 and P_2, is not at all necessary. In order to illustrate, how the length of the P_m's grows with increasing index and to have them at our disposal for some later examples, table 4.2 presents a compilation up to forth order.

The following notation turns out to be quite economic in the sequel,

$$\langle P_k \rangle := \sum_{s\in\mathcal{S}} \mathsf{w}_s\, P_k(\tau,\partial_t,s\partial_x) \qquad \langle sP_k \rangle := \sum_{s\in\mathcal{S}} \mathsf{w}_s\, s\, P_k(\tau,\partial_t,s\partial_x)$$

when we have to compute frequently generalized moments of the equilibrium contained in table 4.3. For the proof of theorem 4.2 we have to know the following differential operators explicitly:

$$\langle P_0 \rangle = 1 \qquad \langle P_2 \rangle = \tau\big(\tfrac{1}{\theta}(\tau-\tfrac{1}{2})\partial_x - \partial_t\big) \qquad \langle sP_1 \rangle = -\tau\tfrac{1}{\theta}\partial_x \tag{4.17}$$

$\sum_{s\in\mathcal{S}} P_{2m}(\tau,\partial_t,s\partial_x)\,\mathcal{D}_s(v)= \quad \langle P_{2m}\rangle\, v$	$\sum_{s\in\mathcal{S}} P_{2m+1}(\tau,\partial_t,s\partial_x)\,\mathcal{D}_s(v)= \quad 0$
$\sum_{s\in\mathcal{S}} P_{2m}(\tau,\partial_t,s\partial_x)\,\mathcal{A}_s(v)= \quad 0$	$\sum_{s\in\mathcal{S}} P_{2m+1}(\tau,\partial_t,s\partial_x)\,\mathcal{A}_s(v)= a\theta\langle sP_{2m+1}\rangle v$
$\sum_{s\in\mathcal{S}} P_{2m}(\tau,\partial_t,s\partial_x)\,\mathcal{R}_s(v)=-\tau c\langle P_{2m}\rangle\, v$	$\sum_{s\in\mathcal{S}} P_{2m+1}(\tau,\partial_t,s\partial_x)\,\mathcal{R}_s(v)= \quad 0$

Table 4.3: Generalized moment relations of the equilibrium.

Evolution equations. The question about the evolution equations for the $u^{(n)}$'s finds a simple answer by computing the zeroth moment of (4.12), i.e. summing it over $s \in \mathcal{S}$. For $n = 0, 1$ we can directly sum (4.13), (4.14). Taking the *structure relations* or the *moments* of the *equilibrium* (cf. table 4.1) into account, we discover the tautologies $u^{(0)} = u^{(0)}$ and $u^{(1)} = u^{(1)}$.
Let us now sum equation (4.12) for $n = 2$ and replace $\mathsf{f}^{(0)}, \mathsf{f}^{(1)}$ by (4.13), (4.14):

$$\begin{aligned}\sum_{s\in\mathcal{S}} \mathsf{f}_s^{(2)} \;&=\; \sum_{s\in\mathcal{S}} \Big[\mathcal{E}_s^{(2)} \;-\; \tau\left[T_1(\partial_t,s\partial_x)\,\mathsf{f}_s^{(1)} \;+\; T_2(\partial_t,s\partial_x)\,\mathsf{f}_s^{(0)}\right] \;+\; \tau\, g\, \mathsf{w}_s\Big]\\ u^{(2)} \;&=\; u^{(2)} \;-\; \tau c u^{(0)} \;-\; \tau\sum_{s\in\mathcal{S}} \left[T_1(\partial_t,s\partial_x)\,\mathsf{f}^{(1)} \;+\; T_2(\partial_t,s\partial_x)\,\mathsf{f}^{(0)}\right] \;+\; \tau\, g\end{aligned}$$

Since $u^{(2)}$ cancels out, we find a *diffusion-advection-reaction* equation for $u^{(0)}$, where the given function g takes the role of an external source.

$$\sum_{s\in\mathcal{S}} \Big[(s\partial_x)\big(u^{(1)}\mathsf{w}_s + \theta\, a u^{(0)}\, s\mathsf{w}_s - \tau\,\partial_x u^{(0)} s\,\mathsf{w}_s\big) + (\partial_t + \tfrac{1}{2}s^2\partial_x^2)\big(u^{(0)}\mathsf{w}_s\big)\Big] = g - cu^{(0)}$$

$$\boxed{\partial_t u^{(0)} \;+\; a\partial_x u^{(0)} \;-\; \tfrac{1}{\theta}(\tau - \tfrac{1}{2})\partial_x^2 u^{(0)} \;+\; cu^{(0)} \;=\; g} \tag{4.18}$$

Thus, $u^{(0)}$ is well determined by (4.18) together with initial conditions and the periodic boundary conditions. Exactly as in the previous chapter, this equation for the *zeroth* order of u emanates from the *second* order of equation (4.11). We refer to this phenomenon as *order-shift*, which is typical for *singularly perturbed* systems.

Remark 4.2. *The evolution equation for $u^{(k)}$, $k \in \mathbb{N}_0$, is derived by summing equation (4.12) over $s \in \mathcal{S}$ setting $n = k + 2$ (**order-shift**). This reproduces the* diffusion-advection-reaction *equation in all orders, i.e. the left hand side corresponds to equation (4.18). For $k = 0$ equation (4.18) is obtained, while for $k = 1$ the equation is homogeneous, see (4.21); higher orders are driven by source terms depending on lower orders. Odd and even orders are strictly separated: If k is odd resp. even, then $u^{(k)}$ depends via the source terms only on preceding odd resp. even orders (**odd-even decoupling**). The external source-term g occurs only in even orders; see equations (4.19), (4.20).*

Proof: Let us sum equation (4.12) over $s \in \mathcal{S}$, for even $k \in \mathbb{N}_0$, i.e. $n = 2\alpha + 2$ with $\alpha \geq 0$. The sum in (4.12) is split into two partial sums, such that the running index attains either only even or only odd values.

$$\sum_{s\in\mathcal{S}} \mathsf{f}_s^{(2\alpha+2)} = \sum_{s\in\mathcal{S}} \Big[\sum_{\substack{\gamma\geq 0\\ \gamma \text{ even}}}^{2\alpha+2} P_{2\alpha+2-\gamma}\, \mathcal{E}_s^{(\gamma)} + \sum_{\substack{\gamma\geq 0\\ \gamma \text{ odd}}}^{2\alpha+2} P_{2\alpha+2-\gamma}\, \mathcal{E}_s^{(\gamma)} + \tau\, P_{2\alpha}\, g\, \mathsf{w}_s \Big]$$

$$u^{(2\alpha+2)} = \sum_{s\in\mathcal{S}} \Big[\sum_{\beta=0}^{\alpha+1} P_{2(\alpha-\beta+1)}\, \mathcal{E}_s^{(2\beta)} + \sum_{\beta=0}^{\alpha} P_{2(\alpha-\beta)+1}\, \mathcal{E}_s^{(2\beta+1)} + \tau\, P_{2\alpha}\, g\, \mathsf{w}_s \Big]$$

Decomposing the equilibrium into its three constituents, we observe, that only the underbraced addends survive according to table 4.3.

$$\text{even:} \quad \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta+1)}\, \mathcal{E}_s^{(2\beta)} = \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta+1)} \Big[\underbrace{\mathcal{D}_s^{(2\beta)}} + \mathcal{A}_s^{(2\beta-1)} + \underbrace{\mathcal{R}_s^{(2\beta-2)}} \Big] = \langle P_{2(\alpha-\beta+1)} \rangle \big(u^{(2\beta)} - \tau c u^{(2\beta-2)} \big)$$

$$\text{odd:} \quad \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta)+1}\, \mathcal{E}_s^{(2\beta+1)} = \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta)+1} \Big[\mathcal{D}_s^{(2\beta+1)} + \underbrace{\mathcal{A}_s^{(2\beta)}} + \mathcal{R}_s^{(2\beta-1)} \Big] = a\theta \langle s P_{2(\alpha-\beta)+1} \rangle\, u^{(2\beta)}$$

Therefore if n is even, only even orders of u_h occur in the zeroth moment equation.

$$u^{(2\alpha+2)} = \sum_{\beta=0}^{\alpha+1} \langle P_{2(\alpha-\beta+1)} \rangle \big(u^{(2\beta)} - \tau c u^{(2\beta-2)} \big) + a\theta \sum_{\beta=0}^{\alpha} \langle s P_{2(\alpha-\beta)+1} \rangle u^{(2\beta)} + \tau \langle P_{2\alpha} \rangle g$$

Writing the terms containing either $u^{(2\alpha+2)}$ or $u^{(2\alpha)}$ outside the sums yields:

$$\begin{aligned} &= \langle P_0 \rangle\, u^{(2\alpha+2)} + \langle P_2 \rangle\, u^{(2\alpha)} - \tau c \langle P_0 \rangle\, u^{(2\alpha)} + a\theta \langle s P_1 \rangle\, u^{(2\alpha)} + \tau \langle P_{2\alpha} \rangle\, g \\ &\quad + \sum_{\beta=0}^{\alpha-1} \Big[\langle P_{2(\alpha-\beta+1)} \rangle + a\theta \langle s P_{2(\alpha-\beta)+1} \rangle \Big] u^{(2\beta)} - \tau c \sum_{\beta=0}^{\alpha} \langle P_{2(\alpha-\beta+1)} \rangle\, u^{(2\beta-2)} \end{aligned}$$

It should be noticed, that the first addend of the second sum ($\beta = 0$) does not contribute, since $u^{(-2)} \equiv 0$. Therefore the sum starts actually with $\beta = 1$. By replacing $\beta \to \beta + 1$, the second sum can be merged with the first one. Moreover, resorting to (4.17), results in:

$$\begin{aligned} &= u^{(2\alpha+2)} + \tau \big[\tfrac{1}{\theta} (\tau - \tfrac{1}{2}) \partial_x^2 - \partial_t \big] u^{(2\alpha)} - \tau\, c\, u^{(2\alpha)} - \tau\, a\, \partial_x u^{(2\alpha)} + \tau\, \langle P_{2\alpha} \rangle\, g \\ &\quad + \sum_{\beta=0}^{\alpha-1} \Big[\langle P_{2(\alpha-\beta+1)} \rangle - \tau c \langle P_{2(\alpha-\beta)} \rangle + a\theta \langle s P_{2(\alpha-\beta)+1} \rangle \Big] u^{(2\beta)} \end{aligned}$$

Subtracting $u^{(2\alpha+2)}$ on both sides, gathering all terms containing $u^{(2\alpha)}$ on the left hand side and dividing by τ gives finally the diffusion-advection-reaction equation for $u^{(2\alpha)}$, which is driven by derivatives of the *lower even* orders of u_h and the external source g.

$$\boxed{\partial_t u^{(2\alpha)} + a \partial_x u^{(2\alpha)} - \tfrac{1}{\theta} (\tau - \tfrac{1}{2}) \partial_x^2 u^{(2\alpha)} + c\, u^{(2\alpha)} = \mathrm{EXT}^{(2\alpha)} + \mathrm{LOW}^{(2\alpha)}} \tag{4.19}$$

$$\begin{aligned}
\mathrm{LOW}^{(2\alpha)} &:= \sum_{\beta=0}^{\alpha-1} \Big[\omega\langle P_{2(\alpha-\beta+1)}\rangle \;+\; \omega a\theta\langle sP_{2(\alpha-\beta)+1}\rangle \;-\; c\langle P_{2(\alpha-\beta)}\rangle\Big]\, u^{(2\beta)} \\
\mathrm{EXT}^{(2\alpha)} &:= \langle P_{2\alpha}\rangle\, g
\end{aligned}$$

Thus we find the assertion corroborated for even $k \in \mathbb{N}_0$.

Repeating the same procedure by taking the zeroth moment of equation (4.12) for odd $k \in \mathbb{N}_0$, i.e. $n = 2\alpha + 3$ $\alpha \geq 0$, yields:

$$\begin{aligned}
\sum_{s\in\mathcal{S}} \mathsf{f}_s^{(2\alpha+3)} &= \sum_{s\in\mathcal{S}} \Big[\sum_{\substack{\gamma\geq 0\\ \gamma \text{ odd}}}^{2\alpha+3} P_{2\alpha+1-\gamma}\, \mathcal{E}_s^{(\gamma)} \;+\; \sum_{\substack{\gamma\geq 0\\ \gamma \text{ even}}}^{2\alpha+3} P_{2\alpha+1-\gamma}\, \mathcal{E}_s^{(\gamma)} \;+\; \tau\, P_{2\alpha+1}\, g\, \mathsf{w}_s\Big] \\
u^{(2\alpha+3)} &= \sum_{s\in\mathcal{S}} \Big[\sum_{\beta=0}^{\alpha+1} P_{2(\alpha-\beta+1)}\, \mathcal{E}_s^{(2\beta+1)} \;+\; \sum_{\beta=0}^{\alpha+1} P_{2(\alpha-\beta)+3}\, \mathcal{E}_s^{(2\beta)} \;+\; \tau\, P_{2\alpha+1}\, g\, \mathsf{w}_s\Big]
\end{aligned}$$

Since $\sum_{s\in\mathcal{S}} P_{2\alpha+1}\mathsf{w}_s = \langle P_{2\alpha+1}\rangle = 0$ according to table 4.3, the terms, involving the external source g, disappear as claimed. The equilibrium is again decomposed into its three parts – only the underbraced terms survive.

$$\begin{aligned}
\text{odd:}\quad \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta+1)}\, \mathcal{E}_s^{(2\beta+1)} &= \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta+1)}\, \Big[\underbrace{\mathcal{D}_s^{(2\beta+1)}} + \mathcal{A}_s^{(2\beta)} \;+\; \underbrace{\mathcal{R}_s^{(2\beta-1)}}\Big] \\
&= \langle P_{2(\alpha-\beta+2)}\rangle\, \big(u^{(2\beta+1)} - c\tau u^{(2\beta-1)}\big) \\
\text{even:}\quad \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta)+3}\, \mathcal{E}_s^{(2\beta)} &= \sum_{s\in\mathcal{S}} P_{2(\alpha-\beta)+3}\, \Big[\mathcal{D}_s^{(2\beta)} \;+\; \underbrace{\mathcal{A}_s^{(2\beta-1)}} + \mathcal{R}_s^{(2\beta-2)}\Big] \\
&= a\theta\langle sP_{2(\alpha-\beta)+3}\rangle\, u^{(2\beta-1)}
\end{aligned}$$

As asserted, the reassembled equation involves only *odd* orders of u_h.

$$u^{(2\alpha+3)} = \sum_{\beta=0}^{\alpha+1}\langle P_{2(\alpha-\beta+2)}\rangle\, \big(u^{(2\beta+1)} - c\tau u^{(2\beta-1)}\big) \;+\; a\sum_{\beta=0}^{\alpha+1}\langle sP_{2(\alpha-\beta)+3}\rangle\, u^{(2\beta-1)}$$

Writing the terms containing $u^{(2\alpha+3)}$ or $u^{(2\alpha+1)}$ outside the sums yields:

$$\begin{aligned}
&= \langle P_0\rangle\, u^{(2\alpha+3)} \;+\; \langle P_2\rangle\, u^{(2\alpha+1)} \;-\; \tau c\langle P_2\rangle\, u^{(2\alpha+1)} \;+\; a\theta\langle sP_1\rangle\, u^{(2\alpha+1)} \\
&\quad + \sum_{\beta=0}^{\alpha-1}\langle P_{2(\alpha-\beta+1)}\rangle u^{(2\beta+1)} - \sum_{\beta=0}^{\alpha}\Big[c\tau\langle P_{2(\alpha-\beta+1)}\rangle - a\theta\langle sP_{2(\alpha-\beta)+3}\rangle\Big] u^{(2\beta-1)}
\end{aligned}$$

Note, that there is no contribution for $\beta = 0$ in the second sum, since $u^{(-1)} \equiv 0$. Therefore the sum begins actually with $\beta = 1$. By replacing $\beta \to \beta + 1$, the second sum can be merged with the first one, since the index β runs through the same range. Furthermore, resolving $\langle P_0\rangle$, $\langle P_2\rangle$ and $\langle sP_1\rangle$ by means of (4.17), leads to:

$$\begin{aligned}
&= u^{(2\alpha+3)} \;+\; \tau\big[\tfrac{1}{\theta}(\tau - \tfrac{1}{2})\partial_x^2 - \partial_t\big] u^{(2\alpha+1)} \;-\; \tau c u^{(2\alpha+1)} \;-\; \tau a\partial_x u^{(2\alpha+1)} \\
&\quad + \sum_{\beta=0}^{\alpha-1}\Big[\omega\langle P_{2(\alpha-\beta+1)}\rangle \;-\; c\langle P_{2(\alpha-\beta)}\rangle \;+\; \omega a\theta\langle sP_{2(\alpha-\beta)+1}\rangle\Big]\, u^{(2\beta-1)}
\end{aligned}$$

Subtracting $u^{(2\alpha+3)}$ on both sides, gathering all terms containing $u^{(2\alpha+1)}$ on the left hand side and dividing by τ gives finally the diffusion-advection-reaction equation for $u^{(2\alpha+1)}$, which is driven by derivatives of the *lower odd* orders of u_h.

$$\boxed{\partial_t u^{(2\alpha+1)} + a\partial_x u^{(2\alpha+1)} - \tfrac{1}{\theta}(\tau - \tfrac{1}{2})\partial_x^2 u^{(2\alpha+1)} + cu^{(2\alpha+1)} = \mathrm{LOW}^{(2\alpha+1)}} \tag{4.20}$$

$$\mathrm{LOW}^{(2\alpha+1)} := \sum_{\beta=0}^{\alpha-1} \Big[\omega\langle P_{2(\alpha-\beta+1)}\rangle \; + \; \omega a\theta\langle sP_{2(\alpha-\beta)+1}\rangle \; - \; c\langle P_{2(\alpha-\beta)}\rangle\Big] \, u^{(2\beta-1)}$$

This shows, that the assertion is also true for odd $k \in \mathbb{N}_0$. Especially, we obtain for $\alpha = 0$ the *homogeneous* diffusion-advection-reaction equation for $u^{(1)}$

$$\partial_t u^{(1)} \; + \; a\,\partial_x u^{(1)} \; - \; \tfrac{1}{\theta}(\tau - \tfrac{1}{2})\partial_x^2 u^{(1)} \; + \; c\,u^{(1)} \; = \; 0\,. \tag{4.21}$$

Since there is nothing to sum on the right hand side, $\mathrm{LOW}^{(1)} \equiv 0$. ∎

Conclusions. In the introduction to this chapter we have claimed, that the lattice-Boltzmann scheme can be used as a numerical solver of equation (4.1). Like all numerical methods, the algorithm does generally not furnish the exact solution, but a result that is more or less slightly perturbed by some (unknown) error. From remark 4.2 it becomes clear where this error originates from. Actually, the lattice-Boltzmann algorithm solves simultaneously two independent hierarchies of diffusion-advection-reaction equations, given by (4.19) and (4.20). Unfortunately, the solutions of these infinitely many equations $u^{(0)}, u^{(1)}, u^{(2)}...$ are not provided separately but in a compound form.

$$\begin{aligned} U_h(n,i) &= u_h(t_n, x_i) \\ &= u^{(0)}(t_n, x_i) + hu^{(1)}(t_n, x_i) + h^2 u^{(2)}(t_n, x_i) + h^3 u^{(3)}(t_n, x_i) + ... \end{aligned}$$

Let us assume, that the lattice-Boltzmann algorithm is initialized so that $u^{(0)}(0,\cdot) = w$ where w is the initial condition specified by the target problem. Then $u^{(0)} \equiv u$ is just the solution of the target problem and we obtain

$$U_h(n,i) = u(t_n, x_i) + \mathrm{O}(h)\ . \tag{4.22}$$

This represents a first justification of assertion (4.6), that the zeroth moment U_h yields an approximate solution of the target problem. The leading error term is given by $hu^{(1)}$. Since the $u^{(k)}$'s are independent of the grid-spacing h, the error should nearly be halved, if the grid size is halved.
However, we can improve the convergence behavior significantly (about one order) by exploiting the odd-even decoupling. If the lattice-Boltzmann algorithm is initialized in such a way, that $u^{(2\alpha+1)}(0,\cdot) = 0$ for all $\alpha \in \mathbb{N}_0$, then these orders vanish completely, provided the external source-term is scaled with h^2 as indicated. This is seen by the following reasoning:
$u^{(1)}$ satisfies a homogeneous equation. So, there are neither source-terms nor boundary conditions (recall that we consider only periodic boundary conditions) that could drive it away from zero. Inductively we infer that all higher odd orders must also vanish, since they are only driven by preceding lower odd orders (odd-even

decoupling) being zero according to the hypothesis. Instead of (4.22), a proper initialization thus leads to

$$\begin{aligned} U_h(n,i) &= u^{(0)}(t_n,x_i) \ + \ h^2 u^{(2)}(t_n,x_i) \ + \ h^4 u^{(4)}(t_n,x_i) \ + \ ... \\ &= u(t_n,x_i) + \mathrm{O}(h^2)\,. \end{aligned} \tag{4.23}$$

Therefore we expect the lattice-Boltzmann algorithm to be a second order scheme. So the error is diminished by a factor of four, if the spatial discretization becomes two times finer.

1st and 2nd Moments. Under the assumption, that the odd orders are switched off, we obtain for the first and second monomial moments:

$$\begin{aligned} \sum_{s\in\mathcal{S}} s\,\mathsf{f}_{h,s} &= h\Big(au^{(0)} - \tfrac{1}{\theta}\tau\partial_x u^{(0)}\Big) + \mathrm{O}(h^3) \\ &= h\Big(au^{(0)} - (\nu + \tfrac{1}{2\theta})\partial_x u^{(0)}\Big) + \mathrm{O}(h^3)\,, \end{aligned}$$

$$\sum_{s\in\mathcal{S}} s^2\,\mathsf{f}_{h,s} \;=\; \tfrac{1}{\theta}\,[u^{(0)} + h^2\,u^{(2)}] \;+\; h^2(\tfrac{1}{\theta}-1)\tau\partial_x \underbrace{\big[au^{(0)} - \nu\partial_x^2 u^{(0)}\big]}_{\text{flux}} \;+ \mathrm{O}(h^4)\,.$$

Remark 4.3. Here we have stressed the *analyzing* aspect of the asymptotic expansion. It has been elucidated how to extract the macroscopic quantities (solution of the target equation and the associated flux) from the mesoscopic quantities by computing suitable moments of the population function.
Nonetheless, we should keep in mind the *synthesizing* aspect being also contained in formula (4.16). It tells, how an approximate population function can be build (approximate solution of the discrete lattice-Boltzmann equation) from a solution of the target problem. This is of particular interest in the following situations:

- *Initialization:* Find a suitable initial population function for a given initial condition of the target problem.

- *Boundary conditions:* At boundaries the lattice-Boltzmann algorithm can not be applied to update all populations, since there are missing neighbor nodes. One ansatz is to construct the missing populations by means of the macroscopic boundary conditions of the target equation.

- *Grid coupling and population transformation:* Similarly to the boundary conditions, mesoscopic *interface conditions* have to be found. How are the population functions on grids of different mesh-size related with each other ?

- *Theoretical studies:* Construct from the solution of the target problem an approximate solution of the discrete lattice-Boltzmann equation to prove convergence.

4.2 Consistency and asymptotic similarity

It has been shown in the preceding section by a formal computation, that the lattice-Boltzmann scheme is related to the diffusion-advection-reaction equation.
Classically, a finite difference scheme is connected to a differential equation by a consistency analysis, where the solution of the differential equation is plugged into the scheme. Under suitable smoothness assumptions, the finite differences are replaced by truncated Taylor expansions. If the solution of the differential equation satisfies the difference scheme up to a remainder that shrinks to zero for the discretization getting finer and finer, the scheme is called *consistent* to the differential equation. Then the output of the scheme is considered to approximate the solution of the differential equation.
In order to prove this approximation in a rigorous way, a stability discussion is necessary. For this the scheme is associated with a linear operator acting on a space, whose dimension is determined by the chosen grid. Stability means, that the family of operators resulting from global grid refinements is bounded independently of the mesh parametrized by the grid spacing. By virtue of the *Lax-Richtmyer equivalence theorem* convergence of a linear scheme is guaranteed as soon as it is consistent and stable.

However, this standard approach – as just outlined – is not directly applicable to the analysis of lattice-Boltzmann schemes, as it involves more variables (unknowns) with the tuple-valued population function than the approximated target equation. So the consistency analysis cannot be performed in the manner described above, since it is not possible to insert the solution of the diffusion-advection-reaction equation into the lattice-Boltzmann algorithm. Nevertheless, we may follow the guideline of the classical analyzing procedure, if we generalize the notion of consistency.

The lattice-Boltzmann algorithm in operator form and further notation. In the sequel we consider the lattice-Boltzmann algorithm as described on page 160 with a terminal time $0 < T < \infty$. We recall that N_h denotes the number of spatial nodes for $h \in \mathcal{H}_L$, where $\mathcal{H}_L$ is the set of admissible grid spacings.
Basically, the lattice-Boltzmann algorithm assigns to a given function corresponding to the state at time-step $n-1$, a new function corresponding to the state at time-step n. So, we can describe the algorithm on the grid $\mathcal{G}_h$ by an *evolution operator*,

$$\mathfrak{B}_h(n) : \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R}) \to \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R}) \qquad n \in \mathbb{T}_h \setminus \{0\}$$

where the argument n indicates, that $\mathfrak{B}_h$ might explicitly depend on the iteration, if for instance the external source depends on time. Since the equilibrium depends linearly on the zeroth moment U_h, the evolution operator is linear as long as there are no external sources. In general, however, the evolution operator is an *affine-linear mapping.* Equivalently, the lattice-Boltzmann equation (4.5) can now be written in operator form,

$$\mathsf{F}_h(n+1, \cdot) \;=\; \mathfrak{B}_h(n+1)\mathsf{F}_h(n, \cdot) \;=\; \mathfrak{E}_h \mathsf{F}_h(n, \cdot) \;+\; \mathsf{G}_h(n, \cdot)$$

where the linear part is given by the *evolution matrix*

$$\mathfrak{E}_h : \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R}) \to \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R}) \; .$$

Supposing that the quantities τ, a, c and $\omega = \tau^{-1}$ are constant, $\mathfrak{E}_h$ does not depend on the iteration. Let $\mathfrak{E}_{h,s}\big(\mathsf{F}_h(n,\cdot), i\big)$ denote the evaluation of $\mathfrak{E}_h \mathsf{F}_h(n,\cdot) \in \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z})$ for $s \in \mathcal{S}$ and $i \in \mathbb{Z}/N_h\mathbb{Z}$. Then the evolution operator is defined by

$$\mathfrak{E}_{h,s}\big(\mathsf{F}_h(n,\cdot), i\big) \ := \ (1-\omega)\mathsf{F}_{h,s}(n, i-s) \ + \ \mathcal{E}_{h,s}\Big(\sum_{\tilde{s}\in\mathcal{S}} \mathsf{F}_{h,\tilde{s}}(n, i-s)\Big) \ .$$

The additive part, containing the external source g, is given by the grid function $\mathsf{G}_h : \mathbb{T}_h \to \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R})$.

$$\mathsf{G}_{h,s}(n,i) \ := \ h^2\, g(t_n, x_i)\, \mathsf{w}_s$$

The *restriction operator*[8]

$$\mathfrak{R}_h : \mathcal{F}([0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R}) \to \mathcal{F}(\mathbb{T}_h \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R})$$

confines every function $f \in \mathcal{F}([0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ to the time-space grid $\mathcal{G}_h$. If $\mathfrak{R}_h(f,n,i)$ denotes the evaluation of $\mathfrak{R}_h f$ at the time-step $n \in \mathbb{T}_h$ and at the mesh-node $i \in \mathbb{Z}/N_h\mathbb{Z}$, then $\mathfrak{R}_h$ is defined by

$$\mathfrak{R}_h(f,n,i) := f(t_n, x_i) \qquad \text{with} \qquad (t_n, x_i) \in \mathcal{G}_h \ .$$

We will mostly use the short notation $\hat{f}(n,i) := \mathfrak{R}_h(f,n,i)$; it should then become clear from the context, to what grid the restriction refers.
Let us finally equip the function space $\mathcal{F}(\mathcal{S} \times \mathbb{Z}/N\mathbb{Z}, \mathbb{R})$ with some norms[9]. Since

$$\dim \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N\mathbb{Z}, \mathbb{R}) \ = \ \#\mathcal{S} \ \cdot \ \#(\mathbb{Z}/N\mathbb{Z}) \ = \ 3N \quad ,$$

the space is isomorphic to $\mathbb{R}^{3N}$. Therefore we obtain canonically the norms:

$$\|\mathsf{f}\|_\infty := \max_{s\in\mathcal{S}} \max_{i\in\mathbb{Z}/N\mathbb{Z}} |\mathsf{f}_s(i)| \qquad \|\mathsf{f}\|_p := \Big(\sum_{s\in\mathcal{S}} \sum_{i\in\mathbb{Z}/N\mathbb{Z}} |\mathsf{f}_s(i)|^p\Big)^{\frac{1}{p}} \quad p \geq 1$$

As we interpret the functions of $\mathcal{F}(\mathcal{S} \times \mathbb{Z}/N\mathbb{Z}, \mathbb{R})$ not as elements of $\mathbb{R}^{3N}$ but as restrictions of functions in $\mathcal{F}(\mathcal{S} \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$, we prefer[10] the L^p-norms.

$$\|\mathsf{f}\|_{L^p} := \sqrt[p]{\tfrac{L}{N}}\|\mathsf{f}\|_p \ = \ \Big(\sum_{s\in\mathcal{S}} \sum_{i\in\mathbb{Z}/N\mathbb{Z}} |\mathsf{f}_s(i)|^p \tfrac{L}{N}\Big)^{\frac{1}{p}}$$

In order to adopt a uniform notation, we write $\|\cdot\|_{L^\infty}$ instead of $\|\cdot\|_\infty$. If not indicated otherwise, we suppose $1 \leq p \leq \infty$. It should be remarked, that $L^p(\mathcal{S} \times \mathbb{Z}/N\mathbb{Z}, \mathbb{R})$ is *isometrically embedded* into $L^p(\mathcal{S} \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ by identifying $i \in \mathbb{Z}/N\mathbb{Z}$ with $x_i := i\frac{L}{N} \in \mathbb{R}/L\mathbb{R}$. The functions which are thus defined only on the nodes x_i

[8]By abuse of notation, we utilize the same symbol also for the restriction operator on $\mathcal{F}(\mathcal{S} \times [0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ that is defined analogously.

[9]It should be clear, that everything what follows has a straight-forward analogon in $\mathcal{F}(\mathbb{Z}/N\mathbb{Z}, \mathbb{R})$. However, we do not state the analogous definitions, to avoid repetitions.

[10]Note that for the constant one-function $\mathbf{1}_N \in \mathcal{F}(\mathbb{Z}/N\mathbb{Z}, \mathbb{R})$ we have $\|\mathbf{1}_N\|_p = \sqrt[p]{N}$, while we have $\|\mathbf{1}_N\|_{L^p} = \sqrt[p]{L}$ being independent of N. In that sense, the L^p norms are better to be compared with each other, if N varies.

are extended by the left neighbor node interpolation (or alternatively by the nearest neighbor node interpolation) to piecewise constant functions defined on the whole domain $\mathbb{R}/L\mathbb{R}$.

$$L^p(\mathcal{S} \times \mathbb{Z}/N\mathbb{Z}, \mathbb{R}) \ni \mathsf{H} \mapsto \bar{\mathsf{H}} \in L^p(\mathcal{S} \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$$

$$\text{Left neighbor node interpolation:} \quad \bar{\mathsf{H}}_s(x) := \begin{cases} \mathsf{H}_s(x_0) & x_0 \leq x < x_1 \\ \vdots & \vdots \leq x < \vdots \\ \mathsf{H}_s(x_{N-1}) & x_{N-1} \leq x < L \end{cases}$$

Note, that the L^p-norm of H is equal to the standard L^p-norm of the tuple-valued function $\bar{\mathsf{H}}$ on $\mathbb{R}/L\mathbb{R}$, i.e. $\|\mathsf{H}\|_{L^p} = \|\bar{\mathsf{H}}\|_{L^p} := \big(\sum_{s\in\mathcal{S}} \int_{\mathbb{R}/L\mathbb{R}} |\bar{\mathsf{H}}_s(x)|^p \, dx \big)^{\frac{1}{p}}$.

Definition 4.1. ***(Consistency)*** *Let $\mathsf{f}_\epsilon \in \mathcal{F}(\mathcal{S} \times [0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ depend on the real parameter ϵ. The function f_ϵ is* consistent *to the discrete lattice-Boltzmann equation with a consistency order $r \in \mathbb{R}_0^+$ referring to the norm $\|\cdot\|_{L^p}$, if there exists a constant C_p independent of h and n, such that for all $h \in \mathcal{H}_L$ and $n \in \mathbb{T}_h \setminus \{0\}$ the* iteration error *satisfies the estimate*

$$\left\| \, \hat{\mathsf{f}}_h(n,\cdot) - \mathfrak{B}_h(n)\hat{\mathsf{f}}_h(n-1,\cdot) \, \right\|_{L^p} \;\leq\; C_p \, h^r \, . \tag{4.24}$$

The reason why we are interested in a consistent population function f_ϵ is to compare it for $\epsilon = h$ with the output F_h of the lattice-Boltzmann algorithm on the grid $\mathcal{G}_h$. Since the solution of the discrete lattice-Boltzmann equation is uniquely defined by the initial values and since $\hat{\mathsf{f}}_h$ solves the discrete lattice-Boltzmann equation approximately, we expect, that F_h and $\hat{\mathsf{f}}_h$ remain close together if they are (almost) equally initialized. Of course, the deviation might increase in time mainly because $\hat{\mathsf{f}}_h$ does not satisfy the discrete lattice-Boltzmann equation exactly. Let us suppose, that the deviation per iteration is of order $\mathrm{O}(h^r)$. In order to reach a fixed time $t \in (0,T]$, $N_t = N_t(h) := [tN^2] = [th^{-2}]$ iterations have to be performed. In the worst case, the deviation per iteration cumulates; thus after $[th^{-2}]$ iterations the mutual deviation of F_h and $\hat{\mathsf{f}}_h$ should be of order $[th^{-2}]\mathrm{O}(h^r) = \mathrm{O}(h^{r-2})$, under the assumption that both quantities agree at the beginning.
Before we go on, let us mention the following lemma, that turns out to be quite helpful in the proof of the next theorem.

Lemma 4.3. *For $k \in \mathbb{N}$ consider the affine-linear mappings $A_k : V \to V$ with V denoting some vector space. Assume $A_k x = L_k x + b_k$ with the linear mapping $L_k \in \mathrm{Hom}(V)$ and some vector $b_k \in V$. Then for $x, y \in V$*

$$\prod_{k=1}^{n} A_k \, x \;-\; \prod_{k=1}^{n} A_k \, y \;=\; \prod_{k=1}^{n} L_k \, (x-y) \tag{4.25}$$

where the product sign denotes the composition, *for instance* $\prod_{k=1}^{3} A_k \, x = (A_3 \circ A_2 \circ A_1)(x) = A_3\Big(A_2\big(A_1(x)\big)\Big)$.

Proof: The proof is done by induction:

$n = 1$: $\quad A_1 x - A_1 y = (L_1 x + b_1) - (L_1 y + b_1) = L_1 x - L_1 y = L_1(x - y)$

$n \Rightarrow n+1$: Suppose (4.25) holds for some $n \in \mathbb{N}$.

$$\begin{aligned}
\prod_{k=1}^{n+1} A_k\, x \;-\; \prod_{k=1}^{n+1} A_k\, y \;&=\; A_{n+1} \prod_{k=1}^{n} A_k\, x \;-\; A_{n+1} \prod_{k=1}^{n} A_k\, y \\
&=\; L_{n+1} \Big(\prod_{k=1}^{n} A_k\, x \;-\; \prod_{k=1}^{n} A_k\, y \Big) \\
&=\; L_{n+1} \prod_{k=1}^{n} L_k\, (x - y) \;=\; \prod_{k=1}^{n+1} L_k\, (x - y)
\end{aligned}$$

■

The following theorem is related to the *Lax-Richtmyer equivalence theorem*, which plays a central role in the theory of linear finite difference schemes. Actually, this theorem comes along in various versions depending on the difference scheme to be considered.
Generally, it states, that a linear finite difference scheme is convergent if and only if it is consistent and stable, i.e. consistency and stability are sufficient and necessary conditions for convergence. The power of the equivalence theorem is based on the fact, that it permits to divide the convergence analysis into two parts that can be done independently: a consistency analysis that is often quite evident by a Taylor expansion and a stability analysis.

In the following theorem we provide the central estimate that connects – under the hypothesis of stability – *consistency* with *asymptotic similarity*, thus paving the way towards the *convergence* of the considered D1P3 lattice-Boltzmann scheme.

Theorem 4.1. *May F_h for $h \in \mathcal{H}_L$ be a solution of the discrete lattice-Boltzmann equation (4.5) initialized by $F_h(0,\cdot)$ and may f_ϵ be consistent of order r to the discrete lattice-Boltzmann equation with respect to the L^p-norm. Then the difference between F_h and $\hat{\mathsf{f}}_h$ can be estimated by*

$$\begin{aligned}
&\|\mathsf{F}_h(N_t,\cdot) - \hat{\mathsf{f}}_h(N_t,\cdot)\|_{L^p} \\
&\qquad \leq \underbrace{\|\mathfrak{E}_h^{N_t}\|_p \cdot \|\mathsf{F}_h(0,\cdot) - \hat{\mathsf{f}}_h(0,\cdot)\|_{L^p}}_{\textit{inherited initial dev.}} \;+\; \underbrace{\max_{0 \leq n \leq N_t} \|\mathfrak{E}_h^{n}\|_p \, C_p\, T\, h^{r-2}}_{\textit{inherent dev.}} \;, \qquad (4.26)
\end{aligned}$$

where $\|\mathfrak{E}_h\|_p$ denotes the operator norm induced by the p-norm on $\mathcal{F}(S \times \mathbb{Z}/N_h\mathbb{Z})$ of the evolution matrix $\mathfrak{E}_h \in \mathcal{L}\big(\mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R}), \mathcal{F}(\mathcal{S} \times \mathbb{Z}/N_h\mathbb{Z}, \mathbb{R})\big)$.

The mutual deviation is obviously split into two parts: The *inherited initial deviation* amplified or attenuated by the repeated application of the evolution operator and the *inherent deviation*, that is due to the iteration error.

Proof: Let us introduce the two grid functions

$$\begin{aligned}
\mathsf{A}_h(\cdot) \;&:=\; \prod_{k=1}^{N_t} \mathfrak{B}_h(k) \mathsf{F}_h(0,\cdot) \;-\; \prod_{k=1}^{N_t} \mathfrak{B}_h(k) \hat{\mathsf{f}}_h(0,\cdot) \\
\mathsf{B}_h(\cdot) \;&:=\; \prod_{k=1}^{N_t} \mathfrak{B}_h(k) \hat{\mathsf{f}}_h(0,\cdot) \;-\; \hat{\mathsf{f}}_h(N_t,\cdot) \qquad .
\end{aligned}$$

Since $\mathsf{F}_h(N_t,\cdot) = \prod_{k=1}^{N_t} \mathfrak{B}_h(k)\mathsf{F}_h(0,\cdot)$, we get by using the triangle inequality:

$$\left\|\mathsf{F}_h(N_t,\cdot) - \hat{\mathsf{f}}_h(N_t,\cdot)\right\| \;\leq\; \|\mathsf{A}_h\| + \|\mathsf{B}_h\|$$

So, we have to estimate A_h and B_h. Application of lemma 4.3 yields:

$$\mathsf{A}_h \;=\; \prod_{k=1}^{N_t} \mathfrak{E}_h \left(\mathsf{F}_h(0,\cdot) - \hat{\mathsf{f}}_h(0,\cdot)\right) \;=\; \mathfrak{E}_h^{N_t} \left(\mathsf{F}_h(0,\cdot) - \hat{\mathsf{f}}_h(0,\cdot)\right)$$

and therefore we obtain: $\quad \|\mathsf{A}_h\|_{L^p} \leq \|\mathfrak{E}_h^{N_t}\|_p \, \|\mathsf{F}_h(0,\cdot) - \hat{\mathsf{f}}_h(0,\cdot)\|_{L^p}$

B_h can be written as a telescopic sum in the following manner.

$$\begin{aligned}\mathsf{B}_h \;&= \sum_{j=1}^{N_t-1} \Big[\prod_{k=j}^{N_t} \mathfrak{B}_h(k)\;\hat{\mathsf{f}}_h(j-1,\cdot) \;-\; \prod_{k=j+1}^{N_t} \mathfrak{B}_h(k)\;\hat{\mathsf{f}}_h(j,\cdot)\Big] \\ &\quad + \mathfrak{B}_h(N_t)\,\hat{\mathsf{f}}_h(N_t-1,\cdot) \;-\; \hat{\mathsf{f}}_h(N_t,\cdot)\end{aligned}$$

Exploiting once more lemma 4.3 gives:

$$\begin{aligned}&= \sum_{j=1}^{N_t-1} \prod_{k=j+1}^{N_t} \mathfrak{E}_h \;\Big[\mathfrak{B}_h(j)\,\hat{\mathsf{f}}_h(j-1,\cdot) \;-\; \hat{\mathsf{f}}_h(j,\cdot)\Big] \\ &\quad + \mathfrak{B}_h(N_t)\,\hat{\mathsf{f}}_h(N_t-1,\cdot) \;-\; \hat{\mathsf{f}}_h(N_t,\cdot) \\ &= \sum_{j=1}^{N_t} \mathfrak{E}_h^{N_t-j} \Big[\mathfrak{B}_h(j)\,\hat{\mathsf{f}}_h(j-1,\cdot) \;-\; \hat{\mathsf{f}}_h(j,\cdot)\Big]\end{aligned}$$

Note, that $\mathfrak{E}_h^0$ stands for the identity. Using now the consistency (4.24) and observing that $N_t \leq tN^2 = th^{-2}$, this results in the following estimate:

$$\begin{aligned}\|\mathsf{B}_h\|_{L^p} \;&\leq\; \sum_{j=1}^{N_t} \|\mathfrak{E}_h^{N_t-j}\|_p \cdot \|\mathfrak{B}_h(j)\,\hat{\mathsf{f}}_h(j-1,\cdot) \;-\; \hat{\mathsf{f}}_h(j,\cdot)\|_{L^p} \\ &\leq\; N_t \max_{1\leq j\leq N_t} \|\mathfrak{E}_h^{N_t-j}\|_p \, C_p\, h^r \;\leq\; \max_{0\leq n\leq N_t} \|\mathfrak{E}_h^n\|_p \, C_p T\, h^{r-2}\end{aligned}$$

Hence, the difference between the lattice-Boltzmann solution F_h and the given function $\hat{\mathsf{f}}_h$ is dominated as asserted. ■

It is desirable to improve estimate (4.26) in the sense that the h-dependence of the right hand side gets more simple. In order to finish the conclusions, that the preceding theorem permits us to draw, we adopt the following **assumption of stability**:

$$\sup_{h\in\mathcal{H}_L} \max_{0\leq n\leq[th^{-2}]} \|\mathfrak{E}_h^n\|_p =: E_p < \infty \tag{4.27}$$

Furthermore let us suppose, that F_h and $\hat{\mathsf{f}}_h$ are *initially consistent* of order ρ (w.r.t. the L^p-norm), meaning that there exists a constant $J_p > 0$ being independent of h with

$$\|\mathsf{F}_h(0,\cdot) - \hat{\mathsf{f}}_h(0,\cdot)\|_{L^p} \;\leq\; J_p\, h^\rho\;. \tag{4.28}$$

Then (4.26) becomes:

$$\|\mathsf{F}_h(N_t,\cdot)-\hat{\mathsf{f}}_h(N_t,\cdot)\|_{L^p} \leq \; E_p\, J_p\, h^\rho \;+\; E_p\, C_p\, T\, h^{r-2} \tag{4.29}$$

As both quantities on the left hand side depend on h, it does not make sense to speak about *convergence* of $\mathsf{F}_h(N_t(h),\cdot)$ towards $\hat{\mathsf{f}}_h(N_t(h),\cdot)$ or conversely for $h\downarrow 0$. Instead, we refer to definition 1.4 where we have introduced the notion of *asymptotic similarity*, characterizing this kind of mutual approximation.

Corollary 4.1. *May F_h for $h\in\mathcal{H}_L$ be a solution of the discrete lattice-Boltzmann equation (4.5) and may f_ϵ be consistent of order r to the discrete lattice-Boltzmann equation w.r.t. the norm $\|\cdot\|_{L^p}$. The stability assumption may be fulfilled and $\mathsf{F}_h,\hat{\mathsf{f}}_h$ are supposed to be consistently initialized up to order $r-2$. Then $\mathsf{F}_h(N_t,\cdot),\ \hat{\mathsf{f}}_h(N_t,\cdot)\in L^p(\mathcal{S}\times\mathbb{Z}/N_h\mathbb{Z},\mathbb{R}) \overset{isom.}{\hookrightarrow} L^p(\mathcal{S}\times\mathbb{R}/L\mathbb{R},\mathbb{R})$ are* asymptotically similar *as sequences in $L^p(\mathcal{S}\times\mathbb{R}/L\mathbb{R},\mathbb{R})$ for $h\downarrow 0$ with residue of order $r-2$, as we have:*

$$\|\mathsf{F}_h(N_t,\cdot)-\hat{\mathsf{f}}_h(N_t,\cdot)\|_{L^p} \leq \; (E_p\, J_p \;+\; E_p\, C_p\, T)\;\, h^{r-2} \tag{4.30}$$

Proof: The assertion is a direct consequence of the isometry and equation (4.30), that results trivially from (4.29). ■

Eventually, we throw a glance at the zeroth moment, which is actually the quantity of interest. Let $1\leq p<\infty$.

$$\|U_h(N_t,\cdot)-\hat{u}_h(N_t,\cdot)\|_{L^p} = \sqrt[p]{\tfrac{L}{N}}\,\Big(\sum_{i\in\mathbb{Z}/N_h\mathbb{Z}}\Big|\sum_{s\in\mathcal{S}}\mathsf{F}_{h,s}(N_t,i)-\hat{\mathsf{f}}_{h,s}(N_t,i)\Big|^p\Big)^{\frac{1}{p}}$$

In order to get the summation over $s\in\mathcal{S}$ out of the modulus, we remember that $x\mapsto|x|^p$ is a convex function[11] for $p\geq 1$. Thus we obtain:

$$\begin{aligned}
&\leq \sqrt[p]{\tfrac{L}{N}}\;3^{\frac{p-1}{p}}\Big(\sum_{i\in\mathbb{Z}/N_h\mathbb{Z}}\sum_{s\in\mathcal{S}}\Big|\mathsf{F}_{h,s}(N_t,i)-\hat{\mathsf{f}}_{h,s}(N_t,i)\Big|^p\Big)^{\frac{1}{p}}\\
&= 3^{\frac{p-1}{p}}\;\|F_h(N_t,\cdot)-\hat{\mathsf{f}}(N_t,\cdot)\|_{L^p} \;\leq\; 3^{\frac{p-1}{p}}\;\big(E_p\,J_p \;+\; E_p\,C_p\,T\big)\;h^{r-2}
\end{aligned}$$

This estimate is also true[12] for $p=\infty$, since $\lim_{p\to\infty}3^{\frac{p-1}{p}}=3$. Therefore we see, that the asymptotic similarity is transmitted from the population functions $\mathsf{F}_h,\hat{\mathsf{f}}_h$ to the associated zero moments $U_h,\hat{u}_h$.

[11] If the function f is convex on the interval $I\subset\mathbb{R}$, then the following inequality holds true for n points $x_1,...,x_n\in I$ and positive numbers $\lambda_1,...,\lambda_n$ with $\lambda_1+...+\lambda_n=1$:

$$f(\lambda_1x_1+...+\lambda_nx_n)\leq\lambda_1f(x_1)+...+\lambda_nf(x_n)$$

Therefore we obtain in our special case:

$$\begin{aligned}
|\sum_{s\in\mathcal{S}}x_s|^p &= 3^p\,\tfrac{1}{3^p}|x_0+x_\oplus+x_\ominus|^p = 3^p\,|\tfrac13x_0+\tfrac13x_\oplus+\tfrac13x_\ominus|^p\\
&\leq 3^p\,\big(\tfrac13|x_0|^p+\tfrac13|x_\oplus|^p+\tfrac13|x_\ominus|^p\big) \;=\; 3^{p-1}\sum_{s\in\mathcal{S}}|x_s|^p
\end{aligned}$$

[12] For $p=\infty$ we have:

$$\begin{aligned}
\|U_h(N_t,\cdot)-\hat{u}_h(N_t,\cdot)\|_{L^\infty} &= \max_{i\in\mathbb{Z}/N_h\mathbb{Z}}\Big|\sum_{s\in\mathcal{S}}\mathsf{F}_{h,s}(N_t,i)-\hat{\mathsf{f}}_{h,s}(N_t,i)\Big|\\
&\leq 3\max_{s\in\mathcal{S}}\max_{i\in\mathbb{Z}/N_h\mathbb{Z}}\Big|\mathsf{F}_{h,s}(N_t,i)-\hat{\mathsf{f}}_{h,s}(N_t,i)\Big| = 3\,\|\mathsf{F}_{h,s}(N_t,\cdot)-\hat{\mathsf{f}}_{h,s}(N_t,\cdot)\|_{L^\infty}
\end{aligned}$$

4.3 Construction of consistent population functions

In this section we show by an explicit construction, that functions being consistent to the lattice-Boltzmann equation of arbitrarily high order do exist if we assume sufficient regularity of the solution to the target problem. In order to avoid the impression, that these consistent functions appear out of hot air we point out that the construction is strongly motivated by the results of the formal analysis. In fact, these approximate population functions are obtained by truncating the asymptotic expansion.
Before we formulate the smoothness requirements and describe the construction process, we state the following preparatory lemma.

Lemma 4.4. *Let $f \in C^{n+1}([0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ and $0 \leq |h| \leq 1$, then:*

$$\Big| f(t+h^2, x+h) \;-\; \sum_{k=0}^{n} h^k \, T_k(\partial_t, \partial_x) f(t,x) \Big| \;\leq\; \|f\|_{C^{n+1}} \, |h|^{n+1}$$

Proof: According to Taylor's theorem, there exist $\theta \in (t, t+h^2)$ and $\xi \in (x, x+h)$, such that:

$$f(t+h^2, x+h) \;=\; \underbrace{\sum_{k=0}^{n} \tfrac{1}{k!} \left[\binom{h^2}{h} \cdot \binom{\partial_t}{\partial_x} \right]^k f(t,x)}_{=:S_n} + \underbrace{\tfrac{1}{(n+1)!} \left[\binom{h^2}{h} \cdot \binom{\partial_t}{\partial_x} \right]^{n+1} f(\theta, \xi)}_{=:R_n}$$

The truncated Taylor series is split into two sums, where the indices β, γ are supposed to be non-negative. The second sum is abbreviated by S_n^*.

$$\begin{aligned} S_n \quad &= \sum_{\substack{2\beta+\gamma \leq n \\ \beta+\gamma \leq n}} \tfrac{1}{\beta!\gamma!} h^{2\beta+\gamma} \partial_t^\beta \partial_x^\gamma f(t,x) \;+ \sum_{\substack{2\beta+\gamma \geq n+1 \\ \beta+\gamma \leq n}} \tfrac{1}{\beta!\gamma!} h^{2\beta} h^{\gamma} \partial_t^\beta \partial_x^\gamma f(t,x) \\ &= \sum_{k=0}^{n} \sum_{2\beta+\gamma=k} \tfrac{1}{\beta!\gamma!} h^{2\beta+\gamma} \partial_t^\beta \partial_x^\gamma f(t,x) \;+\; S_n^* \\ &\overset{(4.10)}{=} \sum_{k=0}^{n} h^k \, T_k(\partial_t, \partial_x) f(t,x) \;+\; S_n^* \end{aligned}$$

If $n = 2k$ respectively $n = 2k+1$, then S_n^* contains $k^2 + k$ respectively $(k+1)^2$ addends. However, we estimate S_n^* simply by dropping the restriction $2\beta+\gamma \geq n+1$ and adding up all derivatives of order $\leq n$.

$$\begin{aligned} |S_n^*| &\leq |h|^{n+1} \sum_{\substack{2\beta+\gamma \geq n+1 \\ \beta+\gamma \leq n}} |\partial_t^\beta \partial_x^\gamma f(t,x)| \;\leq\; |h|^{n+1} \sum_{\beta+\gamma \leq n} |\partial_t^\beta \partial_x^\gamma f(t,x)| \\ &\leq |h|^{n+1} \sum_{\beta+\gamma \leq n} \|\partial_t^\beta \partial_x^\gamma f\|_{C^0} \;=\; |h|^{n+1} \, \|f\|_{C^n} \end{aligned}$$

Estimating R_n yields:

$$\begin{aligned} |R_n| \le\; & |h|^{n+1} \sum_{\beta+\gamma=n+1} |\partial_t^\beta \partial_x^\gamma f(\theta,\xi)| \\ \le\; & |h|^{n+1} \sum_{\beta+\gamma=n+1} \|\partial_t^\beta \partial_x^\gamma f\|_{C^0} \;=\; |h|^{n+1}\, |f|_{C^{n+1}} \end{aligned}$$

Since

$$f(t+h^2, x+h) \;-\; \sum_{k=0}^{n} h^k T_k(\partial_t, \partial_x) f(t,x) \;=\; S_n^* \;+\; R_n$$

the assertion follows by taking the modulus, applying the triangle inequality on the right hand side, using the estimates for R_n and S_n^* and observing, that $\|f\|_{C^{n+1}} = \|f\|_{C^n} \;+\; |f|_{C^{n+1}}$. ■

Existence and Smoothness Hypothesis: Let $\alpha \in \mathbb{N}_0$ be given. Consider the target problem with the initial and source functions

$$w \in C^{2\alpha+4}(\mathbb{R}/L\mathbb{R}, \mathbb{R}) \qquad g \in C^{2\alpha+2}([0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})\ . \tag{4.31}$$

We assume, that there exist functions $u^{(0)}, u^{(2)}, ..., u^{(2\alpha)}$ with

$$u^{(2\beta)} \in C^{2\alpha-2\beta+4}([0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R}) \quad \beta \in \{0,1,...,\alpha\} \tag{4.32}$$

being solutions to the following coupled initial value problems (cf. eqn. (4.19)).

$$u^{(2\beta)}(0,\cdot) \;=\; w\ \delta_{0\beta} \tag{4.33}$$

$$\partial_t u^{(2\beta)} + \big(a\partial_x - \underbrace{\tfrac{1}{\theta}(\tau - \tfrac{1}{2})}_{=\nu} \partial_x^2 + c\big) u^{(2\beta)} \;=\; \mathrm{EXT}^{(2\beta)} + \mathrm{LOW}^{(2\beta)} \tag{4.34}$$

Here $\delta_{0\beta}$ denotes the Kronecker-delta. The two source terms are given by:

$$\begin{aligned} \mathrm{LOW}^{(2\alpha)} \;&:=\; \sum_{\beta=0}^{\alpha-1} \Big[\omega\langle P_{2(\alpha-\beta+1)}\rangle \;+\; \omega a\theta\langle sP_{2(\alpha-\beta)+1}\rangle \;-\; c\langle P_{2(\alpha-\beta)}\rangle\Big]\, u^{(2\beta)} \\ \mathrm{EXT}^{(2\alpha)} \;&:=\; \langle P_{2\alpha}\rangle\ g \end{aligned}$$

We suppose, that the algorithmic parameters θ, τ are related to the diffusivity ν of the target problem according to (4.7). Then $u^{(0)}$ is just the solution of the target problem (4.1). Furthermore let us define

$$u^{(2\alpha+2)} \equiv 0, \qquad u^{(2\beta+1)} \equiv 0 \quad \beta \in \{0,1,...,\alpha+1\}\ .$$

For ease of notation we set $u^{(-1)} \equiv 0$ and $u^{(-2)} \equiv 0$.

Remark 4.4. The smoothness hypothesis is motivated by the pragmatic reason, to obtain the main result (theorem 4.2) in classical sense. Since the polynomials defined by (4.15) mix time and space derivatives, we do not want to distinguish between temporal and spatial regularity. Apart from this, however, we do not assume more regularity than necessary. In particular, the smoothness hypothesis guarantees, that the terms on the right hand side of (4.34) keep their classical meaning. It has to be checked separately, whether the required regularity of the data really implies the assumed regularity of the $u^{(2\beta)}$'s by successively solving the initial value problem (4.33),(4.34). Using the regularizing effect of the diffusive term (second derivative), the assumption may be relaxed; but we do not enter this issue here.

Due to the regularity assumption concerning g and the $u^{(2\beta)}$'s we can construct the following functions $\mathsf{f}^{(n)}$ for $n \in \{0, 1, ..., 2\alpha + 3\}$

$$\mathsf{f}^{(n)} \;:=\; \sum_{k=0}^{n} P_{n-k}(\tau, \partial_t, S\partial_x)\mathcal{E}^{(k)} \;+\; \tau P_{n-2}\, g\, \mathsf{w}\;, \tag{4.35}$$

where $\mathcal{E}^{(k)}$ abbreviates $\mathcal{D}(u^{(k)}) + \mathcal{A}(u^{(k-1)}) + \mathcal{R}(u^{(k-2)})$. It should be noticed, that the degree of the polynomial P_{n-k} is $n-k$. So, the highest (possibly mixed) derivative contained in $P_{n-k}(\tau, \partial_t, S\partial_x)$ is of order $n-k \le 2\alpha+3-k$. Since it is applied to $u^{(k)} \in C^{2\alpha-k+4}$, $u^{(k-1)} \in C^{2\alpha-k+5}$ and $u^{(k-2)} \in C^{2\alpha-k+6}$ the sum is well defined. Likewise, as we have assumed $g \in C^{2\alpha+2}$, the highest derivative acting on g and being of order $2\alpha + 1$, also exists in the classical sense. Actually, we have always presupposed one order of regularity more than necessary for the definition of the $\mathsf{f}^{(n)}$'s. This additional order is required to derive the estimate of lemma 4.6.

The *approximate population function* f_ϵ to be compared with the output of the lattice-Boltzmann algorithm is assembled by the following truncated asymptotic expansion

$$\mathsf{f}_\epsilon \;:=\; \sum_{n=0}^{2\alpha+3} \mathsf{f}^{(n)}\, \epsilon^n\;, \tag{4.36}$$

where the $\mathsf{f}^{(n)}$'s appear as asymptotic orders. By virtue of their construction (cf. 4.16 resp. (4.35)) it is no surprise, that they exhibit the following algebraic property becoming crucial further below.

Lemma 4.5. *$u^{(n)}$ can be recovered from $\mathsf{f}^{(n)}$ ($n = 0, ..., 2\alpha + 3$) by computing its zeroth moment.*

$$\sum_{s\in\mathcal{S}} \mathsf{f}_s^{(n)} = u^{(n)} \qquad \left\{ \begin{array}{lcllr} \sum_{s\in\mathcal{S}} \mathsf{f}_s^{(2\beta)} & = & u^{(2\beta)} & \textit{for} & \beta \in \{0,1,...,\alpha\} \\ \sum_{s\in\mathcal{S}} \mathsf{f}_s^{(2\beta+1)} & = & 0 & \textit{for} & \beta \in \{0,1,...,\alpha+1\} \\ \sum_{s\in\mathcal{S}} \mathsf{f}_s^{(2\alpha+2)} & = & 0 & & \end{array} \right.$$

Proof: For $n = 0, 1$ the assertion can be directly verified using (4.13) and (4.14); for $n \ge 2$ we take recourse to theorem 4.2. From the representation formula (4.35), we infer for even $n = 2\beta + 2$ and $\beta \in \{0, 1, ..., \alpha\}$ by repeating the arithmetic manipulations performed already in the proof of theorem 4.2:

$$\sum_{s\in\mathcal{S}} \mathsf{f}^{(2\beta+2)} = u^{(2\beta+2)} \underbrace{-\big(\partial_t + a\partial_x - \tfrac{1}{\theta}(\tau - \tfrac{1}{2})\partial_x^2 + c\big)u^{(2\beta)} + \mathrm{EXT}^{(2\beta)} + \mathrm{LOW}^{(2\beta)}}_{=0}$$

Since $u^{(2\beta)}$ satisfies (4.34), the underbraced term vanishes and the assertion is obtained. In the same way we get for odd $n = 2\beta + 3$ and $\beta \in \{0, 1, ..., \alpha\}$ the expression[13],

$$\sum_{s\in\mathcal{S}} \mathsf{f}^{(2\beta+3)} = \underbrace{u^{(2\beta+3)} - \big(\partial_t + a\partial_x - \tfrac{1}{\theta}(\tau - \tfrac{1}{2})\partial_x^2 + c\big)u^{(2\beta+1)} + \mathrm{LOW}^{(2\beta+1)}}_{=0}$$

which vanishes since $u^{(1)} = u^{(3)} = ... = u^{(2\alpha+3)} = 0$, thus concluding the proof. ■

Therefore we can write the zeroth moment of f_ϵ

$$u_\epsilon \ := \ \sum_{s\in\mathcal{S}} \mathsf{f}_{\epsilon,s} \ = \ \sum_{\beta=0}^{\alpha} u^{(2\beta)}\ \epsilon^{2\beta}$$

as a truncated asymptotic expansion with the $u^{(2\beta)}$'s as coefficient functions. Finally let us define the following remainder terms,

$$\mathsf{r}^{(n)}_{\epsilon,s}(t,x) \ := \ \mathsf{f}^{(n)}_s(t+\epsilon^2, x+s\epsilon) - \sum_{k=0}^{2\alpha+3-n} \epsilon^k\ T_k(\partial_t, s\partial_x)\ \mathsf{f}^{(n)}_s(t,x) \tag{4.37}$$

satisfying the following estimate.

Lemma 4.6. *There exists a constant $K = K(a, c, \tau, \theta)$ depending only on the parameters of the target problem*[14]*, such that*

$$\sum_{n=0}^{2\alpha+3} \|r^{(n)}_{\epsilon,s}\|_{C^0}\ \epsilon^n \ \le \ K\ \Big(\sum_{\gamma=0}^{\alpha} \|u^{(2\gamma)}\|_{C^{2(\alpha-\gamma)+4}} \ + \ \|g\|_{C^{2\alpha+2}}\Big)\ \epsilon^{2\alpha+4}$$

for all $s \in \mathcal{S}$ and $0 \le \epsilon \le 1$.

Proof: In order to estimate $\mathsf{r}^{(2\beta)}_{\epsilon,s}$ we apply lemma 4.4 to $\mathsf{f}^{(2\beta)}_s$.

$$\begin{aligned} \|\mathsf{r}^{(2\beta)}_{\epsilon,s}\|_{C^0} \ &= \ \sup_{\substack{x\in\mathbb{R}/L\mathbb{R} \\ t\in[0,T-\epsilon^2]}} \Big|\mathsf{f}^{(n)}_s(t+\epsilon^2, x+s\epsilon) - \sum_{k=0}^{2\alpha+3-n} \epsilon^k\ T_k(\partial_t, s\partial_x)\ \mathsf{f}^{(n)}_s(t,x)\Big| \\ &\le \ \|\mathsf{f}^{(n)}_s\|_{C^{2\alpha+4-2\beta}}\ \epsilon^{2\alpha+4-2\beta} \end{aligned}$$

Using the notation of (4.8), equation (4.35) is written for $n = 2\beta$ with $\beta \in \{0, ..., \alpha+1\}$ more explicitly in the following form:

$$\mathsf{f}^{(2\beta)} \ = \ \sum_{\gamma=0}^{\beta} \Big[P_{2\beta-2\gamma}\mathcal{D}^{(2\gamma)} \ + \ P_{2\beta-2\gamma-1}\mathcal{A}^{(2\gamma)} \ + \ P_{2\beta-2\gamma-2}\mathcal{R}^{(2\gamma)}\Big] \ + \ \tau P_{2\beta-2}\, g\, \mathsf{w}$$

[13] Note, that $\mathrm{LOW}^{(2\beta+1)}$ is a linear combination of derivatives of $u^{(1)}, ..., u^{(2\beta-1)}$. For the exact definition see equation (4.20). In particular, $\mathrm{LOW}^{(2\beta+1)}$ vanishes if $u^{(1)}, ..., u^{(2\beta-1)}$ vanish.

[14] Recall: a advection speed, c reactivity, τ collision time, θ weight parameter.

Using the triangle inequality, this leads to:

$$\|\mathsf{f}_s^{(n)}\|_{C^{2\alpha+4-2\beta}} \leq \Big\{ \sum_{\gamma=0}^{\beta} \big\| P_{2\beta-2\gamma}\mathcal{D}_s^{(2\gamma)} + P_{2\beta-2\gamma-1}\mathcal{A}_s^{(2\gamma)} + P_{2\beta-2\gamma-2}\mathcal{R}_s^{(2\gamma)} \big\|_{C^{2\alpha+4-2\beta}} + \big\|\tau P_{2\beta-2}\, g\, \mathsf{w}_s\big\|_{C^{2\alpha+4-2\beta}} \Big\} \quad (4.38)$$

Let us now simplify the right hand side by estimating it further. Since $P_{2\beta-2}\, g$ contains (mixed) derivatives of g up to order $2\beta-2$, we see, that $\|P_{2\beta-2}\, g\|_{C^{2\alpha+4-2\beta}}$ involves derivatives up to order $(2\beta-2)+(2\alpha+4-2\beta) = 2\alpha+2$. Therefore we get

$$\|\tau P_{2\beta-2}\, g\, \mathsf{w}_s\|_{C^{2\alpha+4-2\beta}} \;\leq\; G^{(2\beta)}\; \|g\|_{C^{2\alpha+2}}$$

with $G^{(2\beta)} := \tau\big(\textit{sum of absolute values of polynomial coefficients of } P_{2\beta-2}(\tau,\cdot,\cdot)\big)$. Similarly, we proceed with the first addend. Introducing the polynomials $p_{s,2\gamma}^{(2\beta)}$, whose coefficients depend only on the parameters a, c, τ and θ via $\mathcal{D}_s(1)$, $\mathcal{A}_s(1)$, $\mathcal{R}_s(1)$,

$$p_{s,2\gamma}^{(2\beta)}(\vartheta,\varsigma) \;:=\; \Big(\mathcal{D}_s(1)\; P_{2(\beta-\gamma)} + \mathcal{A}_s(1)\; P_{2(\beta-\gamma)-1} + \mathcal{R}_s(1)\; P_{2(\beta-\gamma)-2}\Big)(\tau,\vartheta,\varsigma)$$

we define the constants $K_{2\gamma}^{(2\beta)}$ analogously to the $G^{(2\beta)}$'s by:

$$K_{2\gamma}^{(2\beta)} \;:=\; \textit{sum of absolute values of polynomial coefficients of } p_{\oplus,2\gamma}^{(2\beta)}$$

So, the $K_{2\gamma}^{(2\beta)}$'s depend also exclusively on a, c, τ and θ. In the representation of $\mathsf{f}^{(2\beta)}$ derivatives of $u^{(2\gamma)}$ occur up to order $2\beta-2\gamma$ (for $\gamma \leq \beta$) contained in $P_{2\beta-2\gamma} = P_{2\beta-2\gamma}(\tau, \partial_t, s\partial_x)$. That is why the highest derivative of $u^{(2\gamma)}$ occurring on the right hand side of (4.38) is of order $(2\alpha+4-2\beta)+(2\beta-2\gamma) = 2\alpha+4-2\gamma$. Hence we obtain:

$$\big\|P_{2\beta-2\gamma}\mathcal{D}_s^{(2\gamma)} + P_{2\beta-2\gamma-1}\mathcal{A}_s^{(2\gamma)} + P_{2\beta-2\gamma-2}\mathcal{R}_s^{(2\gamma)}\big\|_{C^{2\alpha+4-2\beta}} \;\leq\; K_{2\gamma}^{(2\beta)}\; \|u^{(2\gamma)}\|_{C^{2\alpha+4-2\gamma}}$$

So we conclude:

$$\|\mathsf{r}_{\epsilon,s}^{(2\beta)}\|_{C^0} \;\leq\; \Big\{ \sum_{\gamma=0}^{\beta} K_{2\gamma}^{(2\beta)}\; \|u^{(2\gamma)}\|_{C^{2\alpha+4-2\gamma}} \;+\; G^{(2\beta)}\; \|g\|_{C^{2\alpha+2}} \Big\}\; \epsilon^{2\alpha+4-2\beta}$$

The case $n = 2\beta+1$ with $\beta \in \{0, ..., \alpha+1\}$ is treated in the same way, resulting in an analogous inequality:

$$\|\mathsf{r}_{\epsilon,s}^{(2\beta+1)}\|_{C^0} \;\leq\; \Big\{ \sum_{\gamma=0}^{\beta} K_{2\gamma}^{(2\beta+1)}\; \|u^{(2\gamma)}\|_{C^{2\alpha+4-2\gamma}} \;+\; G^{(2\beta+1)}\; \|g\|_{C^{2\alpha+2}} \Big\}\; \epsilon^{2\alpha+3-2\beta}$$

Multiplying these equations by $\epsilon^{2\beta}$ resp. $\epsilon^{2\beta+1}$, adding them together and summing over $\beta \in \{0, ..., \alpha+1\}$ yields, after changing the order of the double summation on the right hand side:

$$\sum_{n=0}^{2\alpha+3} \|\mathsf{r}_{\epsilon,s}^{(n)}\|_{C^0}\,\epsilon^n \;\le\; \Big\{ \sum_{\gamma=0}^{\alpha} \Big[\sum_{\beta=\gamma}^{\alpha+1} K_{2\gamma}^{(2\beta)} + K_{2\gamma}^{(2\beta+1)} \Big] \|u^{(2\gamma)}\|_{C^{2(\alpha-\gamma)+4}} + \sum_{n=0}^{2\alpha+3} G^{(n)}\,\|g\|_{C^{2\alpha+2}} \Big\}\,\epsilon^{2\alpha+4}$$

Note, that the sum over γ reaches only to α and not to $\alpha+1$, since $u^{(2\alpha+2)}$ vanishes by definition. So, we obtain the assertion by choosing

$$K \;:=\; \max\,\Big\{ \sum_{n=0}^{2\alpha+3} G^{(n)},\; \sum_{\beta=\gamma}^{\alpha+1} K_{2\gamma}^{(2\beta)} + K_{2\gamma}^{(2\beta+1)} \;\text{ with }\; \gamma = 0, ..., \alpha \Big\}\,.$$

Recalling the definition of the constants $G^{(n)}$ and $K_{2\gamma}^{(n)}$, we see that K does neither depend on ϵ nor on the $u^{(2\gamma)}$'s and g. In fact, it is only determined by the polynomials P_k ($k \in \{0, 1, ..., 2\alpha+3\}$) and the indicated parameters. ■

We are now well prepared to prove that $\hat{\mathsf{f}}_h$ solves the discrete lattice-Boltzmann equation on the grid $\mathcal{G}_h$ approximately, up to a remainder-term of order $\mathrm{O}(h^{2\alpha+4}) = \mathrm{O}(N_h^{-2\alpha-4})$.

Theorem 4.2. *Suppose that the existence and smoothness hypothesis is satisfied. May* $\mathsf{f}_\epsilon \in \mathcal{F}(\mathcal{S} \times [0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ *be constructed as described above. Then the subsequent estimate holds true for any admissible mesh-size* $h \in \mathcal{H}_L$ *and all time steps* $n \in \mathbb{T}_h \setminus \{N_T(h)\}$ *with* $1 \le p \le \infty$.

$$\big\|\hat{\mathsf{f}}_h(n+1,\cdot) - \mathfrak{B}_h(n+1)\hat{\mathsf{f}}_h(n,\cdot)\big\|_{L^p} \;\le\; C_p\, h^{2\alpha+4} \tag{4.39}$$

The constant C_p *is independent of* h *and* n

$$C_p := \sqrt[p]{3L}\; K\,\Big(\sum_{\gamma=0}^{\alpha} \|u^{(2\gamma)}\|_{C^{2(\alpha-\gamma)+4}} + \|g\|_{C^{2\alpha+2}} \Big)\,, \tag{4.40}$$

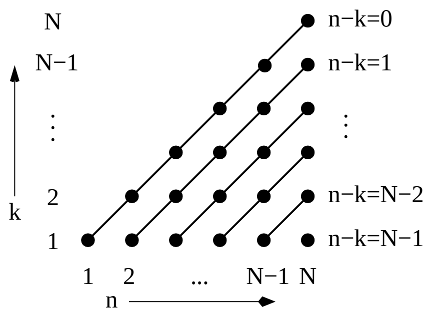
where K *is the constant from lemma 4.6. Thus* f_ϵ *is consistent of order* $2\alpha+4$ *to the lattice-Boltzmann equation with respect to the supremum-norm and* a fortiori *with the weaker* L^p*-norms* ($1 \le p < \infty$).

For the proof of the theorem, the recursive representation formula (4.12) for the asymptotic order functions will be of great help:

$$\mathsf{f}^{(0)} \;=\; \mathcal{E}^{(0)}, \qquad \mathsf{f}^{(n)} \;=\; \mathcal{E}^{(n)} \;-\; \sum_{k=1}^{n} T_k \mathsf{f}^{n-k} \;+\; \delta_{2n}\tau g \mathsf{w} \quad n \ge 1\,. \tag{4.41}$$

Proof: Let us first consider the case of the supremum norm ($p = \infty$). The asserted inequality is just a short notation of

$$|D_h(n,i,s)| \;:=\; \Big| \mathsf{f}_{h,s}(t_n + h^2, x_i) \;-\; \Big\{ (1-\omega)\mathsf{f}_{h,s}(t_n, x_i - sh) + \omega\,\mathcal{E}_{h,s}\Big(\sum_{\tilde{s}\in\mathcal{S}} \mathsf{f}_{h,\tilde{s}}(t_n, x_i - sh) \Big) \;+\; h^2\, g(t_n, x_i - sh)\,\mathsf{w}_s \Big\} \Big| \;\le\; C_p\, h^{2\alpha+4}$$



New 1st index: $\nu := n-k \qquad \nu : 0 \to N-1$
If k 2nd index: $k : 1 \to N-\nu, \quad n = \nu + k$
If n 2nd index: $n : \nu+1 \to N, \quad k = n - \nu$

$$\sum_{n=1}^{N}\sum_{k=1}^{n} a_n b_k c_{n-k} = \sum_{\nu=0}^{N-1}\sum_{k=1}^{N-\nu} a_{\nu+k} b_k c_\nu = \sum_{\nu=0}^{N-1}\sum_{n=\nu+1}^{N} a_n b_{n-\nu} c_\nu$$

Figure 4.3: Change of indexation in a double sum. See equation (4.43).

that is asserted to hold for any $s \in \mathcal{S}, (t_n + h^2, x_i) \in \mathcal{G}_h$. Since we want to prove the estimate independently of $x_i \in \mathcal{L}_h$, i.e. in the supremum norm, we have to verify the following inequality, obtained by replacing x_i with $x_i + sh$, for any $(t_n + h^2, x_i) \in \mathcal{G}_h$ and $s \in \mathcal{S}$.

$$\Big| \mathsf{f}_{h,s}(t_n + h^2, x_i + sh) \; - \; \Big\{ (1-\omega)\mathsf{f}_{h,s}(t_n, x_i) + \omega\, \mathcal{E}_{h,s}\Big(\sum_{\tilde{s}\in\mathcal{S}} \mathsf{f}_{h,\tilde{s}}(t_n, x_i)\Big) \; + \; h^2\, g(t_n, x_i)\, \mathsf{w}_s \Big\} \Big| \;\le\; C_p\, h^{2\alpha+4}$$

Actually, we will show that this equation holds for arbitrary $(t + h^2, x) \in [0, T] \times \mathbb{R}/L\mathbb{R}$, if $|h| \le 1$. By reason of clarity, we omit the arguments (t, x) whenever possible without misinterpretation. Recall in particular $T_k = T_k(\partial_t, s\partial_x)$.

$$(*) := \Big| \mathsf{f}_{h,s}(t + h^2, x + sh) - \Big\{ (1-\omega)\mathsf{f}_{h,s}(t, x) + \omega\mathcal{E}_{h,s}\Big(\sum_{\tilde{s}\in\mathcal{S}} \mathsf{f}_{h,\tilde{s}}(t, x)\Big) + h^2 g(t, x)\mathsf{w}_s \Big\} \Big|$$

Solving (4.37) for the shifted order function, multiplying by h^n and summing over $n \in \{0, 1, ..., \alpha + 3\}$ yields

$$\mathsf{f}_{h,s}(t + h^2, x + sh) \;=\; \sum_{n=0}^{2\alpha+3} h^n \Big[\sum_{k=0}^{2\alpha+3-n} h^k\, T_k \mathsf{f}_s^{(n)}(t, x) + \mathsf{r}_s^{(n)}(t, x) \Big] \,. \tag{4.42}$$

By summing also the representation formula (4.41) over n we obtain

$$\mathsf{f}_{h,s} \;=\; \mathcal{E}_s^{(0)} \;+\; \sum_{n=1}^{2\alpha+3} \mathsf{f}_s^{(n)} h^n \;=\; \mathcal{E}_{h,s}(u_h) \;-\; \tau \sum_{n=1}^{2\alpha+3}\sum_{k=1}^{n} h^n T_k \mathsf{f}_s^{(n-k)} \;+\; h^2 \tau g \mathsf{w}_s \,.$$

As explained in figure 4.3 (set $2\alpha + 3 =: N$ introduce $n - k =: \nu$ as new index, keep k as second index and rename, finally, ν in n), we will change the indexation of the double sum such that it can be put together with the double sum in (4.42).

$$\mathsf{f}_{h,s} \;=\; \mathcal{E}_{h,s}(u_h) \;-\; \tau \sum_{n=0}^{2\alpha+2}\sum_{k=1}^{2\alpha+3-n} h^{n+k}\, T_k \mathsf{f}_s^{(n)} \;+\; h^2 \tau g \mathsf{w}_s \tag{4.43}$$

Let us now plug (4.42) and (4.43) into ($*$). By virtue of lemma (4.5) we have $\mathcal{E}_h\big(\sum_{\tilde{s}\in\mathcal{S}} \mathsf{f}_{h,\tilde{s}}\big) = \mathcal{E}_h(u_h)$. This enables us to put together the equilibrium terms coming directly from the algorithm with those from (4.43). Thus we get:

$$(*) = \Bigg| \sum_{n=0}^{2\alpha+3} \sum_{k=0}^{2\alpha+3-n} h^{n+k}\, T_k \mathsf{f}_s^{(n)} + \sum_{n=0}^{2\alpha+3} h^n\, \mathsf{r}_s^{(n)} + (\omega-1)\mathcal{E}_{h,s}(u_h) + (1-\tau)\, h^2 g \mathsf{w}_s$$
$$+ (\tau-1) \sum_{n=0}^{2\alpha+2} \sum_{k=1}^{2\alpha+3-n} h^{n+k}\, T_k \mathsf{f}_s^{(n)} + \omega \mathcal{E}_{h,s}(u_h) - h^2\, g\, \mathsf{w}_s \Bigg|$$

The double sums can be simplified essentially:

$$\sum_{n=0}^{2\alpha+3} \sum_{k=0}^{2\alpha+3-n} h^{n+k}\, T_k \mathsf{f}_s^{(n)} - \sum_{n=0}^{2\alpha+2} \sum_{k=1}^{2\alpha+3-n} h^{n+k}\, T_k \mathsf{f}_s^{(n)} = \sum_{n=0}^{2\alpha+3-n} h^n\, T_0 \mathsf{f}_s^{(n)} = \mathsf{f}_{h,s}$$

So we obtain:

$$(*) = \Bigg| \mathsf{f}_{h,s} + \sum_{n=0}^{2\alpha+3} h^n\, \mathsf{r}_s^{(n)} - \mathcal{E}_{h,s}(u_h) + \tau \sum_{n=0}^{2\alpha+3} \sum_{k=1}^{2\alpha+3-n} h^n\, T_k \mathsf{f}_s^{(n)} - h^2\, \tau\, g\, \mathsf{w}_s \Bigg|$$

Using again (4.43) for the resubstitution (replace third, fourth and fifth term together by $-\mathsf{f}_{h,s}$) we arrive at

$$(*) = \Bigg| \mathsf{f}_{h,s} + \sum_{n=0}^{2\alpha+3} h^n \mathsf{r}_{h,s}^{(n)} - \mathsf{f}_{h,s} \Bigg| = \Bigg| \sum_{n=0}^{2\alpha+3} h^n\, \mathsf{r}_{h,s}^{(n)} \Bigg| \le \sum_{n=0}^{2\alpha+3} \|\mathsf{r}_{h,s}^{(n)}\|_{C^0}\, h^n .$$

Taking into account the estimate of the remainder terms in lemma (4.6), we find the following inequality implying the assertion.

$$(*) \le \underbrace{K \left(\sum_{\gamma=0}^{\alpha} \|u^{(2\gamma)}\|_{C^{2(\alpha-\gamma)+4}} + \|g\|_{C^{2\alpha+2}} \right)}_{=:C_\infty} h^{2\alpha+4}$$

Note that $\sqrt[\infty]{3L} := \lim\limits_{p\to\infty} \sqrt[p]{3L} = 1$. The case $1 \le p < \infty$ follows easily:

$$\|\hat{\mathsf{f}}_h(n+1) - \mathfrak{B}_h(n+1)\hat{f}_h(n,\cdot)\|_{L^p} = \Big(\sum_{s\in\mathcal{S}} \sum_{i\in\mathbb{Z}/N\mathbb{Z}} \big|D_h(n,i,s)\big|^p \, \tfrac{L}{N} \Big)^{\frac{1}{p}}$$
$$\le \Big(\sum_{s\in\mathcal{S}} \sum_{i\in\mathbb{Z}/N\mathbb{Z}} \big|C_\infty\, h^{2\alpha+4}\big|^p \, \tfrac{L}{N} \Big)^{\frac{1}{p}} = \Big(3N \, \big|C_\infty\, h^{2\alpha+4}\big|^p \, \tfrac{L}{N} \Big)^{\frac{1}{p}} = \underbrace{\sqrt[p]{3L}\, C_\infty}_{=:C_p} h^{2\alpha+4}$$

■

Initialization and Convergence. Next, let us consider the case $\alpha = 0$ in detail. This case is, of course, the simplest as well as the most important one, since it is sufficient to prove convergence of the lattice-Boltzmann scheme. The consideration

of higher orders leads to a truncated asymptotic expansion of the zeroth moment, that additionally describes the leading error terms being not of interest here.
We have already remarked, that $u^{(0)} = u$, where u is just the solution of the target problem (4.1). As no other asymptotic orders are involved, we will always write u instead of $u^{(0)}$. Let us assume, that the existence and smoothness hypothesis is fulfilled for u and g, such that we have $u \in C^4([0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$ and $g \in C^2([0,T] \times \mathbb{R}/L\mathbb{R}, \mathbb{R})$. Then it is possible to construct recursively the following functions by means of the polynomials T_k (see equation (4.10) and table 4.2), where T_k abbreviates $T_k(\partial_t, s\partial_x)$.

$$\begin{aligned} \mathsf{f}^{(0)} &:= \mathcal{D}(u) \\ \mathsf{f}^{(1)} &:= \mathcal{A}(u) \;-\; \tau\;\, T_1\,\mathsf{f}^{(0)} \\ \mathsf{f}^{(2)} &:= \mathcal{R}(u) \;-\; \tau\Big\{\, T_1\,\mathsf{f}^{(1)} \;+\; T_2\,\mathsf{f}^{(0)}\,\Big\} \;+\; \tau\, g\,\mathsf{w} \\ \mathsf{f}^{(3)} &:= \qquad\quad\;\; -\; \tau\Big\{\, T_1\,\mathsf{f}^{(2)} \;+\; T_2\,\mathsf{f}^{(1)} \;+\; T_3\,\mathsf{f}^{(0)}\,\Big\} \end{aligned} \tag{4.44}$$

Note, that this construction procedure is equivalent to (4.35). The approximate population function f_ϵ is then defined by

$$\mathsf{f}_\epsilon \;:=\; \mathsf{f}^{(0)} \;+\; \epsilon\,\mathsf{f}^{(1)} \;+\; \epsilon^2\,\mathsf{f}^{(2)} \;+\; \epsilon^3\,\mathsf{f}^{(3)}\,. \tag{4.45}$$

Explicitly we obtain for the asymptotic orders:

$$\begin{aligned} \mathsf{f}^{(0)}_s(t,x) &= \mathsf{w}_s\; u(t,x) \\ \mathsf{f}^{(1)}_s(t,x) &= \mathsf{w}_s\Big\{\theta\, a\, s \;-\; \tau\, s\,\partial_x\Big\}u(t,x) \\ \mathsf{f}^{(2)}_s(t,x) &= -\tau\mathsf{w}_s\Big\{c \;+\; \theta a s^2\partial_x \;-\; (\tau-\tfrac{1}{2})\,s^2\,\partial_x^2 \;+\; \partial_t\Big\}u(t,x) \;+\; \tau\mathsf{w}_s\, g(t,x) \\ \mathsf{f}^{(3)}_s(t,x) &= \tau^2\mathsf{w}_s\Big\{sc\partial_x + \theta a s^3\partial_x^2 - (\tau-\tfrac{1}{2})s^3\partial_x^3 + s\partial_t\partial_x\Big\}u(t,x) \;+\; \tau\mathsf{w}_s s\partial_x g(t,x) \\ &\quad -\; \tau\mathsf{w}_s\big(\partial_t + \tfrac{1}{2}s^2\partial_x^2\big)\big(\theta a s - \tau s\partial_x\big)u(t,x) \;-\; \tau\mathsf{w}_s\Big\{s\partial_t\partial_x + \tfrac{1}{6}s^3\partial_x^3\Big\}u(t,x) \end{aligned}$$

Considering these functions as polynomials in $s \in \mathcal{S}$, we observe, that $\mathsf{f}^{(1)}_s$ and $\mathsf{f}^{(3)}_s$ have odd parity, while $\mathsf{f}^{(0)}_s$ and $\mathsf{f}^{(2)}_s$ are of even parity. Hence the zeroth moments of $\mathsf{f}^{(1)}_s$ and $\mathsf{f}^{(3)}_s$ vanish. Since u is the solution of the target equation we find

$$\sum_{s\in\mathcal{S}} \mathsf{f}^{(2)}_s \;=\; -\tau\Big\{cu + a\partial_x u - \tfrac{1}{\theta}(\tau-\tfrac{1}{2})\partial_x^2 u + \partial_t u - g\Big\} \;=\; 0\,.$$

So we obtain, without explicitly referring to lemma 4.5,

$$\sum_{s\in\mathcal{S}} \mathsf{f}_{\epsilon,s} \;=\; \sum_{s\in\mathcal{S}} \big[\mathsf{f}^{(0)}_s \;+\; h\,\mathsf{f}^{(1)}_s \;+\; h^2\,\mathsf{f}^{(2)}_s \;+\; h^3\,\mathsf{f}^{(3)}_s\big] \;=\; u\,.$$

Let us now estimate the remainder terms:

$$
\begin{aligned}
|\mathsf{r}_{h,s}^{(0)}(t,x)| &\leq \|u\|_{C^4}\, h^4 \\
|\mathsf{r}_{h,s}^{(1)}(t,x)| &\leq \theta|a|\, \|u\|_{C^3}\, h^3 \;+\; \tau\|\partial_x u\|_{C^3}\, h^3 \;\leq\; \underbrace{(\theta\,|a| + \tau)}_{=:K_1}\, \|u\|_{C^4}\, h^3 \\
|\mathsf{r}_{h,s}^{(2)}(t,x)| &\leq \tau c\|u\|_{C^2} + \tau|a|\theta\|\partial_x u\|_{C^2} + \tau(\tau - \tfrac{1}{2})\|\partial_x^2 u\|_{C^2} + \tau\|\partial_t u\|_{C^2} + \tau\|g\|_{C^2} \\
&\leq \underbrace{\big(\tau c + \theta\tau|a| + \tau(\tau - \tfrac{1}{2}) + \tau\big)}_{=:K_2}\, \|u\|_{C^4}\, h^2 \;+\; \tau\, \|g\|_{C^2}\, h^2 \\
|\mathsf{r}_{h,s}^{(2)}(t,x)| &\leq \underbrace{\Big(\tau^2\big(c + |a|\theta + \tau + \tfrac{1}{2}\big) + \tfrac{3}{2}\tau(|a|\theta + \tau) + \tfrac{7}{6}\tau\Big)}_{=:K_3} \|u\|_{C^4} h + \tau\|g\|_{C^2} h
\end{aligned}
$$

As already mentioned, the lattice-Boltzmann scheme is classically initialized by the equilibrium associated with the initial condition for the target problem. The preceding analysis of the lattice-Boltzmann scheme suggests, however, to initialize the population function by the initial value of the approximated population function. Thus there are four further possibilities of initialization depending on how many orders should be taken into account.

$$
\begin{array}{lclll}
\text{equilibrium: } \mathsf{F}_h(0,i) & = & \mathcal{E}_h\big(w(x_i)\big) & \text{needs: } w & \\
0^{th} \text{ order } : \mathsf{F}_h(0,i) & = & \mathsf{f}_h^{(0)}(0,x_i) & \text{"} & w \\
1^{st} \text{ order } : \mathsf{F}_h(0,i) & = & \sum\limits_{n=0}^{1} \mathsf{f}_h^{(n)}(0,x_i)\, h^n & \text{"} & w, \partial_x w \\
2^{nd} \text{ order } : \mathsf{F}_h(0,i) & = & \sum\limits_{n=0}^{2} \mathsf{f}_h^{(n)}(0,x_i)\, h^n & \text{"} & w, \partial_x w, \partial_x^2 w, g \\
3^{rd} \text{ order } : \mathsf{F}_h(0,i) & = & \sum\limits_{n=0}^{3} \mathsf{f}_h^{(n)}(0,x_i)\, h^n & \text{"} & w, \partial_x w, \partial_x^2 w, \partial_x^3 w, g, \partial_x g
\end{array}
$$

The right column indicates the quantities being necessary for each specific initialization. Usually, the initial value w and the source g are provided in the grid nodes at least. For initializations of first and higher order, additionally spatial derivatives of w and g are needed, which must be supplied numerically (e.g. numeric differentiation) in the worst case, if w and g are not prescribed analytically. Moreover, the expressions for $\mathsf{f}^{(2)}$ and $\mathsf{f}^{(3)}$ contain temporal derivatives of the solution. These can not be computed directly, but they can be procured by taking recourse to the target equation. Solving it for $\partial_t u$ yields with $u(0,\cdot) = w$:

$$
\begin{aligned}
\partial_t u(0,\cdot) &= \nu\partial_x^2 w - a\partial_x w - cu + g(0,\cdot) \\
\partial_t\partial_x u(0,\cdot) &= \nu\partial_x^3 w - a\partial_x^2 w - cu + \partial_x g(0,\cdot)
\end{aligned}
$$

Thus, the knowledge of w, g is sufficient to compute $\mathsf{f}^{(2)}(0,\cdot)$ and $\mathsf{f}^{(3)}(0,\cdot)$.

Obviously, the initialization via equilibrium or zeroth order is consistent of order 1; generally the initialization up to order n is consistent of order $n+1$. After theorem 4.2 our approximated population function is consistent of order 4 to the discrete lattice-Boltzmann equation. In virtue of theorem 4.1, we conclude under the stability assumption, that the *inherent deviation* is for any fixed, finite time of magnitude h^2. Thus it is reasonable to choose the initialization so that the *inherited initial deviation* is of the same size. Any higher order initialization does not considerably decrease the error, whereas a lower order initialization might spoil the result since the inherited initial deviation is then expected to dominate the inherent deviation. Therefore, the initialization should be preferably done up to first order:

$$\mathsf{F}_h(0,i) = w(x_i)\mathsf{w} \;+\; h\left(\theta a - \tau\partial_x\right) w(x_i)\, S\mathsf{w}\,. \tag{4.46}$$

Theorem 4.3. *Consider the target problem (4.1) with $w \in C^4(\mathbb{R}/L\mathbb{R},\mathbb{R})$ and $g \in C^2([0,T]\times\mathbb{R}/L\mathbb{R},\mathbb{R})$ and assume $u \in C^4([0,T]\times\mathbb{R}/L\mathbb{R},\mathbb{R})$. If assumption* (4.27) *is satisfied, then the zeroth moment U_h of the population function converges in the grid nodes to the solution of the target equation:*

$$\|U_h(N_t,\cdot) - \hat{u}(N_t,\cdot)\|_L^p \le 3^{\frac{p-1}{p}}\left(E_pJ_p + E_pC_pT\right)h^2.$$

4.4 Initial behavior

In this section we shortly study the initial behavior of the lattice-Boltzmann algorithm in a phenomenological way by performing numerical experiments that illustrate the principal observations.

In the previous section the population function F_h corresponding to the output of the lattice-Boltzmann algorithm, is approximated by a truncated power series with respect to the grid spacing h. This approach turns out to provide an understanding of how the lattice-Boltzmann algorithm behaves. The validity of the regular expansion can only be justified in a strict sense if the lattice-Boltzmann algorithm is initialized by the expansion. Practically, however, this condition might be quite difficult to fulfill (especially for more complicated lattice-Boltzmann algorithms) since the expansion requires derivatives of the initial data if the first and even higher order terms should be taken into account.

So the question arises whether the accuracy of the lattice-Boltzmann algorithm can be sensibly spoiled by the initialization. In the case of a large diffusivity it is experimentally observed that a crude initialization affects the result only slightly, as if the algorithm tries by itself to behave in a nice manner and to correct the bad initialization. But for small diffusivities (ω close to 2) persistent initial layers appear.

Observation of an initial layer

As a first example, let us consider the D1P2 algorithm obtained from the D1P3 scheme setting $\theta = 1$. Furthermore we take the pure diffusion equation as target problem, i.e. $a = c = 0$ in (4.1). So it is enough to consider only the diffusive part $\mathcal{D}$ of the equilibrium (4.2).

For a given solution u, we obtain from page 186 the prediction function including the zeroth and first order of the regular expansion:

$$\mathsf{f}_{h,s}(t,x) = \mathsf{f}_s^{(0)}(t,x) + h\mathsf{f}_s^{(1)}(t,x) = \tfrac{1}{2}u(t,x) - h\tfrac{1}{2}s\tau\partial_x u(t,x)\ .$$

From theorem 4.1 it can be concluded that

$$\mathsf{F}_{h,s}(n,i) = \mathsf{f}_{h,s}(nh^2, ih) + \mathrm{O}(h^2) \quad \text{if} \quad \mathsf{F}_{h,s}(0,i) = \mathsf{f}_{h,s}(0,ih).$$

Now we consider the lattice-Boltzmann algorithm initialized only the zeroth order corresponding to the diffusive equilibrium

$$\mathsf{F}_s(0,i) = \mathsf{f}^{(0)}(0,ih) = \tfrac{1}{2}u(0,hi).$$

Surely, the first order of the populations will no more be equal to $\mathsf{f}^{(1)}$. Instead, we expect the appearance of an additional term $\mathsf{L}_{h,s}$, the *initial layer*, that should cancel $\mathsf{f}_s^{(1)}$ for $n = 0$ in order to fulfill the initial condition

$$\mathsf{F}_{h,s}(n,i) = \mathsf{f}_s^{(0)}(nh^2, ih)\ +\ h\ \underbrace{\Big(\mathsf{f}_s^{(1)}(nh^2, ih) + \mathsf{L}_{h,s}(n,i)\Big)}_{\overset{!}{=}\,0}\ + \mathrm{O}(h^2)\,. \tag{4.47}$$

Note that this expansion makes only sense if the initial layer is uniformly bounded for $h \in \mathcal{H}_L$ and $(n,i) \in \mathbb{N}_0 \times \mathbb{Z}/N_h\mathbb{Z}$.
Inserting (4.47) into the lattice-Boltzmann equation, exploiting the linearity of the equilibrium and utilizing the fact that $\mathsf{f}_{h,s}$ satisfies the lattice-Boltzmann equation up to second order, we find after dividing by h

$$\mathsf{L}_{h,s}(n+1, i+s) = (1-\omega)\mathsf{L}_{h,s}(n,i) + \tfrac{1}{2}\omega\Big(\mathsf{L}_{h,s}(n,i) + \mathsf{L}_{h,-s}(n,i)\Big) + \mathrm{O}(h)\ .$$

So, we see that $\mathsf{L}_{h,s}$ must solve the lattice-Boltzmann equation up to a residue of first order. For a more detailed examination of initial layers, it is reasonable (due to the linearity of the equilibrium) to *define* $\mathsf{L}_{h,s}$ as solution of the discrete lattice-Boltzmann equation with the initial condition $\mathsf{L}_{h,s}(0,i) = -\mathsf{f}_s^{(1)}(0,ih)$. This will be done in the next paragraph.

To get an impression of how an initial layer looks like, let us take the "damped sinus" solving the diffusion equation for a little experiment

$$u(t,x) := e^{-4\pi^2\nu t}\ \sin(2\pi x) \qquad x \in \mathbb{R}/1\mathbb{R}. \tag{4.48}$$

If the lattice-Boltzmann algorithm is initialized via the equilibrium, the initial layer can be made visible by plotting the difference[15] between the populations and the two leading orders of the expansion:

$$\mathsf{L}_{h,s}(n,i) = \underbrace{\frac{1}{h}\Big[\mathsf{F}_s(n,i)\ -\ \tfrac{1}{2}u(t,x)\ -\ h\tfrac{1}{2}s\tau\partial_x u(t,x)\Big]}_{=:\mathsf{R}_{h,s}(n,i)} + \mathrm{O}(h)\ . \tag{4.49}$$

[15]The difference is divided by the grid spacing h in order to get an unscaled quantity, that is in leading order independent of the grid size. This facilitates the comparison of this quantity on grids of different size.

We call this quantity the *residual of first order*.

Spatial plots for fixed times suggest, that $\mathsf{R}_{h,s}(n,\cdot)$ behaves smoothly in space; it seems to be proportional to $\cos(2\pi x)$ i.e. $\partial_x u(0,x)$. The quantity reveals a different behavior if we look at its evolution in time. For this the residual in a fixed node is plotted versus the iteration number. The following observations are striking:

- There are three qualitatively different modes of residuals distinguished fundamentally by their temporal behavior (see figure 4.4):
  - i) $2 > \omega > 1$ (over-relaxation): damped oscillations (on time-step level)
  - ii) $1 > \omega > 0$ (under-relaxation): exponential decay.
  - iii) $\omega = 1$ (limit case): no initial layer.
- Th residual of the right and left moving population differ by their sign, i.e. $\mathsf{R}_{h,s}(n,i) \approx -\mathsf{R}_{h,-s}(n,i)$. (see figure 4.4)
- Plotting the residual versus the iteration number for different grid spacings $h_1 \neq h_2$, yields almost congruent curves, i.e. $\mathsf{R}_{h_1,s}(n,i_1) \approx \mathsf{R}_{h_2,s}(n,i_2)$ where $i_1 h_1 = i_2 h_2$ denote the same spatial position. See figure 4.5.

Initial-layer-type output of the D1P2 lattice-Boltzmann algorithm

Now, let us consider $\mathsf{L}_{h,s}$ as solution of the following initial value problem for the discrete lattice-Boltzmann equation

$$\begin{aligned} \mathsf{L}_{h,s}(0,i) &= \tfrac{1}{2}\, s\, w(ih) \\ \mathsf{L}_{h,s}(n+1,i+s) &= (1-\omega)\,\mathsf{L}_{h,s}(n,i) \;+\; \omega\; \underbrace{\tfrac{1}{2}\Big(\mathsf{L}_{h,s}(n,i) + \mathsf{L}_{h,-s}(n,i)\Big)}_{\mathcal{D}_s\left(\sum_{\sigma\in\mathcal{S}} \mathsf{L}_{h,\sigma}(n,i)\right)}, \end{aligned}$$

where w denotes a smooth function on the spatial domain. In the case of the preceding example we have $w(x) \propto \cos(2\pi x)$.
As the qualitative behavior of $\mathsf{L}_{h,s}$ practically leads to the same observations as those concerning the first order residual (see above), this motivates the following ansatz to approximate $\mathsf{L}_{h,s}$:

$$\mathsf{L}_{h,s}(n,i) \;\approx\; \mathsf{A}^{(0)}_s(n,i) \;:=\; \tfrac{1}{2}\, g(n)\, s\, w(ih)\;.$$

The unknown function $g : \mathbb{N}_0 \to \mathbb{R}$ shall be defined in such a way, that $\mathsf{A}^{(0)}$ satisfies the lattice-Boltzmann equation up to a residue of order $\mathrm{O}(h)$. Noticing that the diffusive equilibrium vanishes since $\mathsf{A}^{(0)}_\ominus + \mathsf{A}^{(0)}_\oplus = 0$, we obtain

$$\begin{aligned} g(n+1)\, s\, w(ih+sh) &= (1-\omega)\, g(n)\, s\, w(ih) \\ \Rightarrow\quad g(n+1)\, s\, w(ih) \;+\; \mathrm{O}(h) &= (1-\omega)\, g(n)\, s\, w(ih) \end{aligned}$$

This equation is satisfied if $g(n+1) = (1-\omega)g(n)$. Using the initial condition $g(0) = 1$ we find $g(n) = (1-\omega)^n$ and therefore

$$\mathsf{A}^{(0)}_s(n,i) \;=\; \tfrac{1}{2}\,(1-\omega)^n\, s\, w(ih)\;.$$

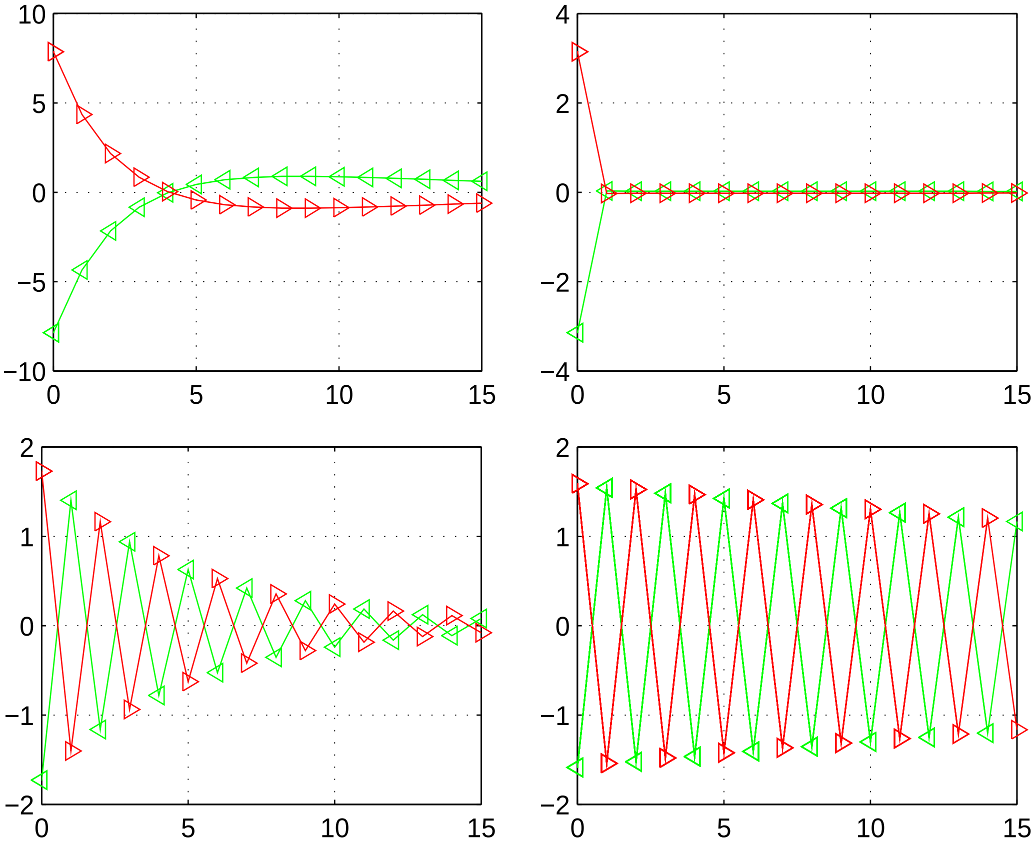


Figure 4.4: The four figures display the (approximate) first order residual (cf. equation (4.49)) in the leftmost grid node plotted versus the iteration number for different values of the diffusivity ν. In each figure the two curves representing the left ($\triangleleft$) and right ($\triangleright$) moving population seem to be equal up to the sign. Moreover the figures indicate, how the residual is influenced by the diffusivity ν related to the collision frequency via $\omega = (\nu + \frac{1}{2})^{-1}$: monotonous exponential decay for $\nu = 2$ (left above), no initial layer for $\nu = 0.5$ (right above), damped oscillations for $\nu = 0.05$ (left below) and $\nu = 0.005$ (right below). The experiments were performed on a grid of 40 nodes.

Since $\mathsf{L}_{h,s}(0, i) = \mathsf{A}_s^{(0)}(0, i)$, this suggests that $\mathsf{L}_{h,s}(n, i) = \mathsf{A}_s^{(0)}(n, i) + \mathrm{O}(h)$.

By means of $\mathsf{A}^{(0)}$ we can explain our observations. For positive diffusivity, i.e. $0 < \omega < 2$ we have $|1 - \omega| < 1$. Therefore the initial layer tends in leading order exponentially with respect to the iteration number n to zero. The *damping rate* is given by

$$\delta := \Big| \log |1 - \omega| \Big| .$$

For $1 - \omega < 0$ i.e. $1 < \omega$ the coefficient $(1 - \omega)^n$ changes from iteration to iteration its sign resulting in the observed oscillations. Table 4.4 gives a summary.

Let us try to approximate $\mathsf{L}_{h,s}$ up to second order. For this we have to find a function $\mathsf{A}^{(1)}$, such that $\mathsf{A}_s^{(0)} + h\mathsf{A}_s^{(1)}$ satisfies the lattice-Boltzmann equation up to

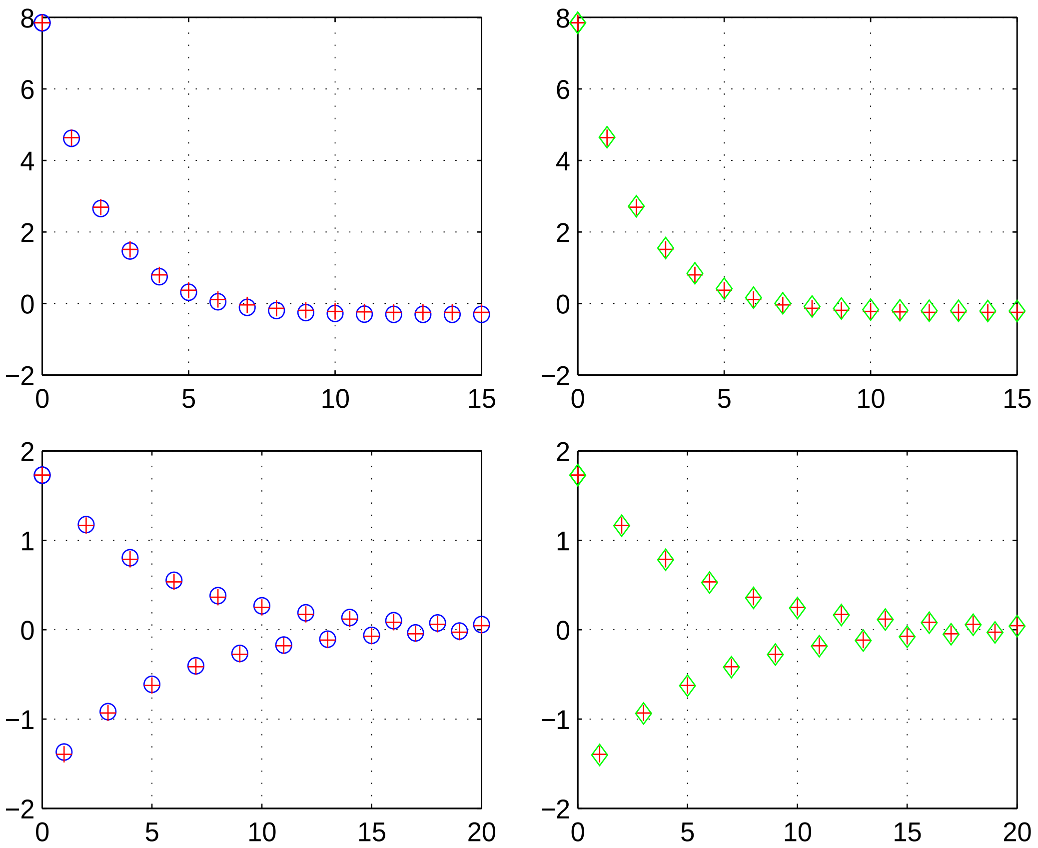


Figure 4.5: Again the (approximate) first order residual (cf. equation (4.49)) is plotted versus the iteration number, but this time only for the right moving population at the leftmost grid. Instead, grids with different number of nodes (resulting in different mesh-spacings) have been considered. Above: $\nu = 2$, number of grid nodes: 80 ($\circ$), 90 ($+$), 100 ($\diamond$). Below: $\nu = 0.05$, number of grid nodes: 20 ($\circ$), 30 ($+$), 40 ($\diamond$). For $\nu = 2$ the second order term of the regular expansion,also contained in the plotted quantity, becomes quite large and even dominating if the grid is not fine enough. Therefore a higher grid resolution has been chosen for the upper figures. Thus the first order of the residual is better recognized. Clearly, it depends only on ν and the iteration number n but not on the grid size h.

mode	behavior
under-relaxed: $0 < \omega < 1$	damped, monotonous: $(1-\omega)^n = e^{-\delta n}$
relaxed : $\omega = 1$	limit case : $(1-\omega)^n = 0$
over-relaxed : $1 < \omega < 2$	damped, oscillating : $(1-\omega)^n = (-1)^n e^{-\delta n}$

Table 4.4: Behavior of the initial layer in dependence of the collision frequency ω.

second order (is consistent up to second order) and $\mathsf{A}_s^{(1)}(0,i) = 0$. We make the following ansatz

$$\mathsf{A}_s^{(1)}(n,i) \;=\; \tfrac{1}{2}\,\mathsf{g}_s(n)\,\partial_x w(ih)$$

with $\mathsf{g} : \mathcal{S} \times \mathbb{N}_0 \to \mathbb{R}$. Plugging $\mathsf{A}^{(0)}(n,i) + h\mathsf{A}^{(1)}(n,i)$ into the lattice-Boltzmann equation

$$\begin{aligned}
&\tfrac{1}{2}(1-\omega)^{n+1} s w(ih+sh) \;+\; \tfrac{1}{2}h\mathsf{g}_s(n+1)\partial_x w(ih+sh)\\
&\qquad = \tfrac{1}{2}(1-\omega)(1-\omega)^n s w(ih) \;+\; \tfrac{1}{2}h(1-\omega)\mathsf{g}_s(n)\partial_x w(ih)\\
&\qquad\qquad +\; \tfrac{1}{2}\omega\big[\tfrac{1}{2}h\mathsf{g}_s(n) + \tfrac{1}{2}h\mathsf{g}_{-s}(n)\big]\partial_x w(ih)
\end{aligned}$$

multiplying by 2 and expanding the terms on the left hand side up to $\mathrm{O}(h^2)$ yields:

$$\begin{aligned}
&(1-\omega)^{n+1} s\Big[w(ih) + sh\partial_x w(ih) + \mathrm{O}(h^2)\Big] + \mathsf{g}_s(n+1)\Big[h\partial_x w(ih) + \mathrm{O}(h^2)\Big]\\
&\quad = (1-\omega)^{n+1} s w(ih) + h(1-\omega)\mathsf{g}_s(n)\partial_x w(ih) + \omega\Big[\tfrac{1}{2}\mathsf{g}_s(n) + \tfrac{1}{2}\mathsf{g}_{-s}(n)\Big]\partial_x w(ih)
\end{aligned}$$

After dividing the equation by $h\partial_x w(ih)$ and sending h to zero we obtain:

$$(1-\omega)^{n+1}\underbrace{s^2}_{=1} + \mathsf{g}_s(n+1) \;=\; (1-\omega)\mathsf{g}_s(n) \;-\; \tfrac{1}{2}\omega\big[\mathsf{g}_s(n) + \mathsf{g}_{-s}(n)\big]$$

Thus we find a recursion relation for g, which takes explicitly the following form:

$$\begin{pmatrix}\mathsf{g}_\ominus(n+1)\\ \mathsf{g}_\oplus(n+1)\end{pmatrix} \;=\; \begin{pmatrix}1-\frac{1}{2}\omega & \frac{1}{2}\omega\\ \frac{1}{2}\omega & 1-\frac{1}{2}\omega\end{pmatrix}\begin{pmatrix}\mathsf{g}_\ominus(n)\\ \mathsf{g}_\oplus(n)\end{pmatrix} \;-\; \begin{pmatrix}(1-\omega)^{n+1}\\ (1-\omega)^{n+1}\end{pmatrix}$$

Obviously, this equation together with the initial condition $\mathsf{g}_s(0) = 0$ is completely symmetric with respect to $\mathsf{g}_\ominus, \mathsf{g}_\oplus$. Therefore $\mathsf{g}_\ominus(n) = \mathsf{g}_\oplus(n) =: g(n)$ and

$$g(0) = 0 \qquad g(n+1) = g(n) + (1-\omega)^{n+1}\;.$$

The solution of this scalar equation is essentially given by the geometric series with respect to the geometric sequence $k \mapsto -(1-\omega)^k$.

$$\begin{aligned}
g(n) \;&=\; 1 \;-\; \sum_{k=0}^{n}(1-\omega)^k \;=\; 1 \;-\; \frac{1-(1-\omega)^{n+1}}{1-(1-\omega)}\\
&=\; (1-\tau) + \tau(1-\omega)^{n+1}
\end{aligned}$$

So we end up with $\;\mathsf{A}_s^{(1)}(n,i) \;=\; \tfrac{1}{2}\left[1 - \tau + \tau(1-\omega)^{n+1}\right]\partial_x w(ih)\;.$

It is striking, that this coefficient does not tend to zero for n going to infinity. Instead

$$\lim_{n\to\infty}\mathsf{A}_s^{(1)} = \tfrac{1}{2}(1-\tau)\partial_x w(ih)\;, \tag{4.50}$$

if $0 < \omega < 2$. This is in contradiction to the numerical experiments and to the dissipative behavior of the lattice-Boltzmann algorithm in general. In order to obtain a more realistic first order coefficient, an heuristic argument will bail us out.

If both of populations of the D1P2 lattice-Boltzmann algorithm are initialized by $\frac{1}{2}(1-\tau)\partial_x w(ih)$, corresponding to (4.50), the leading (zeroth) order is described by the regular expansion. So this amounts to

$$\mathsf{F}_{h,s}(n,i) = \tfrac{1}{2}(1-\tau)v(nh^2, ih) + \mathrm{O}(h)$$

where $v(t,x)$ would be the solution to the subsequent initial value problem for the diffusion equation:

$$\begin{aligned} v(0,x) &= (1-\tau)\,\partial_x w(x)\,, \\ \partial_t v \;-\; \nu\,\partial_x^2 v &= 0\,. \end{aligned}$$

We can imagine that after the initialization of the lattice-Boltzmann algorithm by $\mathsf{A}_s^{(0)} + h\mathsf{A}_s^{(1)}$, the terms containing $(1-\omega)^n$ vanish relatively rapidly, such that (4.50) is left in the first order. This apparently reinitializes the lattice-Boltzmann algorithm once more, with the only difference that now the initial condition is an $\mathrm{O}(h)$-term instead of an $\mathrm{O}(1)$-term. Due to the linearity of the algorithm we therefore suppose[16] that the actual $A^{(1)}$ is given by

$$\mathsf{A}_s^{(1)}(n,i) \;=\; \tfrac{1}{2}\left[1-\tau+\tau(1-\omega)^{n+1}\right] v(nh^2, ih)\;.$$

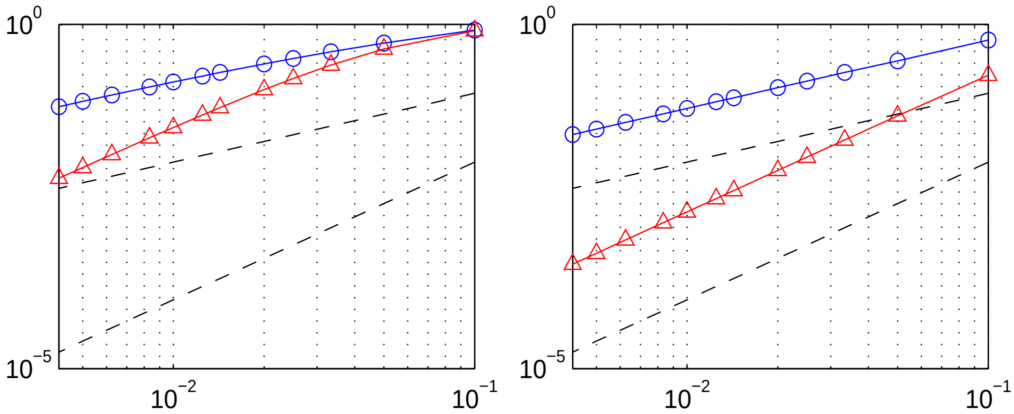


Figure 4.6: The conjecture is backed by a numerical convergence study with $w(x) = \cos(2\pi x)$. Here twelve grids have been used with 10, 20, 30, 40, 50, 70, 80, 100, 120, 160, 200 and 240 nodes. The diffusivity is $\nu = 3$ (left) and $\nu = 0.01$ (right). Circles ($\circ$) indicate the approximation rate for $\mathsf{A}^{(0)}$ while triangles ($\triangle$) mark the approximation rate for $\mathsf{A}^{(0)} + h\mathsf{A}^{(1)}$. Observe that the slope of the curves corresponds to the values in table 4.5 agreeing quite well with the ideal values of 1 and 2 respectively. The small deviations are due to higher order effects diminishing slightly the ideal values.

[16] Alternatively, the following ansatz seems to be reasonable in the same way at the current stage:

$$\mathsf{A}_s^{(1)}(n,i) \;=\; \tfrac{1}{2}\left[(1-\tau)v(nh^2, ih) + \tau(1-\omega)^{n+1}\partial_x w(ih)\right]$$

ν	0^{th} ord. appr.	1^{st} ord. appr.
3	0.9425	1.9348
1.5	0.9786	1.9717
0.7	0.9951	1.9914
0.01	0.9994	1.9997

Table 4.5: Rates of asymptotic similarity.

D1P3 model with diffusion equilibrium

In order to observe the initial behavior of the general D1P3 model, we consider once more the diffusion equation such that the solution is given by (4.48). As in the previous case, the initial behavior is essentially determined by the relaxation parameter ω. For $\omega < 1$ the initial layer decays monotonically, while for $\omega > 1$ a damped oscillating comportment is observed. In both cases the damping rate is also determined by the factor $(1 - \omega)$ and occurs in the discrete time (iteration).

In contrast to the special case of the D1P2 model, we discover *beat-bellies* in the overrelaxed case, i.e. the absolute value of the initial layer swells for several iterations and ebbs then away before it starts to swell again. Considering the same initial value problem on grids of different mesh size, it turns out, that necks and bellies of the beats coincide if they are plotted versus the scaled quantity nh, where n denotes the number of iteration and h corresponds to the grid spacing. This quantity is referred to as the *fast time*, since it advances faster than the standard or slow time given by nh^2.

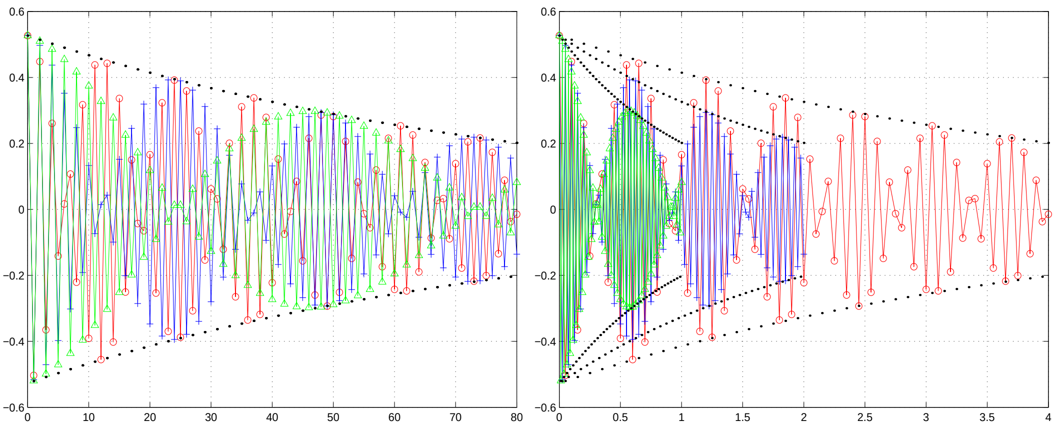


Figure 4.7: Initial layers of the right-moving population at the leftmost node on three different grids are plotted versus the iteration (left) and versus the fast time (right). Note that at left the exponential envelopes for the damping (black dotted lines) coincide for the three grids, whereas the extrema of the beats are attained simultaneously with respect to the fast time (right). Apparently the damping takes place in the iteration scale while the beats are found in the fast time scale. The number of grid nodes is 20 (red $\circ$), 40 (blue $+$), 80 (green $\triangle$). Other parameters: $\nu = 0.001, \theta = 3$.

Chapter 5

Long-term behavior of an advective lattice-Boltzmann scheme

The preceding chapter has shown that *regular asymptotic expansions* are well suited to extract essential properties of lattice-Boltzmann algorithms. In particular, we were able to derive the target equation, to predict the convergence order and to acquire a qualitative understanding of the error production. While shortly considering initial layers, it became also clear that regular expansions can not be used to describe all aspects of lattice-Boltzmann algorithms.

The analyzed D1P3 lattice-Boltzmann algorithm is based on a *parabolically* scaled lattice-Boltzmann equation. In this chapter we focus on a D1P2 algorithm related to a *hyperbolically* scaled lattice-Boltzmann equation. Numerical experiments indicate the occurrence of numerical errors that cannot be explained satisfactorily by means of standard regular expansions alone.
Besides the *advection* equation, which plays the role of the target equation in this case, the hyperbolically scaled lattice-Boltzmann algorithm implicitly solves a *diffusion* equation in the background. As this happens with respect to a shorter time step[1], the effect becomes only apparent after a relatively large number of iterations. In order to capture both the advective foreground process and the diffusive background process being a long-term phenomenon, we set up a prediction function in the form of a two-scale expansion. This turns out to be an appropriate tool elucidating the interaction between the processes. Technically, the two-scale expansion extends the regular one. So it is performed most efficiently after being prepared by the regular expansion. Due to this reason and to assure the self-containment of this chapter, section 5.1 concentrates on the regular expansion. Profiting in section 5.2 from this preparatory work, we motivate the appearance of different time scales in 5.2.1 prior to embark on the asymptotic analysis in 5.2.2. The two-scale expansion for an interesting special case involving *dispersive* effects is presented in subsection 5.2.3.

This chapter remains mainly on a formal level. However, the derived expansions can be justified rigorously if the algorithm is proven to be stable. This is accomplished in section 6.3. There we will also return to the following question shortly touched in 5.2.2: is it possible to draw conclusions about stability on the basis of

[1]Depending on the process, different time increments are associated with one iteration.

formal asymptotics and multiscale expansions in particular? The issue is discussed studying the advective D1P2 lattice-Boltzmann algorithm in an exemplary manner.

Let us finally remark, that the natural time scales inherent in Boltzmann-like equations emerge clearly in this model algorithm. While the advective process is prevailing, the diffusive process counteracts as disturbance. Combining both processes on the slow time scale, leads directly to an advective-diffusive D1P2 scheme whose generalization onto the D1P3 model was subject of the previous chapter.

A few recapitulating words concerning the notation

Lattice-Boltzmann methods are based on discrete velocity particle models. Concretely, we consider fictitious particles that can move with unit speed in one space dimension either to the left or to the right. Hence the discrete velocities are given by $\mathcal{S} := \{s_1, s_2\} = \{-1, 1\}$. The particle distribution is described by a vector-valued function referred to as the *population function*. The first component represents the density of the particles traveling to the left, while the second component is associated with the other species:

$$\mathsf{F}(t,x) = \big[\mathsf{F}_k(t,x)\big]_{k\in\{1,2\}} = \begin{pmatrix} \mathsf{F}_1(t,x) \\ \mathsf{F}_2(t,x) \end{pmatrix}.$$

Observe that the components, which we denote as *populations*, depend on time and space. In order to keep the conception[2] clear, we designate $\mathbb{R}^2$ by $\mathbb{R}^{\mathcal{S}}$ if the indices refer to the discrete velocities. Elements of $\mathbb{R}^{\mathcal{S}}$ are highlighted by sans-serif characters. Let us introduce the following abbreviations, that will simplify the computations in later sections:

$$\mathsf{1} := \binom{1}{1} \in \mathbb{R}^{\mathcal{S}} \qquad \mathsf{s} := \binom{s_1}{s_2} = \binom{-1}{1} \in \mathbb{R}^{\mathcal{S}}$$

The scalar product in $\mathbb{R}^{\mathcal{S}}$ is indicated by brackets. In particular we have:

$$\langle \mathsf{f}, \mathsf{g}\rangle := \mathsf{f}_1\mathsf{g}_1 + \mathsf{f}_2\mathsf{g}_2 \quad \Rightarrow \quad \langle \mathsf{1}, \mathsf{1}\rangle = \langle \mathsf{s}, \mathsf{s}\rangle = 2 \quad \text{and} \quad \langle \mathsf{1}, \mathsf{s}\rangle = 0. \tag{5.1}$$

For $\mathsf{f} \in \mathbb{R}^{\mathcal{S}}$ we call the scalar products $\langle \mathsf{f}, \mathsf{1}\rangle$ and $\langle \mathsf{f}, \mathsf{s}\rangle$ *mass moment* (0^{th} moment) and *flux moment* (1^{st} moment) respectively. Furthermore, we will extensively make use of componentwise multiplication for $\mathsf{f}, \mathsf{g} \in \mathbb{R}^{\mathcal{S}}$:

$$\mathsf{g}\mathsf{f} = \mathsf{f}\mathsf{g} = \binom{\mathsf{f}_1}{\mathsf{f}_2}\binom{\mathsf{g}_1}{\mathsf{g}_2} := \binom{\mathsf{f}_1\mathsf{g}_1}{\mathsf{f}_2\mathsf{g}_2} \quad \Rightarrow \quad \mathsf{1}\mathsf{f} = \mathsf{f}, \quad \mathsf{s}^2 = \begin{pmatrix} (-1)^2 \\ 1^2 \end{pmatrix} = \mathsf{1}, \quad \mathsf{s}^3 = \mathsf{s}\,.$$

By slight abuse[3] of notation, we understand the usually invalid operation $a + \mathsf{f}$ with $a \in \mathbb{R}$ and $\mathsf{f} \in \mathbb{R}^{\mathcal{S}}$ in the sense of $a + \mathsf{f} := a\mathsf{1} + \mathsf{f}$.

[2] In allusion to the *phase-space density* of the *Boltzmann equation*, it would be natural to describe the particle distribution with a function $(t, x, s) \mapsto \tilde{\mathsf{F}}(t, x, s)$ where the variable s can only take the two values s_1 and s_2. However, we abstain from this point of view, using the more familiar vector notation $\mathsf{F}_k(t, x) = \tilde{\mathsf{F}}(t, x, s_k)$ instead.

[3] A reader familiar with MATLAB may notice the analogy with respect to the componentwise multiplication operator (`.*`) for matrices as well as the addition of scalars and matrices.

Presentation of the advective D1P2 lattice-Boltzmann algorithm

The algorithm we are going to investigate has the standard lattice-Boltzmann form

$$\mathsf{F}_k(t+h, x+s_k h) = \mathsf{F}_k(t,x) + \big[J\mathsf{F}(t,x)\big]_k, \qquad k = 1,2 \tag{5.2}$$

where the discretization parameter $h = 1/N$, $N \in \mathbb{N}$ determines the space-time grid. More precisely, t can take the values $t_n = nh$ with $n \in \mathbb{N}_0$ and x ranges in $x_i = ih$ with $i \in \mathbb{Z}$. The collision operator J on the right hand side of (5.2) models the particle interaction. Here, we choose the simple BGK form

$$J\mathsf{F} = \omega(\mathsf{E}(U) - \mathsf{F})$$

with relaxation parameter ω. The equilibrium distribution $\mathsf{E}(U)$ is assumed to depend only on the mass moment U of F

$$\mathsf{E}(U) := \tfrac{1}{2}(1 + a\mathsf{s})U, \qquad U = \langle \mathsf{F}, 1\rangle.$$

The role of the parameter $a \in \mathbb{R}$ will be explained later. In view of (5.1) we observe that

$$\langle \mathsf{E}(U), 1\rangle = \tfrac{1}{2}U\langle 1,1\rangle + \tfrac{1}{2}aU\langle \mathsf{s}, 1\rangle = U$$

so that

$$\langle J\mathsf{F}, 1\rangle = \omega\left\langle \mathsf{E}\big(\langle \mathsf{F}, 1\rangle\big), 1\right\rangle - \omega\langle \mathsf{F}, 1\rangle = \omega\langle \mathsf{F}, 1\rangle - \omega\langle \mathsf{F}, 1\rangle = 0$$

which expresses the mass conservation during the collision process. We remark that the collision operator is linear because the equilibrium distribution depends linearly on U and hence on F. In fact, we can write for a general $\mathsf{f} \in \mathbb{R}^{\mathcal{S}}$

$$\begin{aligned}\mathsf{E}_1(\langle \mathsf{f}, 1\rangle) &= \tfrac{1}{2}(1 + as_1)\langle \mathsf{f}, 1\rangle = \tfrac{1}{2}(1-a)\mathsf{f}_1 + \tfrac{1}{2}(1-a)\mathsf{f}_2\\ \mathsf{E}_2(\langle \mathsf{f}, 1\rangle) &= \tfrac{1}{2}(1 + as_2)\langle \mathsf{f}, 1\rangle = \tfrac{1}{2}(1+a)\mathsf{f}_1 + \tfrac{1}{2}(1+a)\mathsf{f}_2\end{aligned}$$

so that $\mathsf{E}(\langle \mathsf{f}, 1\rangle) = E\mathsf{f}$ with the matrix

$$E := \tfrac{1}{2}\begin{pmatrix} 1-a & 1-a \\ 1+a & 1+a \end{pmatrix} \tag{5.3}$$

which satisfies the projector property $E = E^2$. Using this notation, the collision operator is given by the matrix $J = \omega(E - I)$ and the algorithm (5.2) appears in the form

$$\mathsf{F}_k(t+h, x+s_k h) = (1-\omega)\mathsf{F}_k(t,x) + \omega\big[E\mathsf{F}(t,x)\big]_k. \tag{5.4}$$

Apart from this update equation we require suitable initial and boundary conditions to fully specify the particle dynamics.

A typical initialization rule is based on the equilibrium distribution. Given some function $x \mapsto v_0(x)$, one may set

$$\mathsf{F}(0,x) = \mathsf{E}(v_0(x)). \tag{5.5}$$

Improvements of this basic initialization are derived in the sequel. Let us assume that v_0 is a 1-periodic function, i.e. $v_0(x+1) = v_0(x)$. Since the grid with the nodes

$x_i = ih = i/N$ for $i \in \mathbb{Z}$ is also 1-periodic as $x_i + 1 = x_{i+N}$, we conclude that the population function F is 1-periodic too. In particular, it suffices to compute F on N consecutive nodes $\{x_0, \ldots, x_{N-1}\}$. The required information from the neighbors of the trailing nodes follows from periodicity

$$\mathsf{F}(t, x_{-1}) = \mathsf{F}(t, x_{N-1}), \qquad \mathsf{F}(t, x_N) = \mathsf{F}(t, x_0) \tag{5.6}$$

which can be regarded as a boundary condition.

5.1 Regular expansion

5.1.1 Analysis of the update rule

Implementing the lattice-Boltzmann algorithm introduced above with initial value $v_0(x) = \cos(2\pi x)$, for example, we observe the following dynamics of $U = \langle \mathsf{F}, \mathsf{1} \rangle$: provided $|a|$ is small enough (otherwise the algorithm is unstable), the initial cosine-shape of the graph moves with a speed related to the parameter a. Repeating the simulation with smaller grid size h, we observe similar profiles. It turns out that the amplitude of the cosine reduces during the movement and this reduction is more pronounced for larger values of h.

In order to understand the behavior of the lattice-Boltzmann solution $\mathsf{F}(t, x)$ more thoroughly, we try to approximate it in the form of a regular h-expansion

$$\mathsf{F}(t, x) \approx \mathsf{f}^{[\alpha]}(t, x) := \mathsf{f}^{(0)}(t, x) + h\mathsf{f}^{(1)}(t, x) + \ldots + h^{\alpha}\mathsf{f}^{(\alpha)}(t, x) \tag{5.7}$$

with $t = t_n = nh$ and $x = x_j = jh$, $n \in \mathbb{N}, j \in \mathbb{Z}$. Note that $\mathsf{f}^{[\alpha]}$ is a polynomial[4] of order $\alpha \in \mathbb{N}$ with respect to the grid parameter h. We refer to $\mathsf{f}^{[\alpha]}$ as *prediction function*. The *asymptotic order functions* $\mathsf{f}^{(\beta)}$ with $0 \le \beta \le \alpha$ are supposed to be h-independent. Hence they are defined on the whole unit interval $[0, 1]$. In order to take the boundary conditions (5.6) of the algorithm into account, they should be extendible to smooth, 1-periodic functions on $\mathbb{R}$.

If the prediction function $\mathsf{f}^{[\alpha]}$ nicely approximates[5] the population function F for small h, then $\mathsf{f}^{[\alpha]}$ also satisfies the lattice-Boltzmann algorithm (5.4) up to a small residual. This observation is exploited to determine the coefficients $\mathsf{f}^{(\beta)}$: they are chosen in such a way that the residual has highest possible order in h and thus vanishes with optimal speed as h tends to zero. More specifically, we require for all $h > 0$

$$\mathsf{f}_k^{[\alpha]}(t + h, x + s_k h) = (1 - \omega)\mathsf{f}_k^{[\alpha]}(t, x) + \omega\big[E\mathsf{f}^{[\alpha]}(t, x)\big]_k \;+\; \mathrm{O}(h^{\alpha+1}). \tag{5.8}$$

[4]This particular h-dependence is motivated by the theorem of Weierstraß which states that the polynomials are dense in the space of continuous functions over a compact interval like $[0, 1]$. Hence any continuous function is arbitrarily well approximated by polynomials. This confirms, that the set of polynomials might be ample enough, to approximate the grid dependency of the population function F provided it behaves continuously in h.

[5]Strictly speaking the approximation refers to the restriction of $\mathsf{f}^{[\alpha]}$ onto the corresponding grid.

Under the assumption that the asymptotic order functions are sufficiently smooth, we substitute $\mathsf{f}_k^{[\alpha]}(t+h, x+s_k h)$ by its Taylor expansion around (t,x) up to the order α:

$$\mathsf{f}_k^{[\alpha]}(t+h, x+s_k h) = \sum_{\beta=0}^{\alpha} \tfrac{1}{\beta!} h^\beta \left(\partial_t + s_k \partial_x \right)^\beta \mathsf{f}_k^{[\alpha]}(t,x) \;+\; \mathrm{O}(h^{\alpha+1}). \tag{5.9}$$

A convenient effect of the Taylor expansion is, that it disentangles the index-depending spatial argument at the left hand side. Therefore we may now write vectors instead of components (i.e. f in place of f_k), such that equation (5.8) in combination with the Taylor expansion reads[6]

$$\sum_{\beta=0}^{\alpha} \tfrac{1}{\beta!} h^\beta \,(\underbrace{\partial_t + \mathsf{s}\partial_x}_{=:D})^\beta \,\mathsf{f}^{[\alpha]} = (1-\omega)\mathsf{f}^{[\alpha]} + \omega E \mathsf{f}^{[\alpha]} \;+\; \mathrm{O}(h^{\alpha+1}).$$

If we replace $\mathsf{f}^{[\alpha]}$ by the truncated series in (5.7) and incorporate those terms into $\mathrm{O}(h^{\alpha+1})$ that are proportional to some power of h with an exponent greater than α, we obtain

$$\sum_{\beta=0}^{\alpha} \sum_{\gamma=0}^{\beta} \tfrac{1}{\gamma!} h^\beta \, D^\gamma \, \mathsf{f}^{(\beta-\gamma)} = (1-\omega) \sum_{\beta=0}^{\alpha} h^\beta \mathsf{f}^{(\beta)} + \omega \sum_{\beta=0}^{\alpha} h^\beta E \mathsf{f}^{(\beta)} \;+\; \mathrm{O}(h^{\alpha+1}).$$

This equation, that shall hold true for arbitrary h, enforces the equality of the coefficients referring on both sides to equal powers of h. In particular the *comparison of coefficients* yields for $\alpha \geq 4$ and $0 \leq \beta \leq 4$:

$$\left.\begin{array}{lll}
h^0: & \mathsf{f}^{(0)} & = (1-\omega)\mathsf{f}^{(0)} + \omega E \mathsf{f}^{(0)} \\
h^1: & \mathsf{f}^{(1)} + D\mathsf{f}^{(0)} & = (1-\omega)\mathsf{f}^{(1)} + \omega E \mathsf{f}^{(1)} \\
h^2: & \mathsf{f}^{(2)} + D\mathsf{f}^{(1)} + \frac{1}{2}D^2\mathsf{f}^{(0)} & = (1-\omega)\mathsf{f}^{(2)} + \omega E \mathsf{f}^{(2)} \\
h^3: & \mathsf{f}^{(3)} + D\mathsf{f}^{(2)} + \frac{1}{2}D^2\mathsf{f}^{(1)} + \frac{1}{6}D^3\mathsf{f}^{(0)} & = (1-\omega)\mathsf{f}^{(3)} + \omega E \mathsf{f}^{(3)} \\
h^4: & \mathsf{f}^{(4)} + D\mathsf{f}^{(3)} + \frac{1}{2}D^2\mathsf{f}^{(2)} + \frac{1}{6}D^3\mathsf{f}^{(1)} + \frac{1}{24}D^4\mathsf{f}^{(0)} & = (1-\omega)\mathsf{f}^{(4)} + \omega E \mathsf{f}^{(4)}
\end{array}\right\} \tag{5.10}$$

Next, we gather in each equation the terms of highest index at the left hand side, while the other terms are collected at the right:

$$\left.\begin{array}{lcl}
(I-E)\mathsf{f}^{(0)} & = & 0 \\
(I-E)\mathsf{f}^{(1)} & = & -\frac{1}{\omega}D\mathsf{f}^{(0)} \\
(I-E)\mathsf{f}^{(2)} & = & -\frac{1}{\omega}D\mathsf{f}^{(1)} - \frac{1}{2\omega}D^2\mathsf{f}^{(0)} \\
(I-E)\mathsf{f}^{(3)} & = & -\frac{1}{\omega}D\mathsf{f}^{(2)} - \frac{1}{2\omega}D^2\mathsf{f}^{(1)} - \frac{1}{6\omega}D^3\mathsf{f}^{(0)} \\
(I-E)\mathsf{f}^{(4)} & = & -\frac{1}{\omega}D\mathsf{f}^{(3)} - \frac{1}{2\omega}D^2\mathsf{f}^{(2)} - \frac{1}{6\omega}D^3\mathsf{f}^{(1)} - \frac{1}{24\omega}D^4\mathsf{f}^{(0)}
\end{array}\right\} \tag{5.11}$$

Obviously, we obtain for each asymptotic order function a linear equation with the common system matrix $I-E$. Furthermore the asymptotic order functions depend

[6]We consider D as a special differentiation operator, therefore we should define it by $D := I\partial_t + S\partial_x$, where I is the 2×2 identity matrix and $S := \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Note that the matrix multiplication with I and S is equivalent to the vector multiplication with $\mathsf{1}$ and s respectively.

hierarchically on their "comrades" of lower order, which appear as inhomogeneities. This implies that the equations must be solved successively starting with the homogeneous equation for $\mathsf{f}^{(0)}$. However, the singularity of $I-E$ seems to create a problem. In fact, we have

$$I - E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 1-a & 1-a \\ 1+a & 1+a \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1+a & -1+a \\ -1-a & 1-a \end{pmatrix} \tag{5.12}$$

which shows that the rank of $I-E$ is only 1 (the rows differ only by sign). At first sight this seems to be disturbing, because the lack of an inverse matrix $(I-E)^{-1}$ entails, that the solution for arbitrary right hand sides may not exist or is not unique in the affirmative case. A second glance reveals, however, that the regular expansion ansatz would fail, if $I-E$ were invertible. In this case $\mathsf{f}^{(0)}$ could only be 0 yielding by induction 0 for all the other asymptotic order functions. So $\mathsf{f}^{[\alpha]}$ would have no other choice than to vanish thus trapping the ansatz into a *cul-de-sac*.
Let us now look under what circumstances the 2×2 system $(I-E)\mathsf{x} = \mathsf{y}$ has a solution. As the image space of $I-E$ is spanned[7] by s

$$\mathrm{image}(I-E) = \mathrm{span}\left\{\begin{pmatrix} -1 \\ 1 \end{pmatrix}\right\} = \mathrm{span}\{\mathsf{s}\},$$

the inhomogeneous equations admit only solutions for right hand sides being multiples of s. Since $\{\mathsf{1},\mathsf{s}\}$ forms an orthogonal basis in $\mathbb{R}^{\mathcal{S}}$ every vector and in particular y admits the representation

$$\mathsf{y} = \tfrac{1}{2}\langle \mathsf{y},\mathsf{1}\rangle \mathsf{1} + \tfrac{1}{2}\langle \mathsf{y},\mathsf{s}\rangle \mathsf{s}.$$

where the factor $\frac{1}{2} = \langle \mathsf{1},\mathsf{1}\rangle^{-1} = \langle \mathsf{s},\mathsf{s}\rangle^{-1}$ takes into account that the basis vectors are not normalized. The condition of y being a multiple of s enforces $\langle \mathsf{y},\mathsf{1}\rangle = 0$. Therefore the mass moment of the inhomogeneity y must vanish. Consequently, equations (5.11) are supplemented by the following *solvability conditions*.

$$\left.\begin{array}{rcl} \langle D\mathsf{f}^{(0)},\mathsf{1}\rangle & = & 0 \\ \langle D\mathsf{f}^{(1)},\mathsf{1}\rangle & = & -\frac{1}{2}\langle D^2\mathsf{f}^{(0)},\mathsf{1}\rangle \\ \langle D\mathsf{f}^{(2)},\mathsf{1}\rangle & = & -\frac{1}{2}\langle D^2\mathsf{f}^{(1)},\mathsf{1}\rangle - \frac{1}{6}\langle D^3\mathsf{f}^{(0)},\mathsf{1}\rangle \\ \langle D\mathsf{f}^{(3)},\mathsf{1}\rangle & = & -\frac{1}{2}\langle D^2\mathsf{f}^{(2)},\mathsf{1}\rangle - \frac{1}{6}\langle D^3\mathsf{f}^{(1)},\mathsf{1}\rangle - \frac{1}{24}\langle D^4\mathsf{f}^{(0)},\mathsf{1}\rangle \end{array}\right\} \tag{5.13}$$

In the forthcoming paragraphs we extract from these equations:

- explicit formulas representing the asymptotic order functions in terms of their mass moments,
- evolution equations determining the mass moments.

Before we begin with this program we need to know the nullspace of $I-E$, which is given by the range of E, i.e.:

$$\mathrm{kernel}(I-E) = \mathrm{span}\left\{\frac{1}{2}\begin{pmatrix} 1-a \\ 1+a \end{pmatrix}\right\} = \mathrm{span}\big\{\tfrac{1}{2}(\mathsf{1}+a\mathsf{s})\big\}. \tag{5.14}$$

[7] The column vectors of $I-E$ in (5.12) are multiples of s, more precisely the left column vector is $(1+a)\mathsf{s}$, the right one $(-1+a)\mathsf{s}$.

Note, that the factor $\frac{1}{2}$ serves only to normalize the mass moment.

It shall be emphasized that we have not yet said anything, so far, about the existence of the prediction function. All we do is to derive necessary conditions, that must be fulfilled by a sufficiently smooth prediction function and its asymptotic order functions in order to generate a residual of the desired magnitude. However, as soon as we find functions complying with the obtained conditions, we can infer the existence of the prediction function.

Discussion of the 0^{th} order equation: representation of $\mathsf{f}^{(0)}$

Equation (5.11_0) implies that $\mathsf{f}^{(0)}(t,x)$ must belong to the kernel of $I - E$ independently of its arguments t, x. According to equation (5.14), $\mathsf{f}^{(0)}(t,x)$ should be a multiple of $\frac{1}{2}(1 + a\mathsf{s})$. The factor of proportionality, abbreviated by $u^{(0)}$, may vary with t and x like $\mathsf{f}^{(0)}$ itself and must be considered as a scalar function. So we get:

$$\boxed{\mathsf{f}^{(0)} = \tfrac{1}{2}(1 + a\mathsf{s})u^{(0)}} \tag{5.15}$$

At present, $u^{(0)}$ appears to be completely arbitrary. In the next paragraph we will see, that the solvability condition for equation (5.11_1) imposes the necessary constraints on $u^{(0)}$ to fix it. However, taking the scalar product[8] with 1 we recognize already now that $u^{(0)}$ plays the role of the mass moment pertaining to $\mathsf{f}^{(0)}$.

Discussion of the 1^{st} solvability condition: evolution of $u^{(0)}$

Writing down equation (5.13_0), which is the solvability condition for (5.11_1), yields in detail:

$$\begin{aligned} 0 &= \langle D\mathsf{f}^{(0)}, 1\rangle = \tfrac{1}{2}\big\langle\big(\partial_t + \mathsf{s}\partial_x\big)\big(1 + a\mathsf{s}\big),\ 1\big\rangle u^{(0)} \\ &= \tfrac{1}{2}\big\langle\partial_t + \mathsf{s}(a\partial_t + \partial_x) + a\partial_x,\ 1\big\rangle u^{(0)} \\ &= \partial_t u^{(0)} + a\partial_x u^{(0)}\,, \end{aligned}$$

This shows, that $u^{(0)}$ must be a solution of the advection equation:

$$\boxed{\partial_t u^{(0)} + a\partial_x u^{(0)} = 0} \tag{5.16}$$

Discussion of the 1^{st} order equation: representation of $\mathsf{f}^{(1)}$

From linear algebra it is well known, that the general solution of an inhomogeneous, linear equation like (5.11_1) is composed of a specific solution plus an arbitrary solution of the homogeneous equation. In complete analogy to (5.11_0), the solution of the homogeneous equation $(I - E)\mathsf{f}^{(1)} = 0$ is given by $\frac{1}{2}(1 + a\mathsf{s})u^{(1)}$, where – similarly to $u^{(0)}$ above – $u^{(1)}$ is an unknown function to be determined later.

[8] Observe: $\langle\frac{1}{2}(1 + a\mathsf{s}), 1\rangle = 1$

How does a specific solution of the inhomogeneous equation $(I - E)\mathsf{x} = \mathsf{y}$ look like? From $E^2 = E$ we conclude that $I - E$ is a projector too, because

$$(I - E)^2 \;=\; I^2 - 2E + E^2 \;=\; I - 2E + E \;=\; I - E\,.$$

Hence, if $\mathsf{y} \in \mathrm{image}(I - E)$ which entails $\mathsf{y} = (I - E)\mathsf{z}$ for some $\mathsf{z} \in \mathbb{R}^{\mathcal{S}}$, we conclude that

$$(I - E)\mathsf{y} = (I - E)^2\mathsf{z} = (I - E)\mathsf{z} = \mathsf{y},$$

which proves that y itself solves the equation $(I - E)\mathsf{x} = \mathsf{y}$ and is therefore a specific solution.

Applying this result we get the subsequent representation of $\mathsf{f}^{(1)}$

$$\mathsf{f}^{(1)} = \tfrac{1}{2}(1 + a\mathsf{s})u^{(1)} - \tfrac{1}{\omega}D\mathsf{f}^{(0)},$$

which tacitly assumes the solvability condition $\langle D\mathsf{f}^{(0)}, 1\rangle = 0$ or equivalently equation (5.16) to hold true. From this and $\langle \mathsf{s}, 1\rangle = 0$ we deduce $\langle \mathsf{f}^{(1)}, 1\rangle = u^{(1)}$. Thus $u^{(1)}$ emerges as mass moment of $\mathsf{f}^{(1)}$.

In order to obtain a fully explicit expression we need to compute $D\mathsf{f}^{(0)}$.

$$\begin{aligned} D\mathsf{f}^{(0)} &= \tfrac{1}{2}(\partial_t + \mathsf{s}\partial_x)(1 + a\mathsf{s})u^{(0)} \;=\; \tfrac{1}{2}\overbrace{(\partial_t + a\partial_x)u^{(0)}}^{=0} + \tfrac{1}{2}\mathsf{s}(a\partial_t + \partial_x)u^{(0)} \\ &= \tfrac{1}{2}(1 - a^2)\mathsf{s}\partial_x u^{(0)} \end{aligned}$$

To get rid of the overbraced addend and to transmute the temporal derivative into a spatial one, we have applied equation (5.16) twice. With this preparation, we finally arrive at:

$$\boxed{\mathsf{f}^{(1)} = \tfrac{1}{2}(1 + a\mathsf{s})u^{(1)} - \tfrac{1}{2\omega}(1 - a^2)\mathsf{s}\partial_x u^{(0)}} \tag{5.17}$$

It is also easily checked by this formula, that $u^{(1)}$ corresponds to the mass moment of $\mathsf{f}^{(1)}$.

Extracting further orders: the general procedure

Before we enter the derivation of the evolution equations for $u^{(1)}$ and $u^{(2)}$, which might disguise the underlying thread, let us sketch how the computations proceed. We have seen by solving (5.11_0) and (5.11_1), that we are required to introduce a scalar function $u^{(\beta)}$ for each asymptotic order function. As consequence of the solvability condition it turns out that $u^{(\beta)}$ has the meaning of the mass moment $\langle \mathsf{f}^{(\beta)}, 1\rangle$. Furthermore, $u^{(\beta)}$ is basically determined as a solution of an evolution equation that comes out from the solvability condition pertaining to the next order $\mathsf{f}^{(\beta+1)}$.

Using this knowledge we try to organize the computation of further asymptotic order functions more efficiently. So we introduce the $u^{(\beta)}$'s directly by setting

$u^{(\beta)} := \langle \mathsf{f}^{(\beta)}, 1\rangle$. Abbreviating $E\mathsf{f}^{(\beta)}$ by $\mathsf{E}(u^{(\beta)}) =: \mathsf{E}^{(\beta)}$ we transform (5.11) into

$$\left.\begin{array}{rcl} \mathsf{f}^{(0)} &=& \mathsf{E}^{(0)} \\ \mathsf{f}^{(1)} &=& \mathsf{E}^{(1)} - \frac{1}{\omega}D\mathsf{f}^{(0)} \\ \mathsf{f}^{(2)} &=& \mathsf{E}^{(2)} - \frac{1}{\omega}D\mathsf{f}^{(1)} - \frac{1}{2\omega}D^2\mathsf{f}^{(0)} \\ \mathsf{f}^{(3)} &=& \mathsf{E}^{(3)} - \frac{1}{\omega}D\mathsf{f}^{(2)} - \frac{1}{2\omega}D^2\mathsf{f}^{(1)} - \frac{1}{6\omega}D^3\mathsf{f}^{(0)} \\ \mathsf{f}^{(4)} &=& \mathsf{E}^{(4)} - \frac{1}{\omega}D\mathsf{f}^{(3)} - \frac{1}{2\omega}D^2\mathsf{f}^{(2)} - \frac{1}{6\omega}D^3\mathsf{f}^{(1)} - \frac{1}{24\omega}D^4\mathsf{f}^{(0)} \end{array}\right\}. \tag{5.18}$$

In general we obtain

$$\mathsf{f}^{(\beta)} = \mathsf{E}^{(\beta)} - \tfrac{1}{\omega 1!}D\mathsf{f}^{(\beta-1)} - \ldots - \tfrac{1}{\omega\beta!}D^\beta\mathsf{f}^{(0)}. \tag{5.19}$$

This provides a recursive representation of the asymptotic order functions. After substituting β by $\beta+1$ and taking the scalar product with 1 we recover the solvability condition for $\mathsf{f}^{(\beta+1)}$

$$\langle D\mathsf{f}^{(\beta)}, 1\rangle = -\tfrac{1}{2}\langle D^2\mathsf{f}^{(\beta-1)}, 1\rangle - \ldots - \tfrac{1}{(\beta+1)!}\langle D^{\beta+1}\mathsf{f}^{(0)}, 1\rangle, \tag{5.20}$$

where we have additionally used the identity $\langle \mathsf{E}^{(\gamma)}, 1\rangle = u^{(\gamma)} = \langle \mathsf{f}^{(\gamma)}, 1\rangle$.
This equation is the point of departure for deriving the evolution equation of $u^{(\beta)}$. The first step is to insert the recursive representation of $\mathsf{f}^{(\beta)}$ into (5.20). Shoveling all scalar products on the right hand side with exception of the equilibrium term, we arrive at

$$\langle D\mathsf{E}^{(\beta)}, 1\rangle = c_{\beta-1}\langle D^2\mathsf{f}^{(\beta-1)}, 1\rangle + \ldots + c_0\langle D^{\beta+1}\mathsf{f}^{(0)}, 1\rangle,$$

with $c_{\beta-1}, ..., c_0$ denoting some coefficients. Independently of β, the left hand side becomes the differential operator $\partial_t + a\partial_x$ applied to $u^{(\beta)}$. So $u^{(\beta)}$ satisfies an advection equation that is driven by a source term depending on the mass moments $u^{(0)}, ..., u^{(\beta-1)}$ of the lower-order asymptotic order functions. To make the source more explicit, we employ recursively the representation formulas (counterparts of (5.19) for $\mathsf{f}^{(\beta-1)}, ..., \mathsf{f}^{(1)}$) and obtain eventually

$$\partial_t u^\beta + a\partial_x u^{(\beta)} = \tilde{c}_{\beta-1}\langle D^2\mathsf{E}^{(\beta-1)}, 1\rangle + \ldots + \tilde{c}_0\langle D^{\beta+1}\mathsf{E}^{(0)}, 1\rangle$$

with some new constants $\tilde{c}_{\beta-1}, ..., \tilde{c}_0$. It only remains to expand the powers of D and to evaluate the scalar products. Then the source terms contain both spatial and temporal derivatives of the $u^{(\beta)}$'s. In order to eliminate the time derivatives, one successively resorts to the evolution equations for $u^{(0)}, \ldots, u^{(\beta-1)}$. Actually, this step[9] is the most tedious one of the whole computation. Concluding, we see that the computation can be performed quasi mechanically up to an arbitrary maximal order α.
Note, that we do not meet an equation for the mass moment of the highest order $u^{(\alpha)}$, because this would require to introduce a further asymptotic order function $\mathsf{f}^{(\alpha+1)}$. Without reducing the order of the residual we are free to choose $u^{(\alpha)}$. In accordance with the initial conditions (see next subsection) it can be set to 0.

[9]We remark that this step is not really necessary; furthermore it might be difficult to perform it in other situations, where the target equation is more complicated than the advection equation. The main reason for the replacement of temporal derivatives is to obtain finally expressions for the population functions that depend only on spatial derivatives. If we want to use these expressions to improve the initialization of the algorithm, it will be relatively easy to compute spatial derivatives of the initial data in contrast to time derivatives.

Evolution equation for $u^{(1)}$: Following the instruction from above we insert first (5.18_1) and then (5.18_0) into (5.13_1):

$$\begin{aligned}\langle D\mathsf{E}^{(1)}, 1\rangle &= (\tfrac{1}{\omega} - \tfrac{1}{2})\langle D^2\mathsf{f}^{(0)}, 1\rangle \\ &= (\tfrac{1}{\omega} - \tfrac{1}{2})\langle D^2\mathsf{E}^{(0)}, 1\rangle \,.\end{aligned}$$

Next, we evaluate $\langle D\mathsf{E}^{(1)}, 1\rangle = \partial_t u^{(1)} + a\partial_x u^{(1)}$ and

$$\begin{aligned}\langle D^2\mathsf{E}^{(0)}, 1\rangle &= \tfrac{1}{2}\Big\langle \big(\partial_t^2 + 2\mathsf{s}\partial_t\partial_x + \partial_x^2\big)(1 + a\mathsf{s}),\ 1\Big\rangle\, u^{(0)} \\ &= \big(\partial_t^2 + \partial_x^2 + 2a\partial_t\partial_x\big)u^{(0)} \ = \ (1 - a^2)\partial_x^2 u^{(0)},\end{aligned}$$

where we have employed repeatedly (5.16) $\partial_t u^{(0)} = -a\partial_x u^{(0)}$ to eliminate the time derivatives. Combining the results produces the evolution equation for $u^{(1)}$.

$$\boxed{\partial_t u^{(1)} + a\partial_x u^{(1)} = (\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2)\partial_x^2 u^{(0)}} \tag{5.21}$$

Evolution equation for $u^{(2)}$: Once again we repeat the procedure. Plugging first (5.18_2) into (5.13_2) leads to

$$\langle D\mathsf{E}^{(2)}, 1\rangle \ = \ \big(\tfrac{1}{\omega} - \tfrac{1}{2}\big)\langle D^2\mathsf{f}^{(1)}, 1\rangle + \big(\tfrac{1}{2\omega} - \tfrac{1}{6}\big)\langle D^3\mathsf{f}^{(0)}, 1\rangle \,.$$

Insertion of (5.18_1) for $\mathsf{f}^{(1)}$ and (5.18_2) for $\mathsf{f}^{(2)}$ gives

$$\langle D\mathsf{E}^{(2)}, 1\rangle \ = \ \big(\tfrac{1}{\omega} - \tfrac{1}{2}\big)\langle D^2\mathsf{E}^{(1)}, 1\rangle - \big(\tfrac{1}{\omega^2} - \tfrac{1}{\omega} + \tfrac{1}{6}\big)\langle D^3\mathsf{E}^{(0)}, 1\rangle \,. \tag{5.22}$$

Since the left hand side ends up in the advection operator applied to $u^{(2)}$, it remains to care about the two addends of the right hand side. In order to eliminate finally all time derivatives, we express first $\partial_t^2 u^{(1)}$ by space derivatives of $u^{(1)}$ and $u^{(0)}$. For this we employ the evolution equations (5.21) and (5.16):

$$\begin{aligned}\partial_t^2 u^{(1)} &= -a\partial_x\partial_t u^{(1)} + (\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2)\partial_x^2\partial_t u^{(0)} \\ &= -a\partial_x\Big(- a\partial_x u^{(1)} + (\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2)\partial_x^2 u^{(0)}\Big) - (\tfrac{1}{\omega} - \tfrac{1}{2})a(1 - a^2)\partial_x^3 u^{(0)} \\ &= a^2\partial_x^2 u^{(1)} - 2a(\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2)\partial_x^3 u^{(0)}.\end{aligned}$$

This preparation together with further substitutions using (5.21) and (5.16) then yields:

$$\begin{aligned}\langle D^2\mathsf{E}^{(1)}, 1\rangle &= \tfrac{1}{2}\Big\langle(\partial_t^2 + 2\mathsf{s}\partial_t\partial_x + \partial_x^2)(1 + a\mathsf{s}), 1\Big\rangle u^{(1)} \ = \ \big(\partial_t^2 + 2a\partial_t\partial_x + \partial_x^2\big)u^{(1)} \\ &= a^2\partial_x^2 u^{(1)} - 2a(\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2)\partial_x^3 u^{(0)} \\ &\quad + \partial_x^2 u^{(1)} + 2a\partial_x\Big(- a\partial_x u^{(1)} + (\tfrac{1}{\omega} - \tfrac{1}{2})\partial_x^2 u^{(0)}\Big) \\ &= (1 - a^2)\partial_x^2 u^{(1)}\end{aligned}$$

For the second addend we get; utilizing (5.16) again,

$$\begin{aligned}\langle D^3\mathsf{E}^{(0)}, 1\rangle &= \tfrac{1}{2}\Big\langle\big(\partial_t^3 + 3\mathsf{s}\partial_t\partial_x^2 + 3\partial_t\partial_x^2 + \mathsf{s}\partial_x^3\big)\big(1 + a\mathsf{s}\big),\ 1\Big\rangle u^{(0)} \\ &= \Big(\partial_t^3 + 3a\partial_t^2\partial_x + 3\partial_t\partial_x^2 + a\partial_x^3\Big)u^{(0)} \ = \ \Big(- a^3 + 3a^3 - 3a + a\Big)\partial_x^3 u^{(0)} \\ &= -2a(1 - a^2)\partial_x^3 u^{(0)}.\end{aligned}$$

Combining these results with (5.22) eventually produces the evolution equation for $u^{(2)}$.

$$\boxed{\boldsymbol{\partial_t u^{(2)} + a\partial_x u^{(2)}} = (\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2)\boldsymbol{\partial_x^2 u^{(1)}} + 2a(\tfrac{1}{\omega^2} - \tfrac{1}{\omega} + \tfrac{1}{6})(1 - a^2)\boldsymbol{\partial_x^3 u^{(0)}}} \tag{5.23}$$

Explicit representation of $\mathsf{f}^{(2)}$: In order to provide explicit expressions[10] for the asymptotic order functions we must go through the recursion as given by (5.18) and (5.19). For $\mathsf{f}^{(2)}$ we start with the preliminary calculations to get $D\mathsf{f}^{(1)}$ and $D^2\mathsf{f}^{(2)}$:

$$D\mathsf{f}^{(1)} = \tfrac{1}{2}(1+a\mathsf{s})(\partial_t + \mathsf{s}\partial_x)u^{(1)} - \tfrac{1}{2\omega}(1-a^2)\mathsf{s}(\partial_t + \mathsf{s}\partial_x)\partial_x u^{(0)}$$

Eliminating the temporal derivatives by reverting to (5.16) and (5.21) we continue:

$$\begin{aligned} D\mathsf{f}^{(1)} &= \tfrac{1}{2}(1+a\mathsf{s})\Big(-a\partial_x u^{(1)} + (\tfrac{1}{\omega} - \tfrac{1}{2})(1-a^2)\partial_x^2 u^{(0)}\Big) + \tfrac{1}{2}(1+a\mathsf{s})\mathsf{s}\partial_x u^{(1)} \\ &\quad - \tfrac{1}{2\omega}(1-a^2)\mathsf{s}(-a\partial_x + \mathsf{s}\partial_x)\partial_x u^{(0)} \\ &= \tfrac{1}{2}(1-a^2)\mathsf{s}\partial_x u^{(1)} + \tfrac{1}{2}(1-a^2)\Big((\tfrac{1}{\omega} - \tfrac{1}{2})(1+a\mathsf{s}) - \tfrac{1}{\omega}(1-a\mathsf{s})\Big)\partial_x^2 u^{(0)} \\ &= \tfrac{1}{2}(1-a^2)\mathsf{s}\partial_x u^{(1)} + \tfrac{1}{2}(1-a^2)\Big(\tfrac{2}{\omega}a\mathsf{s} - \tfrac{1}{2}a\mathsf{s} - \tfrac{1}{2}\Big)\partial_x^2 u^{(0)} \end{aligned}$$

Likewise we obtain:

$$\begin{aligned} D^2\mathsf{f}^{(0)} &= \tfrac{1}{2}(1+a\mathsf{s})(\partial_t^2 + 2\mathsf{s}\partial_t\partial_x + \partial_x^2)u^{(0)} = \tfrac{1}{2}(1+a\mathsf{s})(a^2\partial_x^2 - 2a\mathsf{s}\partial_x^2 + \partial_x^2)u^{(0)} \\ &= \tfrac{1}{2}(1-a^2+a^3\mathsf{s} - a\mathsf{s})\partial_x^2 u^{(0)} = \tfrac{1}{2}(1-a^2)(1-a\mathsf{s})\partial_x^2 u^{(0)} \end{aligned}$$

Plugging these expressions into (5.18_2) yields

$$\begin{aligned} \mathsf{f}^{(2)} &= \mathsf{E}^{(2)} - \tfrac{1}{\omega}D\mathsf{f}^{(1)} - \tfrac{1}{2\omega}D^2\mathsf{f}^{(0)} \\ &= \tfrac{1}{2}(1+a\mathsf{s})u^{(2)} - \tfrac{1}{2\omega}(1-a^2)\mathsf{s}\partial_x u^{(1)} \\ &\quad - \tfrac{1}{2\omega}(1-a^2)\Big(\tfrac{2}{\omega}a\mathsf{s} - \tfrac{1}{2}a\mathsf{s} - \tfrac{1}{2}\Big)\partial_x^2 u^{(0)} - \tfrac{1}{4\omega}(1-a^2)(1-a\mathsf{s})\partial_x^2 u^{(0)} \end{aligned}$$

and finally:

$$\boxed{\mathsf{f}^{(2)} = \tfrac{1}{2}(1+a\mathsf{s})u^{(2)} - \tfrac{1}{2\omega}(1-a^2)\mathsf{s}\partial_x u^{(1)} - \tfrac{1}{\omega}(\tfrac{1}{\omega} - \tfrac{1}{2})(1-a^2)a\mathsf{s}\partial_x^2 u^{(0)}} \tag{5.24}$$

5.1.2 Smooth initialization and consistency

Analysis of initial conditions

With the considerations above, we have come quite close to a thorough understanding of the lattice-Boltzmann solution. Starting out with an asymptotic expansion $\mathsf{F}(t,x) \approx \mathsf{f}^{(0)}(t,x) + h\mathsf{f}^{(1)}(t,x) + h^2\mathsf{f}^{(2)}(t,x) + \ldots$, we derived the precise form of the leading order coefficients

$$\left.\begin{aligned} \mathsf{f}^{(0)} &= \tfrac{1}{2}(1+a\mathsf{s})u^{(0)} \\ \mathsf{f}^{(1)} &= \tfrac{1}{2}(1+a\mathsf{s})u^{(1)} - \tfrac{1}{2\omega}(1-a^2)\mathsf{s}\partial_x u^{(0)} \\ \mathsf{f}^{(2)} &= \tfrac{1}{2}(1+a\mathsf{s})u^{(2)} - \tfrac{1}{2\omega}(1-a^2)\mathsf{s}\partial_x u^{(1)} - \tfrac{1}{\omega}(\tfrac{1}{\omega} - \tfrac{1}{2})(1-a^2)a\mathsf{s}\partial_x^2 u^{(0)} \end{aligned}\right\} \tag{5.25}$$

by exploiting the lattice-Boltzmann update rule (5.4). However, the full lattice-Boltzmann algorithm consists of (5.4) *and* the initialization condition (5.5)

$$\mathsf{F}(0,x) = \mathsf{E}(v_0(x)). \tag{5.26}$$

In order to ensure that our prediction function satisfies *all* the constraints imposed on the numerical solution as accurately as possible, we use (5.26) in exactly the same

[10] containing only (spatial) derivatives of the mass moments of equal and lower order

way as the update rule: we replace F in (5.26) by $\mathsf{f}^{[\alpha]}$ and compare h-coefficients. Due to the simplicity of the condition, we quickly find

$$\left.\begin{aligned} h^0: &\quad \mathsf{f}^{(0)}(0,x) = \tfrac{1}{2}(1+a\mathsf{s})v_0(x) \\ h^1: &\quad \mathsf{f}^{(1)}(0,x) = 0 \\ h^2: &\quad \mathsf{f}^{(2)}(0,x) = 0 \end{aligned}\right\} \tag{5.27}$$

Comparing (5.25_0) and (5.27_0), we see that the initialization constraint forces the initial condition $u^{(0)}(0,x) = v_0(x)$ for the leading order mass moment. In view of the fact that $u^{(0)}$ also satisfies (5.16), we conclude that $u^{(0)}$ and thus also $\mathsf{f}^{(0)}$ is now *completely* determined.
In the next order, we similarly find $u^{(1)}(0,x) = 0$ but also $\partial_x u^{(0)}(0,x) = 0$. The latter condition is clearly violated if the function v_0 is not constant because $u^{(0)}(0,x) = v_0(x)$ as a consequence of the previous consideration. In other words, for the lattice-Boltzmann algorithm with initialization (5.26), our prediction function can only describe the numerical solution if v_0 is restricted to the boring case of constant functions. The reason for this behavior can be understood as follows: even if we set $u^{(1)}(0,x) = 0$, the evolution equation requires $\mathsf{f}^{(1)}(t,x) = -\frac{1}{2\omega}(1-a^2)\mathsf{s}\partial_x u^{(0)}(t,x)$ for all $(t,x) \in [0,T] \times \mathbb{R}$ whereas the initialization forces $\mathsf{f}^{(1)}(0,x) = 0$. These two conditions amount to a discontinuity of $\mathsf{f}^{(1)}$ at $t = 0$ unless v_0 is constant. In particular, a *smooth* prediction function as used in the expansion of the update rule, is incapable of describing this situation properly. Since smoothness has been an important ingredient in the derivation (Taylor's theorem), we conclude that our prediction function may properly describe the leading order h-dependence of the numerical solution but fails to describe the first order contribution.
Practically, we can observe the anticipated non-smoothness in first order as a consequence of initialization (5.26). An example is provided by figure 5.1.2.

As a remedy, we propose a modified initialization rule which avoids non-smoothness in the leading orders. We set[11]

$$\mathsf{F}(0,x) = \mathsf{E}(v_0(x)) - h\tfrac{1}{2\omega}(1-a^2)\mathsf{s}\partial_x v_0(x) - h^2\tfrac{1}{\omega}(\tfrac{1}{\omega} - \tfrac{1}{2})(1-a^2)a\mathsf{s}\partial_x^2 v_0(x). \tag{5.28}$$

Analyzing the lattice-Boltzmann algorithm with this initialization rule, we find

$$\left.\begin{aligned} h^0: &\quad \mathsf{f}^{(0)}(0,x) = \tfrac{1}{2}(1+a\mathsf{s})v_0(x) \\ h^1: &\quad \mathsf{f}^{(1)}(0,x) = -\tfrac{1}{2\omega}(1-a^2)\mathsf{s}\partial_x v_0(x) \\ h^2: &\quad \mathsf{f}^{(2)}(0,x) = -\tfrac{1}{\omega}(\tfrac{1}{\omega} - \tfrac{1}{2})(1-a^2)a\mathsf{s}\partial_x^2 v_0(x) \end{aligned}\right\} \tag{5.29}$$

which is compatible with (5.25) if $u^{(0)}(0,x) = v_0(x)$ and $u^{(1)}(0,x) = u^{(2)}(0,x) = 0$. In summary, our prediction function $\mathsf{f}^{[2]}$ satisfies the *full* lattice-Boltzmann algorithm quite accurately, provided the asymptotic order functions are based on the

[11] This idea can be generalized by setting $\mathsf{F}(0,x) = \mathsf{f}^{[\alpha]}(0,x)$ where all mass moments $u^{(\beta)}$ with $1 \le \beta \le \alpha$ are assumed to vanish initially and $u^{(0)}(0,x)$ is assumed to be $v_0(x)$. In this case, the initial condition only depends on v_0 and its derivatives. By construction, the initialization is compatible with the prediction $f^{[\alpha]}$ so that non-smoothness may only appear in order $\alpha + 1$. From a practical point of view, the initial condition enforces $\langle \mathsf{F}(0,x), 1 \rangle = v_0(x)$ so that v_0 has the interpretation as initial value for the mass moment of F.

mass moments satisfying the subsequent initial value problems (5.30), (5.31) and (5.32). For clarity we define the abbreviations:

$$\mu := (\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2), \qquad \lambda := 2a(\tfrac{1}{\omega^2} - \tfrac{1}{\omega} + \tfrac{1}{6})(1 - a^2).$$

IVP for $u^{(0)}$:

$$\boxed{\begin{aligned} &\partial_t u^{(0)} + a\partial_x u^{(0)} = 0 \\ &u^{(0)}(0, x) = v_0(x) \end{aligned}} \tag{5.30}$$

IVP for $u^{(1)}$:

$$\boxed{\begin{aligned} &\partial_t u^{(1)} + a\partial_x u^{(1)} = \mu \partial_x^2 u^{(0)} \\ &u^{(1)}(0, x) = 0 \end{aligned}} \tag{5.31}$$

IVP for $u^{(2)}$:

$$\boxed{\begin{aligned} &\partial_t u^{(2)} + a\partial_x u^{(2)} = \mu \partial_x^2 u^{(1)} + \lambda \partial_x^3 u^{(0)} \\ &u^{(2)}(0, x) = 0 \end{aligned}} \tag{5.32}$$

We remark that $u^{(0)}$, $u^{(1)}$, $u^{(2)}$ are automatically 1-periodic if v_0 has this property. Furthermore these equations determine $u^{(0)}$, $u^{(1)}$ and $u^{(2)}$ unambiguously.

Consistency

We have seen that the prediction function $\mathsf{f}^{[2]}$ approximately satisfies the lattice-Boltzmann algorithm (5.4) with initialization (5.28) if the asymptotic order functions and their mass moments are chosen according to (5.25) and (5.30), (5.31), (5.32). We can formally[12] conclude that $\mathsf{f}^{[2]}$ captures the h-behavior of F up to the expanded order, i.e. $\mathsf{F}(t, x) - \mathsf{f}^{[2]}(t, x) = \mathrm{O}(h^3)$. Summing over the components, we find for the mass moment at every grid point (t, x)

$$U(t, x) = u^{(0)}(t, x) + hu^{(1)}(t, x) + h^2 u^{(2)}(t, x) + \mathrm{O}(h^3). \tag{5.33}$$

In particular, U coincides with the solution $u^{(0)}$ of (5.30) up to an error which is at least proportional to h. In this sense, the advective lattice-Boltzmann algorithm is *consistent* to the advection equation (5.30). The order of consistency can also be deduced from (5.33). If $\omega \neq 2$ and $a^2 \neq 1$ and hence $\mu \neq 0$, equation (5.31) for $u^{(1)}$ involves a non-zero source term (unless v_0 is constant which would imply $\partial_x^2 u^{(0)} = 0$). Thus $u^{(1)}$ will be different from zero and the coincidence of U and $u^{(0)}$ is of first order

$$U(t, x) - u^{(0)}(t, x) = hu^{(1)}(t, x) + \mathrm{O}(h^2).$$

We say that the algorithm is *first order consistent* to (5.30) in that case. If, however, $\omega = 2$ or $a^2 = 1$, the source term in (5.31) vanishes and since $u^{(1)}(0, x) = 0$, the

[12] The conclusion can be rigorously justified provided the lattice-Boltzmann evolution operator can be shown to be *stable* which essentially requires all powers of the evolution matrix to be uniformly bounded in a suitable norm.

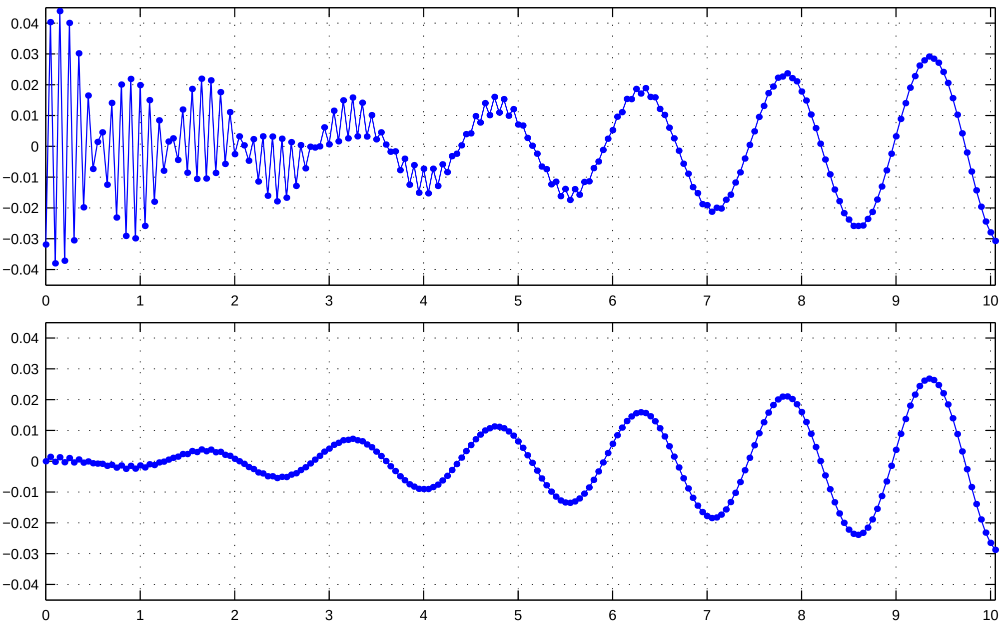


Figure 5.1:
The curves represent the quantity $\mathsf{F}_1(t,x) - \mathsf{f}_1^{[2]}(t,x)$ versus the time t in a fixed grid point x over 200 iterations on a mesh with 20 nodes. The relaxation frequency is set to $\omega = 1.97$. In the upper plot, the lattice-Boltzmann algorithm has been initialized using the equilibrium with $v_0(x) = \cos(2\pi x)$. The lower curve refers to an improved, consistent initialization by means of the regular expansion, i.e. $\mathsf{F}(0,\cdot) = \mathsf{f}^{[2]}(0,\cdot)$. The deviation between F and $\mathsf{f}^{[2]}$ starts with 0 and increases quite slowly in time. Furthermore we observe a smoothly oscillating behavior which is explained by the advection of the cosine ($a = 0.66$). In contrast, the upper curve seems to be a superposition of the lower one with a strongly oscillating (from iteration to iteration) initial deviation due to the inconsistent initialization violating the smoothness assumptions concerning the asymptotic order functions and their mass moments. The initial layer ebbs down and disappears after around 140 iterations.

solution $u^{(1)}(t,x)$ turns out to be zero everywhere. In this case,

$$U(t,x) - u^{(0)}(t,x) = h^2 u^{(2)}(t,x) + \mathrm{O}(h^3)$$

where $u^{(2)}$ is non-zero for non-trivial $u^{(0)}$ and $a^2 \notin \{0,1\}$ (note that for $\omega = 2$, the factor $\lambda = 2a(1-a^2)(\frac{1}{\omega^2} - \frac{1}{\omega} + \frac{1}{6})$ equals $-\frac{1}{6}a(1-a^2)$). Hence, the lattice-Boltzmann method is a second order accurate method for (5.30) in the case $\omega = 2$.
For $a^2 = 1$, the $u^{(2)}$-equation has zero source and zero initial value so that $u^{(2)} = 0$ in that case. Then, the algorithm is at least third order accurate[13].
We conclude that the proposed lattice-Boltzmann algorithm gives rise to approximate solutions of the advection problem (5.30) with first, second, or higher order of consistency, depending on the chosen parameters.
In exactly the same way, other lattice-Boltzmann algorithms [34, 36, 38, 39, 54, 5] or, more generally, any finite difference scheme [37] can be analyzed.

Explicit computation of the mass moments

Due to the simplicity of the scenario considered here, we are able to solve the equations for the mass moments analytically. This step is actually not necessary for the consistency analysis which can be carried out just based on the homogeneous or non-homogeneous structure of the equations as demonstrated above. However, the explicit solution reveals some interesting structure which naturally leads to a multiscale analysis of the algorithm. The equations (5.30), (5.31), and (5.32) are advection problems which can be solved with the method of characteristics.

To explain the idea, let us consider a problem of the general form

$$\partial_t w + a\partial_x w = b, \qquad w(0,x) = w_0(x). \tag{5.34}$$

The equation states that the directional derivative of w along the constant vector $(1,a)$ in (t,x) space is given by the function b. Considering a characteristic line $\sigma \mapsto Y(\sigma)$ starting in the point $(0,x_0)$ with direction vector $(1,a)$, i.e.

$$\dot{Y}(\sigma) = (1,a), \quad Y(0) = (0,x_0) \qquad \Rightarrow \qquad Y(\sigma) = (\sigma, x_0 + \sigma a),$$

we know that the value $z(\sigma)$ of the solution to (5.34) evolves along the line according to the law

$$\dot{z}(\sigma) = b(Y(\sigma)), \qquad z(0) = w_0(x_0).$$

In particular, the solution w at a given point (t,x) is obtained by determining x_0 and σ such that $Y(\sigma) = (t,x)$ and setting $w(t,x) = z(\sigma)$. Specifically, we have $\sigma = t$ and $x_0 = x - at$, so that

$$w(t,x) = z(t), \qquad \dot{z}(\sigma) = b(\sigma, x - (t-\sigma)a), \qquad z(0) = w_0(x - at).$$

Applying this methodology to (5.30), we have to deal with the simple source $b^{(0)}(t,x) = 0$ so that z turns out to be constant along the characteristic line, i.e. $z(\sigma) = w_0(x - at)$ for all σ. Hence:

$$\boxed{u^{(0)}(t,x) = v_0(x - at)} \tag{5.35}$$

[13] One can show that for $a^2 = 1$, the lattice-Boltzmann algorithm even reproduces the *exact* solution of (5.30) at the grid points.

Using this result, we can compute the right hand side of (5.31)

$$b^{(1)}(t,x) = \mu v_0''(x-at), \qquad \mu = (\tfrac{1}{\omega} - \tfrac{1}{2})(1-a^2).$$

Along the characteristic line, we have

$$b^{(1)}(\sigma, x-(t-\sigma)a)) = \mu v_0''(x-(t-\sigma)a - a\sigma) = \mu v_0''(x-at)$$

which is independent of σ. Hence, the function value z grows with constant rate $v_0''(x-at)$ along the characteristic, starting with $u^{(1)}(0, x-at) = 0$ so that:

$$\boxed{u^{(1)}(t,x) = \mu t v_0''(x-at)} \tag{5.36}$$

The source in (5.32) has the form

$$b^{(2)}(t,x) = \mu\partial_x^2 u^{(1)}(t,x) + \lambda\partial_x^3 u^{(0)}(t,x), \qquad \lambda = 2a(\tfrac{1}{\omega^2} - \tfrac{1}{\omega} + \tfrac{1}{6})(1-a^2).$$

Similarly to the previous case, the behavior is simple along the characteristics with the only difference, that the linear growth of $u^{(1)}$ is integrated to a quadratic behavior. More precisely,

$$b^{(2)}(\sigma, x-(t-\sigma)a)) = \mu^2\sigma v_0''''(x-at) + \lambda v_0'''(x-at)$$

and thus:

$$\boxed{u^{(2)}(t,x) = \tfrac{1}{2}\mu^2 t^2 v_0''''(x-at) + \lambda t v_0'''(x-at)} \tag{5.37}$$

Using this explicit form, we can write

$$U(t,x) = v_0(y) + (ht)\mu v_0''(y) + \tfrac{1}{2}(ht)^2\mu^2 v_0''''(y) + (h^2t)\lambda v_0'''(y) + \mathrm{O}(h^3)\big|_{y=x-at}.$$

An important observation is that the expansion *breaks down* if t is chosen large. For example, if $t = 1/h$, the formerly first and second order terms obviously turn into zero order contributions and we may expect that higher order terms which we collected in $\mathrm{O}(h^3)$ also need to be considered since they may be of the form $(ht)^3$, $(ht)^4$, etc. In other words, the regular expansion fails to represent the leading order behavior for large t since it is non-uniformly bounded in the time variable (it contains so-called *secular*[14] terms). Hence, if we are interested in the long time behavior of our scheme, we have to choose a different asymptotic ansatz. Such an ansatz can be motivated by restructuring the expansion of U

$$U(t,x) = v_0(y) \;+\; \left[\mu v_0''(y) + h\lambda v_0'''(y)\right] ht \;+\; \tfrac{1}{2}\mu^2 v_0''''(y)\,(ht)^2 \;+\; \cdots\big|_{y=x-at}.$$

In this form U appears as some sort of truncated power series expansion in the variable (ht) where the coefficients themselves are regularly expanded with respect to h and depend on t. Clearly, this truncation is reasonable if ht is small but fails

[14] Derived from the Latin word *saeculum* = century. The term originates from celestial mechanics, more exactly from asymptotic expansions describing the motion of planets, and refers to long term perturbations. In our context we call terms in an asymptotic expansion secular (with respect to the time variable) if they are not uniformly bounded over $\mathbb{R}^+$ but only over compact intervals. However there is no strict definition.

for $ht > 1$. The way out is to *avoid* expanding the distinct ht behavior with respect to h. Instead, the dependence of the numerical solution on the quantity ht should be resolved directly. This can be achieved with an ansatz of the form

$$U(t,x) = u^{(0)}(t,ht,x) + hu^{(1)}(t,ht,x) + h^2u^{(2)}(t,ht,x) + \dots$$

which is a so called two-scale expansion. An additional motivation for this approach can be drawn from direct observations of the numerical solution. This is explained in the following section.

5.2 Multiscale Expansion

5.2.1 A numeric test to detect different time scales

Consider the lattice-Boltzmann algorithm on two different spatial grids $\mathcal{G}_1 \subset \mathcal{G}_2$ with h_1 being a multiple of h_2. Both algorithms are supposed to be equally initialized by means of a smooth function v_0 representing the initial mass moment. Let $U_1(n)$ and $U_2(n)$ denote the mass moments on the coarse and fine grid after n iterations. For a moment, we want to ignore the analysis done in section 5.1. Then it is not clear *a priori*, how U_1 and U_2 are related to each other. To be less sketchy we may ask, for which indices n_1, n_2 the mass moments $U_1(n_1)$ and $U_2(n_2)$ look especially similar.

We attempt to answer the question by an experiment. If the algorithm is initialized choosing a peak-like profile for v_0, one might discover[15] that the peaks almost coincide with each other if the iteration indices satisfy $n_1h_1 = n_2h_2$. More generally, we can say

$$U_1(n_1) \approx U_2(n_2) \qquad \text{whenever } n_1h_1 = n_2h_2. \tag{5.38}$$

Therefore it is reasonable to introduce a new quantity T_1 defined by $T_1(n_i, \mathcal{G}_i) := n_ih_i$ called *time*. Obviously, the evolution of the algorithm reveals a grid independent dynamics, if iterations on different grids are identified with respect to their time.

The strong resemblance (quasi-agreement) of the mass moments for $T_1(n_1, \mathcal{G}_1) = T_1(n_2, \mathcal{G}_2)$ is explained by the consistency analysis. This tells, that for $h \to 0$ the mass moments converge towards the solution of the advection equation (initialized by v_0), which, of course, does not know anything about the lattice-Boltzmann algorithm and has a grid independent evolution.

If the two algorithms run over long periods with $T_1 \gg 1$, we realize that the peaks start to deform their shape (see figure 5.2). Although this phenomenon is predicted by the regular expansion (specifying the leading error terms $u^{(1)}, u^{(2)}$) the forecast becomes totally wrong after a relatively short while. Furthermore we observe that the flattening of the peak is more distinct on the coarser grid $\mathcal{G}_1$ than on $\mathcal{G}_2$, in spite of comparing iterations corresponding to the same time. This insinuates that the deformation process takes place in a time scale different from that of the advection. In order to study the deformation alone without being "disturbed" by the advection it is possible to switch it off by setting $a = 0$. It turns out that the peaks

[15] The effect becomes the more evident the smaller h_1 and h_2 are chosen.

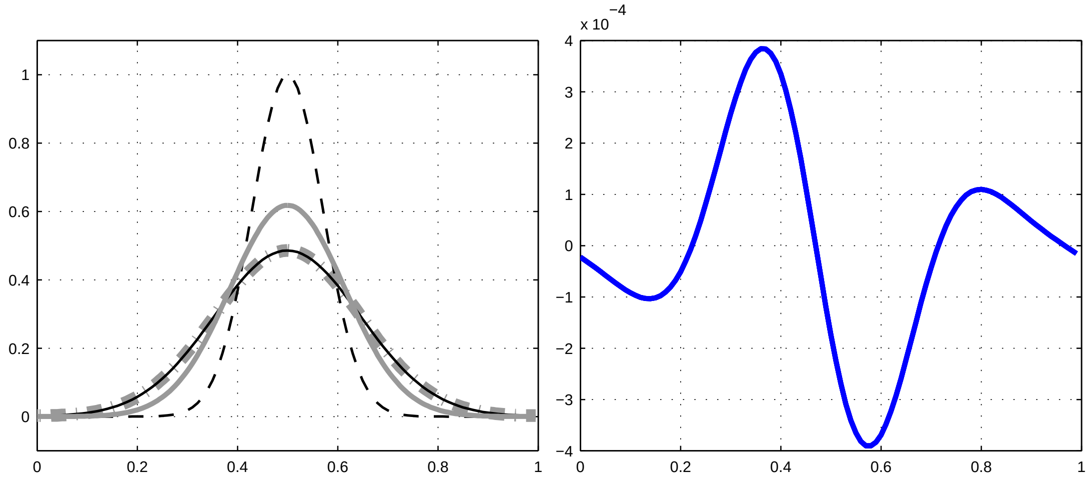


Figure 5.2:
Left: Advective(linear) versus diffusive(quadratic) time scale. Generally, the flattening of the mass moment increases with the number of iterations. Here we have set $\omega = 1.3$. The advection velocity $a = 0.5$ has been chosen in such a way that the peak returns to its initial position at the center after 200(400) iterations on a grid with 100(200) nodes. So we can blind out the displacement by the advection avoiding to switch it off totally. The subsequent table indicates to which grid and to which iteration the lines (representing the mass moment) refer.

line	#grid nodes	#time steps
gray, dash-dotted	100 ($\mathcal{G}_1$)	400
gray, solid	200 ($\mathcal{G}_2$)	800
black solid	200 ($\mathcal{G}_2$)	1600

The fact that the black solid curve agrees with the gray dash-dotted curve suggests that the flattening occurs in the *quadratic time scale*. The black dashed peak corresponds to the initial condition. Right: Difference between the dash-dotted and the thin black solid line (projected onto the coarse grid).

undergo the same deformation on both grids, if the iteration indices fulfill the condition $n_1 h_1^2 = n_2 h_2^2$ (cf. figure 5.2). Similarly to T_1 we therefore define another time T_2 with $T_2(n, \mathcal{G}) := nh^2$, which elapses slower than T_1 provided $h < 1$. In order to distinguish both of the times we refer to T_1 as *linear time* and designate T_2 as *quadratic time*. Moreover we also speak of *time scales* in this context.

A special situation becomes noticeable for $\omega = 2$. Instead of a simple flattening, the deformation of the mass moment displays a distortion due to oscillations (cf. figure 5.3). It turns out, that these oscillations evolve (emerge and modify themselves) even slower than the quadratic time. This urges the conjecture that their evolution can be described by means of the cubic time scale $T_3(n, \mathcal{G}) = nh^3$. This hypothesis is backed by figure 5.3.

Recapitulating we can say that the the evolution of the lattice-Boltzmann algorithm appears as a superposition of several[16] processes. Each one is governed by a time scale of its own and can be interpreted as a grid independent process if regarded in this scale. In the next subsection we apply the technique of multiscale expansions to separate these processes from each other and to derive equations determining their evolution.

5.2.2 Additional quadratic time scale

The previous subsection gives rise to a modification of the ansatz for the prediction function. In order to capture both the advection in the linear time scale and the flattening deformation in the quadratic time scale we equip the new prediction function with two time arguments:

$$\mathsf{f}^{[\alpha]}(t_1, t_2, x) := \mathsf{f}^{(0)}(t_1, t_2, x) \;+\; h\mathsf{f}^{(1)}(t_1, t_2, x) \;+\; \dots \;+\; h^{\alpha}\mathsf{f}^{(\alpha)}(t_1, t_2, x), \tag{5.39}$$

Inspired by the observations, we want to choose $\mathsf{f}^{[\alpha]}$ in such a way that we have

$$\mathsf{F}(t, x) \approx \mathsf{f}^{[\alpha]}(t, ht, x) \tag{5.40}$$

for $t = t_n = nh$ and $x = x_j = jh$ with $n \in \mathbb{N}_0$ and $j \in \mathbb{Z}$. In particular, $\mathsf{f}^{[\alpha]}$ shall satisfy the update rule (5.2) of the lattice-Boltzmann algorithm with a small residual.

$$\mathsf{f}_k^{[\alpha]}(t + h, ht + h^2, x + s_k h) = \mathsf{f}_k^{[\alpha]}(t, ht, x) + \left[J\mathsf{f}^{[\alpha]}(t, ht, x)\right]_k + \mathrm{O}(h^{\alpha+1}) \tag{5.41}$$

We suppose that $\mathsf{f}^{[\alpha]}$ and its asymptotic order functions fulfill the same assumptions concerning the regularity, 1-periodicity etc. as in subsection 5.1.1. Next, we substitute (5.39) into (5.41). In order to find determining conditions for the $\mathsf{f}^{(\beta)}$'s, we perform on the left hand side a Taylor expansion of each $\mathsf{f}^{(\beta)}$ around (t, ht, x). More precisely, the formal Taylor series

$$\left[\mathsf{f}_k^{(\beta)}(t + h, ht + h^2, x + s_k h)\right]_k = \sum_{\gamma \geq 0} \frac{1}{\gamma!}(hD + h^2\partial_{t_2})^{\gamma}\, \mathsf{f}^{(\beta)}(t, ht, x)$$

is truncated after the $(\alpha - \beta)^{\text{th}}$ order. Terms being of h-order greater than α (including the remainder terms of the Taylor expansion) are incorporated into the $\mathrm{O}(h^{\alpha+1})$-term. There is now some liberty to decide how the terms shall cancel each other. A reasonable way[17] is to equate those terms at the left and right hand

[16] Here we could also broach the issue of *initial layers* representing a typical example of superposed processes. Numerical initial layers are often triggered by crude initializations and evolve on a faster time scale than the dominant one. In particular, they evolve with respect to the *discrete time scale* $(T_0(n, \mathcal{G}) = n)$.

[17] We proceed as if we could perform the comparison of coefficients. But actually, the $\mathsf{f}^{(\beta)}(t, ht, x)$'s at the left hand side are not completely expanded with respect to h. Therefore the equality of terms containing the same power of h is not compellent if h is varied. Hence we are not lead automatically to equations that determine the asymptotic order functions as in the case of the regular expansion. Other settings are imaginable. This insinuates that multiscale expansions are not unique in contrast to regular expansions. To illustrate this by a trivial example consider

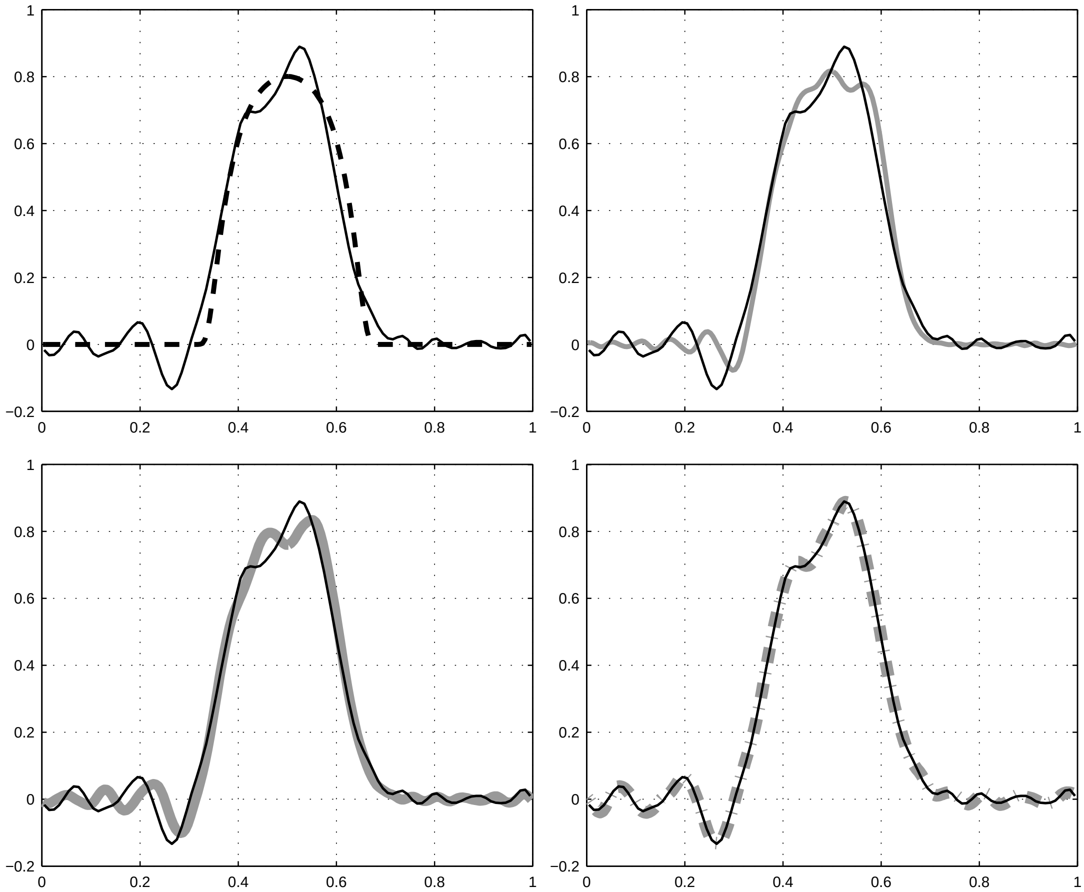


Figure 5.3:

Advective(linear) versus diffusive(quadratic) and dispersive(cubic) time scale: The plots display the mass moment on two different grids at different iterations (see table below). For clarity each diagram shows two curves only. The initial mass moment is given by the dashed line in the upper left subfigure. The flanks of this curve are quite steep near the transition points where the flat zone passes over to the peak. Although it seems to have kinks, the function is everywhere arbitrarily smooth. As this feature provokes relatively strong oscillations, it is well suited to highlight the effect. Observe that the curves in the subfigure right below match so well, that they appear virtually identical. As $3200 = 2^3 \cdot 200$, there are 8 times more iterations necessary to reach the same configuration of oscillations on a grid with half the spacing size. This prompts us to assume that the evolution of the oscillations becomes grid-independent if considered in the cubic time scale.

line	#grid nodes	#time steps
black solid	100	400
gray solid	200	800
gray solid (thick)	200	1600
gray dash-dotted	200	3200

Analogously to the example in figure 5.2 the advection velocity is set to $a = 0.5$, while we have selected the special value $\omega = 2$ for the relaxation frequency.

side which carry the same power of h as prefactor. Then we obtain the following equations from the $0^{\text{th}}, 1^{\text{st}}, 2^{\text{nd}}$ and 3^{rd} order:

$$\left.\begin{array}{lll} h^0: & \mathsf{f}^{(0)} & = (I+J)\mathsf{f}^{(0)} \\ h^1: & \mathsf{f}^{(1)} + D\mathsf{f}^{(0)} & = (I+J)\mathsf{f}^{(1)} \\ h^2: & \mathsf{f}^{(2)} + D\mathsf{f}^{(1)} + \frac{1}{2}D^2\mathsf{f}^{(0)} + \partial_{t_2}\mathsf{f}^{(0)} & = (I+J)\mathsf{f}^{(2)} \\ h^3: & \mathsf{f}^{(3)} + D\mathsf{f}^{(2)} + \frac{1}{2}D^2\mathsf{f}^{(1)} + \frac{1}{6}D^3\mathsf{f}^{(0)} + \partial_{t_2}\mathsf{f}^{(1)} + D\partial_{t_2}\mathsf{f}^{(0)} & = (I+J)\mathsf{f}^{(3)} \end{array}\right\} \quad (5.42)$$

Subtracting $\mathsf{f}^{(\beta)}$ on each side of (5.42_β) results in an equation for $\mathsf{f}^{(\beta)}$ with the system matrix $J = \omega(E - I)$. So the equations (5.42) correspond to (5.10). In particular we obtain the same equations in the 0^{th} and 1^{st} order; only the equations from the 2^{nd} order on are affected by the modified ansatz, producing additional inhomogeneities. As the structure of the equations stays unaltered, we can take over the results from subsection 5.1.1 and only track the additional terms.

Introducing the mass moments $u^{(0)}, u^{(1)}, u^{(2)}$ and taking the scalar product of equations (5.42) with the vector $\mathbf{1}$ finally gives the subsequent equations[18]:

$$\left.\begin{array}{lll} (5.42_1) & \Rightarrow & \partial_{t_1}u^{(0)} + a\partial_x u^{(0)} = 0 \\ (5.42_2) & \Rightarrow & \partial_{t_1}u^{(1)} + a\partial_x u^{(1)} = \mu\partial_x^2 u^{(0)} - \boldsymbol{\partial_{t_2} u^{(0)}} \\ (5.42_3) & \Rightarrow & \partial_{t_1}u^{(2)} + a\partial_x u^{(2)} = \mu\boldsymbol{\partial_x^2 u^{(1)}} + \lambda\partial_x^3 u^{(0)} - \boldsymbol{\partial_{t_2} u^{(1)}} \end{array}\right\} \quad (5.43)$$

In subsection 5.1.2 we have seen that the integration of the evolution equations for $u^{(1)}$ and $u^{(2)}$ lead to *secular* terms, that do not reflect the behavior of the algorithm on a long time scale. These terms are generated by the sources. A particular reason, why secular terms show up, is that the sources are themselves solutions of (inhomogeneous) advection equations with equal advection velocity causing the unbounded factors t and t^2.
A view at (5.43_2) shows that we should encounter the same problem unless we require the right hand side of (5.43_2) to vanish. In fact, by this setting[19], we kill two birds with one stone: Firstly, we avoid the arising of secular terms. Secondly, we find a well known homogeneous diffusion equation, that determines the evolution of $u^{(0)}$ with respect to t_2.

the function $f(n) = nh + nh^2 + n^2h^5$. Introducing the variables $t_1 = nh$ and $t_2 = nh^2$ there is an infinite multitude of possibilities. As the replacement of the third addend is not obvious we can express it using either t_1 or t_2 or both of them by writing it in the from $n^2h^5 = qn^2h^5 + (1-q)n^2h^5$ for some $q \in \mathbb{R}$. Hence we obtain the following parametrized family of representations
$$f_q(t_1, t_2) = (t_1 + t_2) + hqt_2^2 + h^3(1-q)t_1^3,$$
all of them satisfying $f(n) = f_q(nh, nh^2)$.

[18] Observe that the scalar product $\langle D\partial_{t_2}\mathsf{f}^{(0)}, \mathbf{1}\rangle = \partial_{t_2}\langle D\mathsf{f}^{(0)}, \mathbf{1}\rangle = \partial_{t_2}\big(\underbrace{\partial_{t_1}u^{(0)} + a\partial_x u^{(0)}}_{=0}\big) = 0$ does not contribute.

[19] Notice that we only give a motivation and no strict justification for our approach, which has the appeal to break up a complicated, unconventional system into standard partial differential equations – similarly to the separation of variables method decomposing the Laplace equation into ODEs. To reassure a reader being alarmed about this frivolity, we recall that asymptotics often requires some ingenuity to select a good ansatz. A rigorous proof is mostly based on a stability analysis and rarely reflects the idea why a certain ansatz has been chosen. Furthermore it should be kept in mind that multiscale expansions need not be unique.

Analogously we proceed with the next equation (5.43_3) which is decoupled into a homogeneous advection equation for $u^{(2)}$ and a diffusion equation for $u^{(1)}$. This time, the source term containing $u^{(0)}$ drives the diffusion equation for $u^{(1)}$ instead of the advection equation for $u^{(2)}$. We will see that this prevents the appearance of secular terms.

To summarize the reasoning, let us write down the initial value problems that determine $u^{(0)}$ and $u^{(1)}$, where the initial conditions result from the requirement $v_0(x) = \langle \mathsf{F}(0,x), 1\rangle = \langle \mathsf{f}^{[\alpha]}(0,0,x), 1\rangle$:

$$\boxed{\begin{array}{rcrcl} \partial_{t_1} u^{(0)}(t_1,t_2,x) & + & a\partial_x u^{(0)}(t_1,t_2,x) & = & 0 \\ \partial_{t_2} u^{(0)}(t_1,t_2,x) & - & \mu\partial_x^2 u^{(0)}(t_1,t_2,x) & = & 0 \\ \multicolumn{5}{c}{u^{(0)}(0,0,x) = v_0(x)} \end{array}} \tag{5.44}$$

$$\boxed{\begin{array}{rcrcl} \partial_{t_1} u^{(1)}(t_1,t_2,x) & + & a\partial_x u^{(1)}(t_1,t_2,x) & = & 0 \\ \partial_{t_2} u^{(1)}(t_1,t_2,x) & - & \mu\partial_x^2 u^{(1)}(t_1,t_2,x) & = & \lambda\partial_x^3 u^{(0)}(t_1,t_2,x) \\ \multicolumn{5}{c}{u^{(1)}(0,0,x) = 0} \end{array}} \tag{5.45}$$

A reason explaining the flattening observed in figure 5.2 emerges from the occurrence of the diffusion equation in (5.44) and (5.45). Indeed, the smoothing of sharp extrema and the leveling of steep slopes ranks among the characteristic dynamics encoded in this equation and distinguishes it fundamentally from the advection equation. A necessary condition to display this behavior is the positivity of the *diffusivity coefficient* $\mu = (\frac{1}{\omega} - \frac{1}{2})(1-a^2)$ determining how fast the smoothing process takes place. For $\mu < 0$ the smoothing behavior is reversed. As the resulting IVP for the backward diffusion equation is ill-posed, it might not admit a solution at all. This 'bad' behavior may be reflected by the lattice-Boltzmann algorithm revealing instabilities. In allusion to physical diffusion processes the artificial smoothing effect of numerical schemes is referred to as *numeric diffusion*[20].

For each mass moment we interpret the equation systems (5.44) and (5.45) as two initial value problems that are coupled via a common initial condition and the spatial variable x. The solution can be obtained consecutively starting with the diffusion equation.

In order to acquire some feeling, how possible solutions of (5.44) and (5.45) are constructed and look like, let us consider an example. Since we are primarily interested in 1-periodic solutions we choose a cosine[21] as initial mass moment:

$$v_0(0,0,x) = \cos(2\pi m x) \qquad m \in \mathbb{N}.$$

[20]Numeric diffusion may be regarded as an undesired and annoying property. However, in many situations it has a stabilizing side action.

[21]Actually, this example is of much importance, since $\cos(2\pi m x)$ represents the general Fourier mode. Thanks to the linearity of both equations, the solution for an arbitrary 1-periodic initial function v_0 can be obtained by decomposing it into its Fourier modes. Solving the equations for each Fourier mode separately, the solution pertaining to the original initial value is recovered by superposition.

Solving the diffusion equation at first, yields:

$$u^{(0)}(0,t_2,x) = \exp\big(-4\pi^2 m^2 \mu t_2\big)\,\cos\big(2\pi m x\big)$$

Initializing the advection equation by $u^{(0)}(0,t_2,x)$, where t_2 is to be treated as mere parameter, results in the final solution:

$$\boxed{u^{(0)}(t_1,t_2,x) = \exp\big(-4\pi^2 m^2 \mu t_2\big)\,\cos\big(2\pi m(x-at_1)\big)} \tag{5.46}$$

The diffusion equation for $u^{(1)}$ is inhomogeneous. The source itself is a solution of a homogeneous diffusion equation with equal diffusivity μ. Therefore $u^{(1)}(0,t_2,x)$ is just obtained by multiplying $\lambda\partial_x^3 u^{(0)}(0,t_2,x)$ with t_2:

$$u^{(1)}(0,t_2,x) = 8\pi^3 m^3 \lambda\,\boldsymbol{t_2}\,\exp\big(-4\pi^2 m^2 \mu t_2\big)\,\sin\big(2\pi m x\big)$$

Substituting x by $x - at_1$

$$\boxed{u^{(1)}(t_1,t_2,x) = 8\pi^3 m^3 \lambda\,\boldsymbol{t_2}\,\exp\big(-4\pi^2 m^2 \mu t_2\big)\,\sin\big(2\pi m(x-at_1)\big)} \tag{5.47}$$

makes $u^{(1)}$ satisfy all three equations in (5.45).

Similar to (5.36), we notice a factor t_2 in the explicit representation of $u^{(1)}$. However, in opposition to the regular expansion, the factor t_2 is accompanied by the exponential function $\exp(-4\pi^2 m^2 \mu t_2)$ damping $u^{(1)}$ to zero while t_2 grows towards infinity, if the diffusion coefficient μ is positive. So, under this condition, $u^{(1)}$ is uniformly bounded, i.e. there are upper and lower bounds for all $t \geq 0$.

The situation in the first order is characteristic for higher orders, because the structure of the evolution equations

$$\begin{aligned}
\partial_{t_1} u^{(\beta)}(t_1,t_2,x) \;&+\; a\partial_x u^{(\beta)}(t_1,t_2,x) \;=\; 0\\
\partial_{t_2} u^{(\beta)}(t_1,t_2,x) \;&-\; \mu\partial_x^2 u^{(\beta)}(t_1,t_2,x) \;=\; \text{inhomogeneity}\\
&u^{(\beta)}(0,0,x) = 0
\end{aligned}$$

is repeated for all $\beta \in \mathbb{N}$. The only difference is given by the inhomogeneities that become more complicated for higher orders. This entails particularly that t_2^β occurs as factor in the explicit formula for $u^{(\beta)}$. However, also higher powers of t_2 are tamed by the exponential function which decays more rapidly than any polynomial can increase. So $t_2^\beta \exp(-4\pi^2 m^2 \mu h t_2)$ is uniformly bounded for any β and vanishes at infinity ($t \uparrow \infty$). Therefore, we are allowed to conclude that the two-scale expansion avoids secular terms.

The computations give rise to the hypothesis that the mass moment $U = \langle \mathsf{F}, 1\rangle$ admits the representation

$$U(t,x) = u^{(0)}(t,ht,x) + hu^{(1)}(t,ht,x) \;+\; \text{H.O.T.}$$

where H.O.T. stands for ‘higher order terms’ and

$$u^{(0)}(t,ht,x)+hu^{(1)}(t,ht,x)=$$
$$=\Big[\cos\big(2\pi m(x-at)\big)+h^2t\,8\pi^3m^3\lambda\sin\big(2\pi m(x-at)\big)\Big]\exp(-4\pi^2m^2\mu ht).$$

Note that the factor h^2 in front of the sine comes from the first order and the substitution $t_2=ht$. Under the assumption that the higher order terms are well-behaved, we infer:

- $u^{(0)}(t,ht,x)$ describes the behavior of $U(t,x)$ with 2^{nd} order accuracy in h on time intervals whose length is of magnitude $\mathrm{O}(1)$. This means that the maximal iteration index $n_{\max}$ satisfies the estimate $n_{\max}h<C$, where h denotes the grid spacing and C is a given constant.
- $u^{(0)}(t,ht,x)$ describes the behavior of U with 1^{st} order accuracy in h on ‘long’ time intervals whose length is of magnitude $\mathrm{O}(\frac{1}{h})$, i.e. the maximal iteration index complies with the condition $n_{\max}h^2<C$.

In particular, our results for $u^{(0)}$ and $u^{(1)}$ in (5.46) and (5.47) foreshadow that U remains bounded independently of the number of iterations, stability of the algorithm provided.

Indeed, all hypotheses are positively validated by numerical experiments which justifies the two-scale expansion as an adequate tool to analyze the behavior of the lattice-Boltzmann scheme.

5.2.3 Emergence of a cubic time scale

Now let us look what happens with the multiscale expansion if we set $\omega=2$. In this case the diffusivity coefficient $\mu=(\frac{1}{\omega}-\frac{1}{2})(1-a^2)$ becomes 0. So the diffusion equation in (5.44) degenerates to $\partial_{t_2}u^{(0)}=0$ implying that $u^{(0)}$ is not dependent on t_2 at all. Taking the initial condition $u^{(0)}(0,0,x)=v_0(x)$ and the advection equation into account, this amounts to $u^{(0)}(t_1,t_2,x)=v_0(x-at_1)$. Thus $u^{(0)}$ does not capture a damping exponential factor, that could guarantee uniform boundedness of the $u^{(\beta)}$’s with respect to t_2 similarly to the example in the preceding subsection. Computing $u^{(1)}(t_1,t_2,x)=t_2\lambda v_0'''(x-at_1)$ we see that it gets unbounded for $t_2\to\infty$. Hence the two-scale expansion with the additional quadratic time scale is not able to avoid secular terms.

In subsection 5.2.1, we have noticed (cf. figure 5.3), that for $\omega=2$ oscillations emerge and modify themselves in the cubic time scale. This suggests to approximate the population function by a prediction function $\mathsf{f}^{[\alpha]}(t_1,t_3,x)$ having two time arguments as well, but differently related to the n^{th} iteration:

$$\mathsf{F}(nh,jh)\approx\mathsf{f}^{[\alpha]}(nh,nh^3,jh)\qquad\text{or}\qquad\mathsf{F}(t,x)\approx\mathsf{f}^{[\alpha]}(t,h^2t,x).$$

In order to keep this slight modification with respect to (5.40) in mind, we denote the second temporal argument by t_3. Again $\mathsf{f}^{[\alpha]}$ stands for a polynomial of order

α in h. Then, under analogous assumptions as in the previous subsection, we can derive the following equations for $\alpha \geq 4$ to determine the leading asymptotic order functions.

$$\left.\begin{aligned}
h^0 &: \mathsf{f}^{(0)} &&= (I+J)\mathsf{f}^{(0)} \\
h^1 &: \mathsf{f}^{(1)} + D\mathsf{f}^{(0)} &&= (I+J)\mathsf{f}^{(1)} \\
h^2 &: \mathsf{f}^{(2)} + D\mathsf{f}^{(1)} + \tfrac{1}{2}D^2\mathsf{f}^{(0)} &&= (I+J)\mathsf{f}^{(2)} \\
h^3 &: \mathsf{f}^{(3)} + D\mathsf{f}^{(2)} + \tfrac{1}{2}D^2\mathsf{f}^{(1)} + \tfrac{1}{6}D^3\mathsf{f}^{(0)} + \partial_{t_3}\mathsf{f}^{(0)} &&= (I+J)\mathsf{f}^{(3)} \\
h^4 &: \mathsf{f}^{(4)} + D\mathsf{f}^{(3)} + \tfrac{1}{2}D^2\mathsf{f}^{(2)} + \tfrac{1}{6}D^3\mathsf{f}^{(1)} + \tfrac{1}{24}D^4\mathsf{f}^{(0)} + \partial_{t_3}\mathsf{f}^{(1)} + D\partial_{t_3}\mathsf{f}^{(0)} &&= (I+J)\mathsf{f}^{(4)}
\end{aligned}\right\} \quad (5.48)$$

Similarly to the procedure in 5.1.1 and 5.2.2, these equations are written in terms of the mass moments associated to the asymptotic order functions

$$\left.\begin{array}{lcl}
(5.48_1) & \Rightarrow & \partial_{t_1}u^{(0)} + a\partial_x u^{(0)} = 0 \\
(5.48_2) & \Rightarrow & \partial_{t_1}u^{(1)} + a\partial_x u^{(1)} = 0 \\
(5.48_3) & \Rightarrow & \partial_{t_1}u^{(2)} + a\partial_x u^{(2)} = \tilde{\lambda}\partial_x^3 u^{(0)} - \partial_{t_3}u^{(0)} \\
(5.48_4) & \Rightarrow & \partial_{t_1}u^{(3)} + a\partial_x u^{(3)} = \tilde{\lambda}\partial_x^3 u^{(1)} - \partial_{t_3}u^{(1)}
\end{array}\right\} \quad (5.49)$$

where $\tilde{\lambda} := -\frac{1}{6}a(1-a^2)$ shortcuts the coefficient λ for $\omega = 2$. Splitting (5.49_3) into separate equations for the left and right hand side, we get a complete[22] set of equations for $u^{(0)}$.

$$\boxed{\begin{array}{ccccc}
\partial_{t_1}u^{(0)}(t_1,t_3,x) & + & a\partial_x u^{(0)}(t_1,t_3,x) & = & 0 \\
\partial_{t_3}u^{(0)}(t_1,t_3,x) & - & \tilde{\lambda}\partial_x^3 u^{(0)}(t_1,t_3,x) & = & 0 \\
\multicolumn{5}{c}{u^{(0)}(0,0,x) = v_0(x)}
\end{array}} \quad (5.50)$$

Moreover, combining (5.49_2) and (5.49_4) we find in the same way:

$$\boxed{\begin{array}{ccccc}
\partial_{t_1}u^{(1)}(t_1,t_2,x) & + & a\partial_x u^{(1)}(t_1,t_2,x) & = & 0 \\
\partial_{t_3}u^{(1)}(t_1,t_2,x) & - & \tilde{\lambda}\partial_x^3 u^{(1)}(t_1,t_3,x) & = & 0 \\
\multicolumn{5}{c}{u^{(1)}(0,0,x) = 0}
\end{array}} \quad (5.51)$$

The second equation of (5.50) states that the dynamics of $u^{(0)}$ in the slow time variable is described by a *dispersive* equation. A characteristic property of this equation is to disperse a peak into a sequence of oscillations similarly to what we have observed (cf. figure 5.3). The reason for this behavior is found if we study the evolution of a single Fourier mode. Let us therefore consider a sinusoidal initial condition:

$$v_0(x) = \cos(2\pi m x).$$

Integrating the dispersion equation in (5.50) with this initialization yields

$$u^{(0)}(0,t_3,x) = \cos\left(8\pi^3 m^3 \tilde{\lambda} t_3 + 2\pi m x\right). \quad (5.52)$$

[22]Observe that we must expand up to the third order to extract the full set of evolution equations determining $u^{(0)}$.

Differently from the diffusion equation, the sinusoidal profile is shifted and not damped. In contrast to the advection equation the transport velocity depends on the spatial frequency of the cosine. Hence the Fourier modes of an arbitrary initial profile travel with different speeds. This effect destroys the original shape of the profile, which is conserved under pure advection.
Taking (5.52) as initial condition for the advection equation in (5.50) yields the final result for $u^{(0)}$.

$$\boxed{u^{(0)}(t_1, t_3, x) = \cos\left(8\pi^3 m^3 \tilde{\lambda} t_3 + 2\pi m(x - at_1)\right)}$$

Let us now turn to (5.51) that differs from the analogous equations (5.21) and (5.45), because of the missing source term. Due to the zero initial condition, $u^{(1)}$ must vanish.

$$\boxed{u^{(1)}(t_1, t_3, x) = 0}$$

Thus, the asymptotic expansion of the mass moment has a trivial first order if $\omega = 2$, wherefore the lattice-Boltzmann scheme reaches higher accuracy as in the general case $\omega \neq 2$ discussed in subsection 5.1.2.
How does the expansion continue? In general the second order $u^{(2)}$ does not vanish. Due to the resonance effect[23] it will be of the form $t_3 w(t_1, t_3, x)$ with some function w. Resubstituting t_1 by t and t_3 by $h^2 t$ produces a factor h^4 in front of $u^{(2)}$. This indicates that $u^{(0)}$ should approximate the mass moment U up to a residue of magnitude $\mathrm{O}(h^4)$ over time intervals of constant length. This hypothesis is supported by numerical tests. The approximation order is reduced respectively by 1 and 2 if we couple the number of iterations quadratically or cubically to the inverse of the grid spacing h (time intervals of length $\mathrm{O}(h^{-2})$ or $\mathrm{O}(h^{-3})$).

To understand better how the vanishing of the first order comes about, let us close this subsection by delivering the computation in detail:

In order to obtain (5.49_3) and from this (5.51) we must take the scalar product of (5.48_4) with $\mathbf{1}$. Since

$$\langle J\mathsf{f}^{(4)}, \mathbf{1}\rangle = 0, \qquad \text{and} \qquad \langle D\partial_{t_3}\mathsf{f}^{(0)}, \mathbf{1}\rangle = \partial_{t_3}\langle D\mathsf{f}^{(0)}, \mathbf{1}\rangle = \partial_{t_3}(\partial_t u^{(0)} + a\partial_x u^{(0)}) = 0$$

we get

$$0 = \overbrace{\langle \partial_{t_3}\mathsf{f}^{(1)}, \mathbf{1}\rangle}^{=\partial_{t_3} u^{(1)}} + \overbrace{\langle D\,\mathsf{f}^{(3)}, \mathbf{1}\rangle + \tfrac{1}{2}\langle D^2\mathsf{f}^{(2)}, \mathbf{1}\rangle + \tfrac{1}{6}\langle D^3\mathsf{f}^{(1)}, \mathbf{1}\rangle + \tfrac{1}{24}\langle D^4\mathsf{f}^{(0)}, \mathbf{1}\rangle}^{=:A}.$$

[23] Generally: If A is some linear operator (e.g. spatial derivative) and $x(t,\cdot)$ a solution of the homogeneous equation $\partial_t x + Ax = 0$ then $y(t,\cdot) = tx(t,\cdot)$ is a solution of the inhomogeneous equation $\partial_t y + Ay = x$. So the solution of the equation imitates the source while amplified in time due to the multiplication with t. This phenomenon is called *resonance*. Moreover, y becomes unbounded if x does not decay faster to 0 than $1/t$ as $t \to \infty$ (resonance catastrophe).

Concretely: The dispersive equations arising in the asymptotic orders do not differ in the left hand side, since the spatial differential operator A always corresponds to $\tilde{\lambda}\partial_x^3$. As the source terms solve inhomogeneous equations with the same A, the solutions contain powers of t_3 as factors because of the resonance effect. Since there is no exponential damping as in the case of the diffusion equation, the expansion is expected to produce secular terms similarly to the regular expansion but in a much slower time scale.

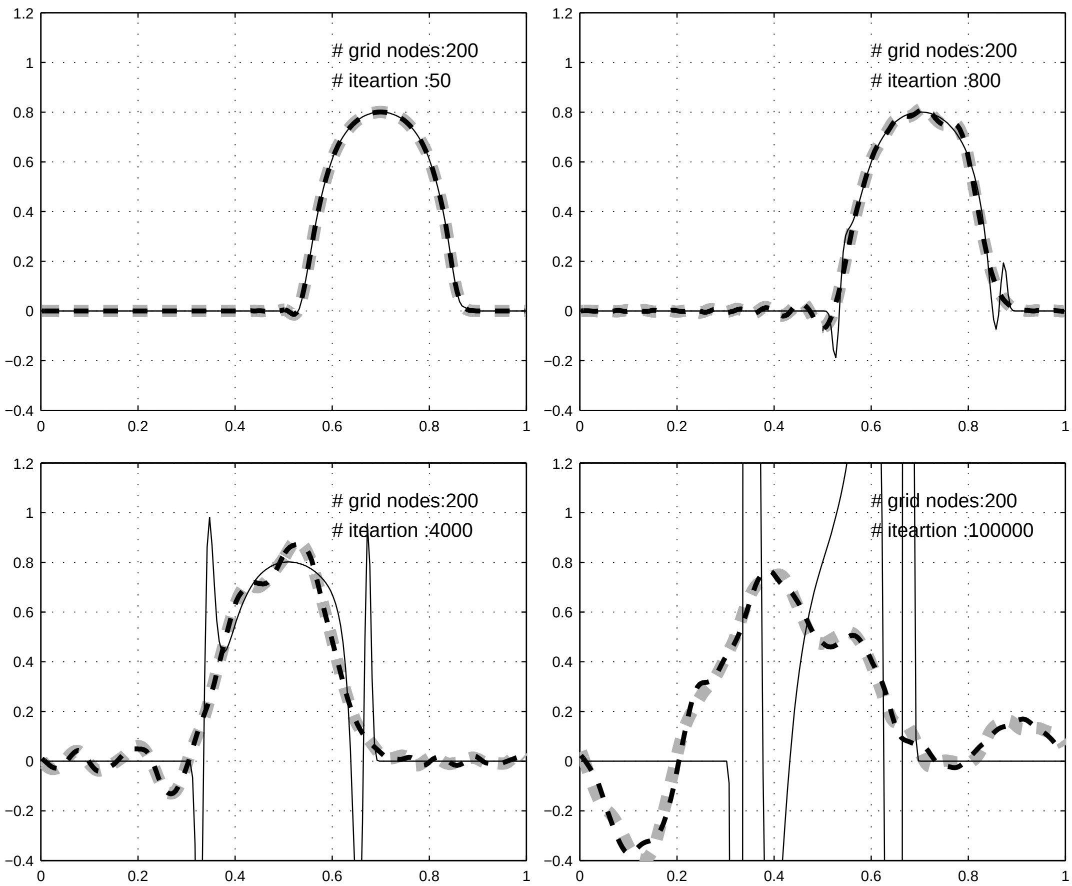


Figure 5.4:

The figure shows four snapshots of the mass moment $U(nh,\cdot)$ (gray dashed). The thin solid line indicates the prediction $u^{(0)}(nh,\cdot)+hu^{(1)}(nh,\cdot)+h^2u^{(2)}(nh,\cdot)$ obtained by the *regular* expansion. In contrast, the black dashed line refers to the prediction $u^{(0)}(nh,nh^3,\cdot)$ that results from the *multiscale* expansion.
After 50 iterations ($t=0.25$) the initial profile is shifted rightwards by 0.2 length units ($a=0.8$); the distortion is hardly visible. 750 iterations later ($t=4$) the distortion becomes noticeable. The regular expansion displays serious problems at the two knees where very sharp peaks are formed that evoke a certain similarity to the second and third derivative of the initial profile (see figure 5.2.3). Due to the secular terms in the regular expansion the derivatives of the initial profile are amplified in time, what makes the peaks grow quickly outside the range of the coordinate system. While the multiscale expansion imitates the oscillations of U very well; even after 100000 iterations the resemblance between U and the prediction is not lost.

It remains to compute A. Applying D^4, D^3, D^2 and D to (5.48_{0-3}) and taking the scalar product with $\mathsf{1}$ yields the following identities

$$\begin{aligned}
\langle D\ \mathsf{f}^{(3)}, \mathsf{1}\rangle &= \langle D\ \mathsf{E}^{(3)}, \mathsf{1}\rangle - \tfrac{1}{\omega}\langle D^2\mathsf{f}^{(2)}, \mathsf{1}\rangle - \tfrac{1}{2\omega}\langle D^3\mathsf{f}^{(1)}, \mathsf{1}\rangle - \tfrac{1}{6\omega}\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle \\
\langle D^2\mathsf{f}^{(2)}, \mathsf{1}\rangle &= \langle D^2\mathsf{E}^{(2)}, \mathsf{1}\rangle - \tfrac{1}{\omega}\langle D^3\mathsf{f}^{(1)}, \mathsf{1}\rangle - \tfrac{1}{2\omega}\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle \\
\langle D^3\mathsf{f}^{(1)}, \mathsf{1}\rangle &= \langle D^3\mathsf{E}^{(1)}, \mathsf{1}\rangle - \tfrac{1}{\omega}\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle \\
\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle &= \langle D^4\mathsf{E}^{(0)}, \mathsf{1}\rangle
\end{aligned}$$

where we have once again used $\langle D\partial_{t_3}\mathsf{f}^{(0)}, \mathsf{1}\rangle = 0$ for the first equation and $J\mathsf{f}^{(\beta)} = \omega\mathsf{E}^{(\beta)} - \omega\mathsf{f}^{(\beta)}$. Thereby we get for A:

$$\begin{aligned}
A &= \langle D\mathsf{f}^{(3)}, \mathsf{1}\rangle + \tfrac{1}{2}\langle D^2\mathsf{f}^{(2)}, \mathsf{1}\rangle + \tfrac{1}{6}\langle D^3\mathsf{f}^{(1)}, \mathsf{1}\rangle + \tfrac{1}{24}\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle \\
&= \langle D\mathsf{E}^{(3)}, \mathsf{1}\rangle + \underbrace{(\tfrac{1}{2} - \tfrac{1}{\omega})}_{=0 \text{ for } \omega=2}\langle D^2\mathsf{f}^{(2)}, \mathsf{1}\rangle + (\tfrac{1}{6} - \tfrac{1}{2\omega})\langle D^3\mathsf{f}^{(1)}, \mathsf{1}\rangle + (\tfrac{1}{24} - \tfrac{1}{6\omega})\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle \\
&= \langle D\mathsf{E}^{(3)}, \mathsf{1}\rangle + (\tfrac{1}{6} - \tfrac{1}{2\omega})\langle D^3\mathsf{E}^{(1)}, \mathsf{1}\rangle + \big[(\tfrac{1}{24} - \tfrac{1}{6\omega}) - \tfrac{1}{\omega}(\tfrac{1}{6} - \tfrac{1}{2\omega})\big]\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle \\
&= \langle D\mathsf{E}^{(3)}, \mathsf{1}\rangle + \underbrace{(\tfrac{1}{6} - \tfrac{1}{2\omega})}_{=-1/12 \text{ for } \omega=2}\langle D^3\mathsf{E}^{(1)}, \mathsf{1}\rangle + \underbrace{\big[\tfrac{1}{24} - \tfrac{1}{3\omega} + \tfrac{1}{2\omega^2}\big]}_{=0 \text{ for } \omega=2}\langle D^4\mathsf{f}^{(0)}, \mathsf{1}\rangle \\
&= \langle D\mathsf{E}^{(3)}, \mathsf{1}\rangle - \tfrac{1}{12}\langle D^3\mathsf{E}^{(1)}, \mathsf{1}\rangle
\end{aligned}$$

Next, we employ $\langle D\mathsf{E}^{(3)}, \mathsf{1}\rangle = \partial_{t_1}u^{(3)} + a\partial_x u^{(3)}$ and

$$\langle D^3\mathsf{E}^{(1)}, \mathsf{1}\rangle = \tfrac{1}{2}\big\langle(\partial_{t_1}^3 + 3\mathsf{s}\partial_{t_1}^2\partial_x + 3\partial_{t_1}\partial_x^2 + \mathsf{s}\partial_x^3)(1 + a\mathsf{s})u^{(1)}, \mathsf{1}\big\rangle = -2a(1-a^2)\partial_x^3 u^{(1)}$$

where the latter formula holds true as $u^{(1)}$ also solves the homogeneous advection equation. Thus, we finally arrive at

$$\partial_{t_3}u^{(1)} + \partial_{t_1}u^{(3)} + a\partial_x u^{(3)} - \tfrac{1}{12}\big(-2a(1-a^2)\big)\partial_x^3 u^{(1)} = 0$$

which is equivalent to (5.49_3).

Appendix: Details of the bump-shaped test function

Here we give the analytic description of the bump profile that we have used as initial condition to illustrate the cubic time scaling (cf. figure 5.3) and to validate the two-scale expansion with cubic time (cf. figure 5.4).

$$\Phi(x) := \begin{cases} x \le 0.3 & : \ 0 \\ 0.3 < x < 0.7: & \exp\left(\dfrac{-15^{-2}}{\big((x+0.2)-0.5\big)^2}\right)\cdot\exp\left(\dfrac{-15^{-2}}{\big((x-0.2)-0.5\big)^2}\right) \\ 0.7 \le x & : \ 0 \end{cases}$$

The function Φ is arbitrarily smooth (no kinks) but not analytic.
Note that each derivative is at least one order of magnitude larger than the preceding one. This explains why the regular expansion containing higher order derivatives 'explodes' rapidly. Furthermore Φ is identically zero outside of the bump, which means that all derivatives vanish there too. By this reason the regular expansion has not even a rudimentary chance to mimic the oscillations in figure 5.3 that extend over the whole interval.

The computation of the two-scale prediction for the mass moment in figure 5.4 is based on a Fourier decomposition of Φ.

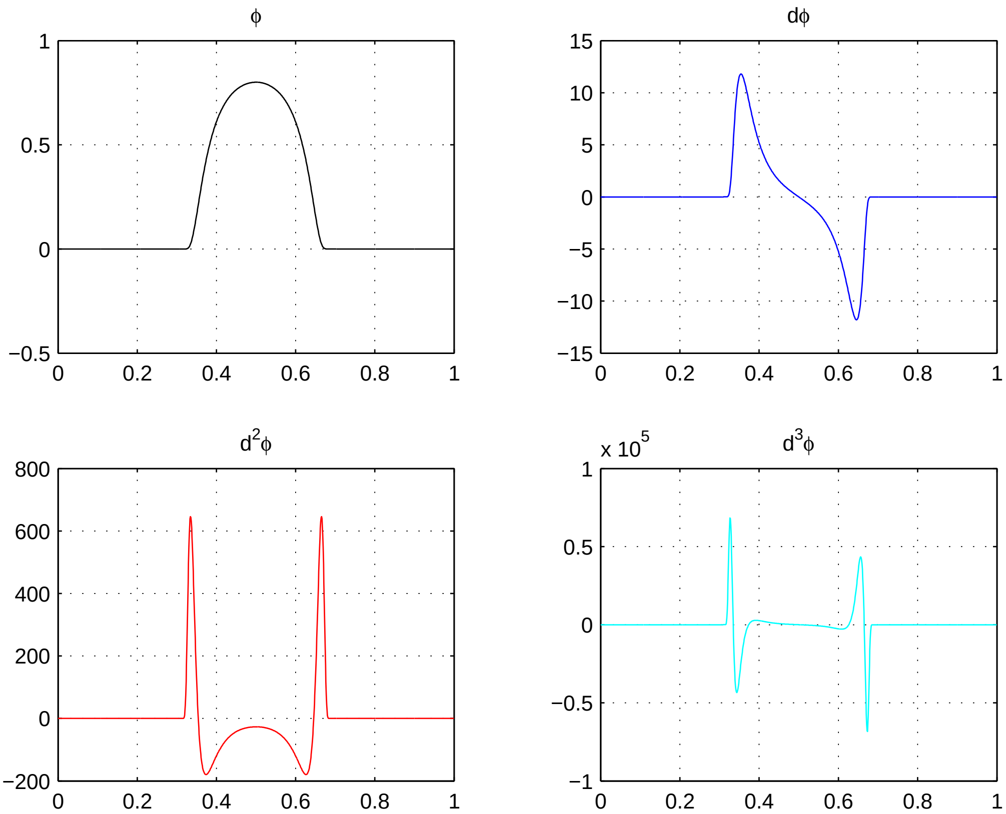


Figure 5.5: Left above: Test profile for initializing the mass moment. The other figures display the first, second and third derivative. Compare the scales of the ordinates axes differeing in orders of magnitude.

Chapter 6

Stability investigations around the D1P2 model

While chapter 4 and 5 are mainly concerned with the consistency analysis of an D1P3 and D1P2 lattice-Boltzmann scheme, we now broach the issue of stability being the second pillar besides consistency in a convergence proof.

Section 6.2 deals with a D1P2 algorithm emanating from a parabolically scaled lattice-Boltzmann equation. The considered algorithm is a special case ($\theta = 1$) of the D1P3 scheme analyzed in chapter 4, which discretizes the advection-diffusion equation with periodic boundary conditions. Theorem 6.1 states the main result establishing stability with respect to the ℓ_∞-norm. From this, uniform convergence in time and space can be inferred. For a purely diffusive scheme, stability is easily obtained (see proposition 6.1); therefore the principal challenge is the part of the equilibrium being responsible for the advection term in the target equation. Since the proof utilizes the matrix representation of the lattice-Boltzmann algorithm, shift matrices are introduced in the preliminary section 6.1.
In subsection 6.2.2 and 6.2.3 the spectrum of the evolution matrix is computed followed by a discussion of its properties. In particular, the eigenvalues near 1 are compared with those of the evolution operator pertaining to the target equation.

A slight modification of the advective-diffusive algorithm studied in 6.2 leads to a scheme approximating the pure advection equation. It is shown in 6.3.1 that the CFL-condition is necessary for stability of this algorithm. This result parallels a classical statement concerning the stability of basic finite difference methods. In 6.3.2 the observation of the preceding subsection is extended by formulating a precise criterion for ℓ_2-norm stability. Finally, the actual stability behavior of the lattice-Boltzmann scheme is compared with a prediction relying on the formal multiscale analysis performed in section 5.2.

Bounce-back-like boundary conditions are considered in section 6.4. Since non-periodic boundary conditions 'break the symmetry' of the evolution matrix, the analytic computation of eigenvalues is restricted to a purely diffusive lattice-Boltzmann scheme. This differs from the scheme in 6.2 just by ignoring the advective part of the equilibrium. In subsection 6.4.1 it is shown that the spectra of lattice-

Boltzmann evolution matrices imitate several features[1] which are known for the continuous counterparts of the target equations (diffusion equation with either periodic, Dirichlet or Neumann boundary conditions). Subsection 6.4.2 reports on the computation of eigenvectors associated to the evolution matrix. The chapter is closed by an outlook sketching possible future work.

6.1 Basics concerning shift matrices

Permutation matrices are distinguished by the property that their application to a vector effects a permutation of its components. Therefore these matrices must have a single element equal to one in each row and each column, while all other entries vanish. This accounts for both *orthogonality* (*unitarity*) and *bistochasticity*.
We speak of *shift matrices* if the underlying permutation corresponds to a *shift*. Note that shift matrices constitute a *cyclic* subgroup within the group of permutation matrices.

In numerical schemes for solving partial differential equations, the nodes of the mesh exchange numbers with their neighbors to approximate, for instance, certain derivatives. Locally, the transport of information between the nodes might be described by finite difference equations. However, in many examples a global representation of the transport operation is constructed by means of permutation matrices. Therefore permutation matrices are naturally involved in finite difference discretizations of differential operators.
Standard finite difference matrices are primarily build up as special linear combinations of shift matrices for the one-dimensional case. Finite difference matrices on domains in several dimensions being the Cartesian product of intervals (rectangular domains, parallelepipeds) are constructed effectively by taking appropriate tensor products of one-dimensional finite difference matrices. A nice example is given by the discretized Laplacian[2] on a domain with periodic boundary conditions.
As classical lattice-Boltzmann algorithms fit into the setup of finite difference schemes, the associated evolution matrices on geometrically simple domains are assembled using shift matrices too.

Before getting more concrete, let us settle some notation. In the sequel we consider a uniform grid with $N \in \mathbb{N}$ nodes discretizing a spatial interval. Furthermore we introduce the *unit roots* of order N as the N complex-valued solutions of the equation $z^N - 1 = 0$, forming a multiplicative group generated by

$$w := \mathrm{e}^{2\mathrm{i}\pi/N}. \tag{6.1}$$

Observe that many quantities (dimension of matrices, w etc.) occurring throughout this section depend on N without being explicitly marked for ease of notation. In

[1]For instance: multiplicity of eigenvalues, (almost) invariance of the spectrum under change of boundary conditions.

[2]The second derivative $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ can be approximated by the 3-point stencil, which becomes in matrix form on a uniform grid: $\Delta_{1,h} = h^{-2}(\mathbf{R} - 2\mathbf{I} + \mathbf{L})$. On a square domain in $\mathbb{R}^2$, the 5-point stencil for the Laplacian $\partial_x^2 + \partial_y^2$ is obtained by $\Delta_{2,h} = \Delta_{1,h} \otimes \mathbf{I} + \mathbf{I} \otimes \Delta_{1,h}$ with an appropriate indexation of the grid nodes.

the context of periodic boundary conditions (for the lattice-Boltzmann algorithm) it is advantageous to take the additive group

$$\mathbb{Z}_N := \mathbb{Z}/N\mathbb{Z} \text{ represented by } \{0, 1, \dots, N-1\}$$

as index set.

Those shift matrices being of primary interest for us are given elementwise by

$$\mathbf{L}_{k,\ell} := \delta_{k+1,\ell} \,, \qquad \mathbf{R}_{k,\ell} := \delta_{k-1,\ell} \,, \qquad k, \ell \in \mathbb{Z}_N$$

or equivalently by

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \\ 1 & 0 & \dots & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 & 0 \\ 0 & \dots & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{6.2}$$

Applied to a column vector, $\mathbf{L}$ performs an *up-shift* while $\mathbf{R}$ effectuates a *down-shift*. Visualizing the vector and its image by plotting its components versus their indices (corresponding to the abscissa values), the up-shift and down-shift look like a left-shift and right-shift respectively. Therefore we call $\mathbf{L}$ the *leftshift* and $\mathbf{R}$ the *rightshift*. Since $\mathbf{L}$ and $\mathbf{R}$ represent a shift by one step, the cyclic group of shift matrices is generated by $\mathbf{L}$ as well as by $\mathbf{R}$. Moreover the following equalities hold:

$$\mathbf{L} = \mathbf{R}^{N-1} = \mathbf{R}^{-1} \qquad \text{and} \qquad \mathbf{R} = \mathbf{L}^{N-1} = \mathbf{L}^{-1}.$$

Because of $\mathbf{L}^N = \mathbf{R}^N = \mathbf{I}$, it is possible to express the inverse matrices as $(N-1)^{\text{th}}$ power. Thanks to the orthogonality we get additionally

$$\mathbf{L}^{-1} = \mathbf{L}^\top = \mathbf{R} \quad \text{and} \quad \mathbf{R}^{-1} = \mathbf{R}^\top = \mathbf{L}. \tag{6.3}$$

The orthogonality requires also, that the spectrum of $\mathbf{L}$ and $\mathbf{R}$ lies on the unit circle. An extension of the Perron-Frobenius theorem (due to Wielandt, [61], page 363) states that the *peripheral spectrum*[3] of a *nonnegative* and *irreducible* matrix corresponds to the set of unit roots of a certain order $p \in \mathbb{N}$ multiplied by the spectral radius. p is also known as *index of cyclicity*; it is at most equal to the matrix dimension. Obviously, $\mathbf{L}$ and $\mathbf{R}$ are cyclic[4] matrices with maximal index of cyclicity, as each block has size 1×1. Combining this, we deduce at once, that the spectrum of $\mathbf{L}$ and $\mathbf{R}$ is given by the unit roots of order N. This can also be found by the following elementary computation of the characteristic polynomial presented for the sake of self-containment.

[3]Eigenvalues of maximal modulus, i.e. whose modulus is equal to the spectral radius.

[4]A matrix is called cyclic, if its column vectors can be transformed into one another by cyclic permutations of the elements. This is not related to the index of cyclicity, which is given as the number of eigenvalues with maximal modulus. The index of cyclicity determines also the number of nonzero-subblocks that are obtained if the matrix is transformed to a certain standard form with nonzero blocks on the first upper diagonal and in the lower left corner (similar to L).

Lemma 6.1. *The characteristic polynomial of the $N \times N$ leftshift and rightshift matrix is given by*

$$\lambda \mapsto (-1)^N(\lambda^N - 1).$$

Hence the unit roots of N^{th} order are the eigenvalues of $\mathbf{L}$ *and* $\mathbf{R}$*, i.e.* $\operatorname{spec}(\mathbf{L}) = \operatorname{spec}(\mathbf{R}) = \{1, w, w^2, \ldots, w^{N-1}\}$.

Proof: Since the characteristic polynomial is the same for a matrix and its transposed, we may restrict ourselves by (6.3) to the case of $\mathbf{L}$. Writing out the determinant and expanding it by *Leibniz' rule* with respect to the first column yields:

$$\det(\mathbf{L} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 & 0 & 0 & \ldots & 0 \\ 0 & -\lambda & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & -\lambda & 1 & 0 \\ 0 & 0 & \ldots & 0 & -\lambda & 1 \\ 1 & 0 & \ldots & 0 & 0 & -\lambda \end{vmatrix}$$

$$= -\lambda \cdot \begin{vmatrix} -\lambda & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & -\lambda & 1 & 0 \\ 0 & \ldots & 0 & -\lambda & 1 \\ 0 & \ldots & 0 & 0 & -\lambda \end{vmatrix} + (-1)^{N+1} \cdot \begin{vmatrix} 1 & 0 & 0 & \ldots & 0 \\ -\lambda & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & -\lambda & 1 & 0 \\ 0 & \ldots & 0 & -\lambda & 1 \end{vmatrix}$$

$$= (-\lambda) \cdot (-\lambda)^{N-1} + (-1)^{N+1} \cdot 1$$

$$= (-1)^N(\lambda^N - 1).$$

As the eigenvalues are just the roots of the characteristic polynomial, the assertion is verified. ■

Lemma 6.2. $\mathbf{L}$ *and* $\mathbf{R}$ *are* diagonalizable*: To each eigenvalue $w^m, m \in \mathbb{Z}_N$ belongs a one-dimensional eigenspace spanned by $(w^{\ell m})_{\ell \in \mathbb{Z}_N}$ in the case of* $\mathbf{L}$ *and by $(w^{(N-\ell-1)m})_{\ell \in \mathbb{Z}_N}$ or equivalently by $(w^{-\ell m})_{\ell \in \mathbb{Z}_N}$ in the case of* $\mathbf{R}$.

Proof: The claim is directly checked remembering that $w^N = w^{Nm} = 1$:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 1 \\ 1 & 0 & \ldots & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ w^m \\ \vdots \\ w^{(N-3)m} \\ w^{(N-2)m} \\ w^{(N-1)m} \end{pmatrix} = \begin{pmatrix} w^m \\ w^{2m} \\ \vdots \\ w^{(N-2)m} \\ w^{(N-1)m} \\ 1 \end{pmatrix} = \begin{pmatrix} w^m \\ w^{2m} \\ \vdots \\ w^{(N-2)m} \\ w^{(N-1)m} \\ w^{Nm} \end{pmatrix} = w^m \begin{pmatrix} 1 \\ w^m \\ \vdots \\ w^{(N-3)m} \\ w^{(N-2)m} \\ w^{(N-1)m} \end{pmatrix}$$

$$\begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & 1 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} w^{(N-1)m} \\ w^{(N-2)m} \\ w^{(N-3)m} \\ \vdots \\ w^m \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ w^{(N-1)m} \\ w^{(N-2)m} \\ \vdots \\ w^{2m} \\ w^m \end{pmatrix} = \begin{pmatrix} w^{Nm} \\ w^{(N-1)m} \\ w^{(N-2)m} \\ \vdots \\ w^{2m} \\ w^m \end{pmatrix} = w^m \begin{pmatrix} w^{(N-1)m} \\ w^{(N-2)m} \\ w^{(N-3)m} \\ \vdots \\ w^m \\ 1 \end{pmatrix}$$

Finally: $(w^{(N-\ell-1)m})_\ell = (w^{-\ell m-m})_\ell = w^{-m}(w^{-\ell m})_\ell$ being collinear to $(w^{-\ell m})_\ell$. ■

The knowledge of the eigenvectors (forming a basis in $\mathbb{C}^N$) permits us to diagonalize $\mathbf{L}$ and $\mathbf{R}$ by means of a *similarity transformation* where the columns of the transformation matrix must be composed of the eigenvectors. In the case of $\mathbf{R}$ we obtain the *Fourier matrix* $\mathbf{F}$, that is defined as the *Vandermonde* matrix $\mathbf{V}(1, w^{-1}, w^{-2}, ..., w^{1-N})$ multiplied by the normalization factor $N^{-1/2}$:

$$\mathbf{F} := \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & w^{-1} & w^{-2} & \dots & w^{-(N-1)} \\ 1 & w^{-2} & w^{-4} & \dots & w^{-2(N-1)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & w^{-(N-1)} & w^{-2(N-1)} & \dots & w^{-(N-1)(N-1)} \end{pmatrix} .$$

Since $\mathbf{L}$ and $\mathbf{R}$ are unitary, their eigenvectors associated to different eigenvalues are mutually orthogonal, so that $\mathbf{F}$ is unitary too. Noticing the symmetry of $\mathbf{F}$, we obtain its inverse as the complex-conjugated matrix:

$$\mathbf{F}^{-1} = \mathbf{F}^* \equiv \overline{\mathbf{F}}^\top = \overline{\mathbf{F}} = \tfrac{1}{\sqrt{N}} \mathbf{V}(1, \overline{w}^{-1}, \overline{w}^{-2}, ..., \overline{w}^{1-N}) .$$

Using the relation $\overline{w} = w^{-1}$ or $\overline{w}^{-1} = w$ we get

$$\mathbf{F}^{-1} = N^{-1/2} \mathbf{V}(1, w^1, w^2, ..., w^{N-1}),$$

which is the matrix whose columns are the eigenvectors of $\mathbf{L}$. So we can summarize the reasoning by the following corollary.

Corollary 6.1. *Let* $\mathbf{W} := \mathrm{diag}(1, w, w^2, \dots, w^{N-1})$*, then:*

$$\begin{aligned} \mathbf{FLF}^{-1} = \mathbf{W} \quad &\Leftrightarrow \quad \mathbf{F}^{-1}\mathbf{LF} = \overline{\mathbf{W}} , \\ \mathbf{F}^{-1}\mathbf{RF} = \mathbf{W} \quad &\Leftrightarrow \quad \mathbf{FRF}^{-1} = \overline{\mathbf{W}} . \end{aligned}$$

6.2 LB advection-diffusion scheme with periodic boundary conditions

Matrix formulation of the algorithm

Let us briefly recall some fundamentals concerning the object of investigation: the D1P2 lattice-Boltzmann algorithm with periodic boundary conditions discretizing the advection-diffusion equation.

Adopting the notation of the previous chapters we denote by $\mathsf{F} : \mathcal{T}_h \times \mathcal{X}_h \to \mathbb{R}^2$ the two-component population function that is defined on the temporal grid $\mathcal{T}_h = \{0, \Delta t, 2\Delta t, 3\Delta t, ...\}$ and the spatial mesh $\mathcal{X}_h$. The latter one discretizes the spatial domain $[0, L]$ by setting $\mathcal{X}_h = \{\frac{1}{2}h, \frac{3}{2}h, ..., \frac{2N-1}{2}h\}$ or alternatively $\mathcal{X}_h = \{0, h, ..., (N-1)h\}$ with the grid spacing $\Delta x = h = \frac{L}{N}$ for some $N \in \mathbb{N}$. Since the algorithm is derived from a parabolically scaled lattice-Boltzmann equation (see end of section 1.1.2) spatial and temporal increments shall be coupled via $\Delta t = \Delta x^2 = h^2$.

The populations in the nodes evolve according to the difference equation

$$\mathsf{F}_s(t_{n+1}, x_j) = \underbrace{\mathsf{F}_s(t_n, x_{j-s}) + \big[J\mathsf{F}(t_n, x_{j-s})\big]_s}_{= \mathsf{C}_s(t_{n+1}, x_{j-s})} . \tag{6.4}$$

Here the population function F is indexed by the D1P2 velocity set $\mathcal{S} = \{-1, 1\}$. F_{-1} and F_1 denote the distributions of virtual particles moving respectively to the left and right. Observe that (6.4) encodes implicitly periodic boundary conditions, since the spatial index j is assumed to be in $\mathbb{Z}_N$, such that $j + s$ and $j - s$ are understood modulo N.

The BGK collision operator J is given by

$$J\mathsf{F} = \omega(\mathcal{E}(U) - \mathsf{F}), \tag{6.5}$$

involving the *equilibrium* $\mathcal{E}(U)$, which is considered here as function of the mass moment $U = \mathsf{F}_{-1} + \mathsf{F}_1$:

$$\mathcal{E}_s(U) := \tfrac{1}{2}(1 + ahs - \tfrac{1}{\omega}ch^2)U. \tag{6.6}$$

Besides the collision frequency ω in (6.5), two macroscopic parameters a and c appear in the equilibrium figuring also in the target equation

$$\partial_t v + a\partial_x v - (\tfrac{1}{\omega} - \tfrac{1}{2})\partial_x^2 v + cv = 0, \tag{6.7}$$

which has been found by the consistency analysis in chapter 4.
The initialization of the lattice-Boltzmann algorithm is a topic of its own. The simplest possibility is to set $\mathsf{F}(0, x_j) = \mathcal{E}\big(v_0(x_j)\big)$ where $v_0 = v(0, \cdot)$ denotes the initial value pertaining to the target equation (6.7).

Let us now cast the algorithm (6.4) in matrix form. For this we notice that each iteration (evolution step) is decomposed into two substeps. The first one is referred to as the *collision*[5] executed as purely nodal[6] operation. It consists in computing the collision product C (underbraced RHS in (6.4)) from the current population function F at each node. Representing the equilibrium operator and J as 2×2 matrices acting on the two-component population function F we obtain

$$\mathsf{C}(t,x) := (I + J)\mathsf{F}(t,x) \qquad \Leftrightarrow \qquad \begin{pmatrix} \mathsf{C}_{-1}(t,x) \\ \mathsf{C}_{+1}(t,x) \end{pmatrix} := \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} \mathsf{F}_{-1}(t,x) \\ \mathsf{F}_{+1}(t,x) \end{pmatrix},$$

where the coefficients of the 2×2 *collision block* are read off from (6.4) (6.5) and (6.6):

$$\begin{aligned} \alpha &:= 1 - \tfrac{1}{2}\omega - \tfrac{1}{2}\omega ah - \tfrac{1}{2}ch^2, & \beta &:= \tfrac{1}{2}\omega - \tfrac{1}{2}\omega ah - \tfrac{1}{2}ch^2, \\ \gamma &:= \tfrac{1}{2}\omega + \tfrac{1}{2}\omega ah - \tfrac{1}{2}ch^2, & \delta &:= 1 - \tfrac{1}{2}\omega + \tfrac{1}{2}\omega ah - \tfrac{1}{2}ch^2. \end{aligned} \tag{6.8}$$

The iteration is completed by the *propagation* or *transport* of the collision products to their corresponding destination nodes. This second substep is non-nodal, for the basic operation of copying a collision product involves two different nodes simultaneously. At this point the shift matrices come into play.

[5] We commit here a slight abuse of terminology. From a kinetic point of view the collision corresponds to the application of the operator J, whilst we understand it now in the sense of evaluating the right hand side of (6.4). To avoid misunderstandings, J may be referred to as the relaxation operator.

[6] The collision can be implemented by a loop running over all nodes. Inside the loop only populations pertaining to the same node are manipulated.

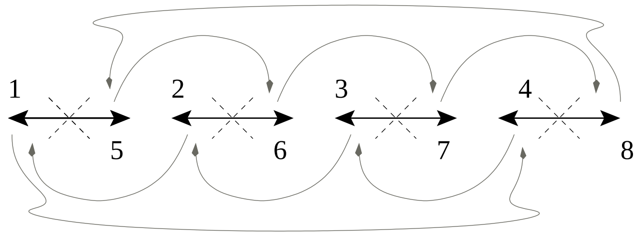


Figure 6.1: The configuration of the D1P2 grid with $N = 4$. The mesh-nodes are represented by dashed crosses $\times$. The arrows symbolize two populations located in each node. After the collision step, every node sends one of the collision products to its left and the other to its right neighbor. Imposing periodic boundary conditions means that the left neighbor of the leftmost node is the rightmost node and the other way round.

For the implementation of the algorithm one may introduce a column vector F embracing the populations of all nodes. Since the equilibrium $\mathcal{E}$ depends linearly on the mass moment U and thus on the populations, the algorithm on the whole grid can be formulated in matrix-vector notation. For our further investigations it is advantageous to index the populations as indicated in figure 6.1. So the populations are arranged as components of the vector F in such a way, that the upper half contains the left-moving populations from left to right, whereas the right-traveling populations are put into the lower half (from left to right as well). Without external sources the population vector $\mathsf{F}(n+1)$ of the $(n+1)^{\text{th}}$ iteration is obtained from $\mathsf{F}(n)$ by applying an evolution matrix $\mathbf{E}$. Hence (6.4) shall be written in the form

$$\mathsf{F}(n+1) = \mathbf{E}\,\mathsf{F}(n) = \mathbf{E}^{n+1}\,\mathsf{F}(0). \tag{6.9}$$

How is the evolution matrix $\mathbf{E}$ set up explicitly? As already mentioned, every iteration is executed by performing two substeps. This entails, that $\mathbf{E}$ should be obtained as a product of a collision matrix $\mathbf{C}$ with a transport matrix $\mathbf{T}$. Since the collision is a nodal operation, $\mathbf{C}$ attains the form

$$\mathbf{C} = \left(\begin{array}{cccc|cccc}
\alpha & \cdot & \cdot & \cdot & \beta & \cdot & \cdot & \cdot \\
\cdot & \alpha & \cdot & \cdot & \cdot & \beta & \cdot & \cdot \\
\cdot & \cdot & \alpha & \cdot & \cdot & \cdot & \beta & \cdot \\
\cdot & \cdot & \cdot & \alpha & \cdot & \cdot & \cdot & \beta \\
\hline
\gamma & \cdot & \cdot & \cdot & \delta & \cdot & \cdot & \cdot \\
\cdot & \gamma & \cdot & \cdot & \cdot & \delta & \cdot & \cdot \\
\cdot & \cdot & \gamma & \cdot & \cdot & \cdot & \delta & \cdot \\
\cdot & \cdot & \cdot & \gamma & \cdot & \cdot & \cdot & \delta
\end{array}\right)$$

in the case of a configuration with four nodes (figure 6.1). Observe that each block is a multiple of the identity matrix. This is due to the fact that the coefficients of the target equation are supposed to be spatially constant[7]. In general, we get

$$\mathbf{C} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \otimes \mathbf{I} = \begin{pmatrix} \alpha\mathbf{I} & \beta\mathbf{I} \\ \gamma\mathbf{I} & \delta\mathbf{I} \end{pmatrix} , \tag{6.10}$$

where $\otimes$ denotes the *Kronecker* or *tensor product* for matrices and $\mathbf{I}$ the $N \times N$ identity matrix.

The second step is done by the *transport matrix.* For the four-node grid in figure 6.1 it becomes

$$\mathbf{T} = \left(\begin{array}{cccc|cccc} \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \end{array} \right) .$$

Thanks to the specific numbering of the populations, the transport matrix is 2×2 block-diagonal with a leftshift and rightshift on its diagonal. The general construction rule for the transport matrix is given by

$$\mathbf{T} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes \mathbf{L} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes \mathbf{R} = \begin{pmatrix} \mathbf{L} & 0 \\ 0 & \mathbf{R} \end{pmatrix} .$$

Multiplying $\mathbf{T}$ with $\mathbf{C}$ from the right, finally yields the *evolution* or *iteration matrix* $\mathbf{E}$ of the algorithm.

$$\mathbf{E} = \mathbf{T} \cdot \mathbf{C} = \begin{pmatrix} \mathbf{L} & 0 \\ 0 & \mathbf{R} \end{pmatrix} \begin{pmatrix} \alpha\mathbf{I} & \beta\mathbf{I} \\ \gamma\mathbf{I} & \delta\mathbf{I} \end{pmatrix} = \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix} . \tag{6.11}$$

Notice the order of the factors, such that $\mathbf{C}$ is applied to the population vector F *prior to* $\mathbf{T}$.

The eigenvalues of the collision block influence the stability of the evolution matrix essentially. They are listed here for later considerations.

Lemma 6.3. *The eigenvalues of the collision block* $\left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$ *are*

$$1 - \omega \quad \textit{with} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \qquad \textit{and} \qquad 1 - ch^2 \quad \textit{with} \begin{pmatrix} \omega(1 - ah) - ch^2 \\ \omega(1 + ah) - ch^2 \end{pmatrix}$$

spanning the associated eigenspaces.

It is noteworthy that the eigenvalues do not depend on the advection velocity a.

[7]Furthermore it is assumed that a, c and ω are also constant in time. Therefore $\mathbf{E}$ does not depend on the iteration index n

6.2.1 An ℓ_∞-stability result

The notion of stability plays a crucial role in the theory of dynamical systems regardless whether they are of continuous or discrete nature. There is no "ultimate" definition of stability; rather there is a multitude of specialized notions sharing a basic idea but differing in details depending on the context they are tailored for.
The subsequent definitions refer to a family of iterations like (6.9). It is supposed that the family is parameterized by a discretization parameter $\Delta t > 0$, which is interpreted as time step. Generally, the spaces that $\mathbf{E}$ and F belong to depend on Δt (compare with the setup described on page 59 and the following ones).

Definition 6.1. *An explicit iteration scheme like* (6.9) *is called* ***stable*** *with respect to the operator norm* $\|\cdot\|$*, if the subsequent condition is fulfilled:*

$$\exists C > 0 : \ \forall \, \text{admissible } \Delta t > 0 : \ \forall n \in \mathbb{N}_0 : \quad \|\mathbf{E}^n\| < C.$$

It should be noted, that the evolution operator $\mathbf{E}$ as well as the norm $\|\cdot\|$ may depend on Δt; for instance via the space it operates on. In contrast, the constant C is required to be independent of the time step Δt and of the iteration n.

In many situations this property is too restrictive. In particular, it does not make sense if the iteration scheme shall approximate the solution of a differential equation that gets unbounded as time tends to infinity. Euler's forward polygon method $x(t_{n+1}) = (1 + a\Delta t)x(t_n)$ discretizing the ODE $\frac{\mathrm{d}}{\mathrm{d}t}x(t) = ax(t)$ may serve as example. In this case the evolution operator corresponds to a multiplication with the number $1 + a\Delta t$. Obviously, we get for $a > 0$: $\lim_{n\to\infty}(1 + a\Delta t)^n = \infty$. So the scheme is not stable in the above sense – a feature, which is of course inherited from its continuous counterpart. On the other hand $(1 + a\Delta t)^{[T/\Delta t]}$ is bounded for $\Delta t \downarrow 0$, since the limit $\lim_{\Delta t\to 0}(1 + a\Delta t)^{[T/\Delta t]} = \mathrm{e}^{aT}$ does exist for any fixed time $T > 0$. It is desirable to dispose of a suitably weaker stability notion for such situations.

Definition 6.2. *An explicit iteration scheme in the form of* (6.9) *is said to be* ***stable over bounded time intervals***[8] $[0, T]$ *with respect to the operator norm* $\|\cdot\|$*, if the following condition holds true:*

$$\forall T > 0 : \ \exists C > 0 : \ \forall \, \text{admissible } \Delta t > 0 : \ \forall n \in \mathbb{N}_0, \ n\Delta t \le T : \quad \|\mathbf{E}^n\| < C.$$

As in definition 6.1, C is independent of Δt but depends on T. Notice that stability in the sense of definition 6.1 implies stability over bounded time intervals.

For the lattice-Boltzmann algorithm the actual discretization parameter, the time step Δt is linked to the spatial grid constant $\Delta x = h$ by setting $\Delta t = h^2$, where $h = \frac{L}{N}$ with $N \in \mathbb{N}$ can only attain integer fractions of the interval length L.
The subsequent stability result refers to the matrix infinity-norm $\|\cdot\|_\infty$ which gives rise to uniform convergence of the algorithm in space. Remember that the matrix infinity-norm is induced by the maximum norm for vectors. As operator norm, $\|\cdot\|_\infty$ is submultiplicative. For a given matrix the $\|\cdot\|_\infty$-norm is easily obtained by

[8]Also referred to as Lax-Richtmyer stability, see [45] page 106.

row-wise summing the moduli of the matrix elements and taking the maximum of the resulting sums. Prior to the main result let us first consider the purely diffusive case.

Proposition 6.1. ***(purely diffusive case)*** *For $a = 0$ and $c = 0$ the lattice-Boltzmann algorithm (6.4) with equilibrium (6.6) is stable in the sense of definition 6.1 with respect to $\|\cdot\|_\infty$, if the relaxation frequency ω satisfies: $0 \le \omega \le 2$.*

The assertion becomes quickly comprehensible, if we notice, that the evolution matrix is *stochastic* under the requirements of the proposition.

Proof: The proof works with the standard estimate based on the submultiplicativity of $\|\cdot\|_\infty$ yielding $\|\mathbf{E}^n\|_\infty \le \|\mathbf{E}\|_\infty^n$. So the assertion is verified, if we show $\|\mathbf{E}\|_\infty \le 1$. Using again the submultiplicativity we get by (6.11) $\|\mathbf{E}\|_\infty \le \|\mathbf{T}\|_\infty \, \|\mathbf{C}\|_\infty$. Now the transport-matrix $\mathbf{T}$ is a permutation matrix and contains in each row exactly a single non-vanishing element equal to 1. Therefore $\|\mathbf{T}\|_\infty = 1$. If $a = 0$ and $c = 0$ the coefficients of the collision block reduce to

$$\begin{aligned} \alpha &= 1 - \tfrac{1}{2}\omega, & \beta &= \tfrac{1}{2}\omega, \\ \gamma &= \tfrac{1}{2}\omega, & \delta &= 1 - \tfrac{1}{2}\omega. \end{aligned}$$

Evidently all coefficients are non-negative for $\omega \in [0, 2]$. Regarding the structure of $\mathbf{C}$ in (6.10) we obtain

$$\|\mathbf{C}\|_\infty = \max\left\{|\alpha| + |\beta|,\, |\gamma| + |\delta|\right\} = (1 - \tfrac{1}{2}\omega) + \tfrac{1}{2}\omega = 1\,.$$

Hence $\|\mathbf{E}\|_\infty \le \|\mathbf{T}\|_\infty \|\mathbf{C}\|_\infty = 1 \cdot 1 = 1$ which finishes the proof. ■

The stability condition with respect to ω is well known. In fact, it extends to many other linear BGK lattice-Boltzmann algorithms, amongst others the D2P9 model with Stokes equilibrium. The condition $0 \le \omega \le 2$ is not only sufficient but also necessary for stability as shown in [63] on page 208. The reasoning therein is based on the idea to consider a spatially constant population function. In this case the lattice-Boltzmann algorithm simplifies, such that the application of the evolution matrix corresponds to a multiplication by $1 - \omega$. The modulus of this number is larger than 1 if the condition on ω is violated.
At the end of this paragraph the necessity is also proven by another argument.

Now, we come to the principal result of this subsection:

Theorem 6.1. ***(general case)*** *Assume that the relaxation frequency satisfies the* stability condition $0 < \omega < 2$. *Then the lattice-Boltzmann algorithm (6.4) with the equilibrium (6.6) is* stable *in the sense of definition 6.2 with respect to $\|\cdot\|_\infty$ for any advection velocity $a \in \mathbb{R}$ and any reactivity $c \ge 0$.*

The proof is based on the following lemma:

Lemma 6.4. *Let α, β, γ and δ be non-negative numbers. Then the following equality holds for all $n \in \mathbb{N}_0$*

$$\left\| \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix}^n \right\|_\infty = \left\| \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^n \right\|_\infty ,$$

where $\mathbf{L}$ *and* $\mathbf{R}$ *are the shift matrices of arbitrary but equal dimension defined in (6.2).*

Proof: Actually, we will prove another claim, from which the statement of the lemma follows easily due to the assumed non-negativity of the four numbers. For this let us introduce the abbreviations

$$\begin{pmatrix} \mathbf{P}_{11}^{(n)} & \mathbf{P}_{12}^{(n)} \\ \mathbf{P}_{21}^{(n)} & \mathbf{P}_{22}^{(n)} \end{pmatrix} := \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix}^n \qquad \text{and} \qquad \begin{pmatrix} \psi_{11}^{(n)} & \psi_{12}^{(n)} \\ \psi_{21}^{(n)} & \psi_{22}^{(n)} \end{pmatrix} := \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^n .$$

Here $\mathbf{P}_{ij}^{(n)}$ with $i,j \in \{1,2\}$ is a block-matrix having the same size $N \times N$ as the shift matrices $\mathbf{L}$ and $\mathbf{R}$. We assert now, that every row-sum of $\mathbf{P}_{ij}^{(n)}$ is equal to $\psi_{ij}^{(n)}$, i.e. the following equation holds for all row indices $k \in \{1, ..., N\}$ and all $i,j \in \{1,2\}$:

$$\mathrm{RowSum}_k(\mathbf{P}_{ij}^{(n)}) = \psi_{ij}^{(n)} .$$

The verification of this claim is performed inductively over the exponent n.

Start: Evidently, the assertion is true for $n = 0$, since we obtain in this case identity matrices.

$$\begin{pmatrix} \mathbf{P}_{11}^{(0)} & \mathbf{P}_{12}^{(0)} \\ \mathbf{P}_{21}^{(0)} & \mathbf{P}_{22}^{(0)} \end{pmatrix} = \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix}^0 = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix} \leftrightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^0 = \begin{pmatrix} \psi_{11}^{(0)} & \psi_{12}^{(0)} \\ \psi_{21}^{(0)} & \psi_{22}^{(0)} \end{pmatrix}$$

Step: By hypothesis of induction the claim holds for $n \in \mathbb{N}_0$. We have to show that it is also true for $n+1$. Now we have

$$\begin{aligned} \begin{pmatrix} \mathbf{P}_{11}^{(n+1)} & \mathbf{P}_{12}^{(n+1)} \\ \mathbf{P}_{21}^{(n+1)} & \mathbf{P}_{22}^{(n+1)} \end{pmatrix} &= \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix} \begin{pmatrix} \mathbf{P}_{11}^{(n)} & \mathbf{P}_{12}^{(n)} \\ P_{21}^{(n)} & \mathbf{P}_{22}^{(n)} \end{pmatrix} \\ &= \begin{pmatrix} \alpha\mathbf{L}\mathbf{P}_{11}^{(n)} + \beta\mathbf{L}\mathbf{P}_{21}^{(n)} & \alpha\mathbf{L}\mathbf{P}_{12}^{(n)} + \beta\mathbf{L}\mathbf{P}_{22}^{(n)} \\ \gamma\mathbf{R}\mathbf{P}_{11}^{(n)} + \delta\mathbf{R}\mathbf{P}_{21}^{(n)} & \gamma\mathbf{R}\mathbf{P}_{12}^{(n)} + \delta\mathbf{R}\mathbf{P}_{22}^{(n)} \end{pmatrix} \end{aligned}$$

and

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^{n+1} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} \psi_{11}^{(n)} & \psi_{12}^{(n)} \\ \psi_{21}^{(n)} & \psi_{22}^{(n)} \end{pmatrix} = \begin{pmatrix} \alpha\psi_{11}^{(n)} + \beta\psi_{21}^{(n)} & \alpha\psi_{12}^{(n)} + \beta\psi_{22}^{(n)} \\ \gamma\psi_{11}^{(n)} + \delta\psi_{21}^{(n)} & \gamma\psi_{12}^{(n)} + \delta\psi_{22}^{(n)} \end{pmatrix} .$$

Observing that multiplication by $\mathbf{L}$ respectively $\mathbf{R}$ from the left acts on a matrix like an upshift or downshift of the rows[9] respectively, we recognize that the row-sums of $\mathbf{P}_{ij}^{(n)}$ remain invariant under multiplication with the shift matrices. Using the linearity of the 'RowSum-operator', we get finally for any row index $k \in \{1, \ldots, N\}$:

$$\begin{aligned} \mathrm{RowSum}_k(\mathbf{P}_{11}^{(n+1)}) &= \mathrm{RowSum}_k(\alpha\mathbf{L}\mathbf{P}_{11}^{(n)} + \beta\mathbf{L}\mathbf{P}_{21}^{(n)}) \\ &= \alpha\,\mathrm{RowSum}_k(\mathbf{L}\mathbf{P}_{11}^{(n)}) + \beta\,\mathrm{RowSum}_k(\mathbf{L}\mathbf{P}_{21}^{(n)}) \\ &= \alpha\,\mathrm{RowSum}_k(\mathbf{P}_{11}^{(n)}) + \beta\,\mathrm{RowSum}_k(\mathbf{P}_{21}^{(n)}) \\ &= \alpha\psi_{11}^{(n)} + \beta\psi_{21}^{(n)} \;=\; \psi_{11}^{(n+1)} . \end{aligned}$$

[9]Just to have a concrete example before one's eyes: $\begin{pmatrix} \cdot & 1 & \cdot \\ \cdot & \cdot & 1 \\ 1 & \cdot & \cdot \end{pmatrix} \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix} = \begin{pmatrix} d & e & f \\ g & h & k \\ a & b & c \end{pmatrix}$

This computation can be analogously repeated for the other three index combinations, which concludes the proof of the claim.

Let us now revert to the assertion of the lemma. As α, β, γ and δ are non-negative, and trivially equal to their modulus, the infinity-norm $\|\cdot\|_\infty$ corresponds to the maximal row-sum. Hence:

$$\begin{aligned}\left\|\begin{pmatrix}\alpha\mathbf{L} & \beta\mathbf{L}\\ \gamma\mathbf{R} & \delta\mathbf{R}\end{pmatrix}^n\right\|_\infty &= \max\left\{\mathrm{RowSum}\big(\mathbf{P}_{11}^{(n)}+\mathbf{P}_{12}^{(n)}\big),\ \mathrm{RowSum}\big(\mathbf{P}_{21}^{(n)}+\mathbf{P}_{22}^{(n)}\big)\right\}\\ &= \max\left\{\psi_{11}^{(n)}+\psi_{12}^{(n)},\ \psi_{21}^{(n)}+\psi_{22}^{(n)}\right\} = \left\|\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}^n\right\|_\infty .\end{aligned}$$

■

Proof: (Theorem 6.1) According to definition (6.2) we have to justify the boundedness of the set

$$M := \left\{\|\mathbf{E}\|^n : \ \ h \in \mathcal{H}_L,\ \ 0 \le n \le [T/h^2]\right\}.$$

From our assumption concerning ω follows

$$1 - \tfrac{1}{2}\omega > 0 \qquad \text{and} \qquad \tfrac{1}{2}\omega > 0.$$

For given a and c we can therefore find some grid spacing $h_0 \in \mathcal{H}_L$, such that the four coefficients of the collision block in (6.8) are *positive* for all admissible $h \le h_0$. This suggests to decompose M in the following manner:

$$M = \underbrace{\left\{\|\mathbf{E}\|^n : \ \ h \in \mathcal{H}_L,\ h > h_0\ \ 0 \le n \le [T/h^2]\right\}}_{\text{finite set}} \ \ \cup \ \ \underbrace{\left\{\|\mathbf{E}\|^n : \ \ h \in \mathcal{H}_L,\ h \le h_0\ \ 0 \le n \le [T/h^2]\right\}}_{\text{bounded by } D} .$$

Clearly, the first subset is finite and hence trivially bounded. In order to show that the second subset is bounded too, we take recourse to lemma (6.4). This gives for all $n \in \mathbb{N}$

$$\|\mathbf{E}^n\|_\infty = \left\|\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}^n\right\|_\infty ,$$

whenever $h < h_0$ or equivalently the number of grid nodes N is greater than $N_0 = L/h_0$. Lemma 6.3 provides the eigenvalues of the collision block, whose moduli are less than 1, if ω satisfies the stability condition, i.e. $\omega \in (0,2)$, and $h < h_0$. This entails that the collision block is *zero-convergent*. Therefore there exists

$$D := \sup\left\{\left\|\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}^n\right\|_\infty \ : \quad n \in \mathbb{N}_0\right\} < \infty$$

and this yields finally

$$\sup_{\substack{h\in\mathcal{H}_L\\ h<h_0}} \max_{0\le n\le[T/h^2]} \|\mathbf{E}^n\|_\infty \ \le \ \sup_{\substack{h\in\mathcal{H}_L\\ h<h_0}} \sup_{n\in\mathbb{N}} \|\mathbf{E}^n\|_\infty \le D ,$$

which establishes the boundedness of the second subset. ■

What has been shown in the proof is actually more than asserted by the theorem: If we declare only those grid spacings $h \in \mathcal{H}_L = \{L/N : N \in \mathbb{N}\}$ for admissible such that the h-dependent coefficients α, β, γ and δ become positive for fixed a and c, then even stability according to definition 6.1 is obtained.

Lemma 6.5. *Let α, β, γ and δ arbitrary real numbers. Then the subsequent inequality holds for all $n \in \mathbb{N}_0$:*

$$\left\| \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}^n \right\|_\infty \le \left\| \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix}^n \right\|_\infty \le \left\| \begin{pmatrix} |\alpha| & |\beta| \\ |\gamma| & |\delta| \end{pmatrix}^n \right\|_\infty .$$

Proof: Since the matrices share the same formal structure, the four elements which belong to the matrices raised to the n^{th} power are composed of analogous products with respect to the four coefficients. Therefore the assertion follows directly from the triangle inequality and the definition of the matrix infinity norm.
Nota bene: Actually, the matrix in the center is not 2×2 but $2N \times 2N$. The products of coefficients that appear as addends in the same element if the matrix is 2×2 (left and right), may be distributed over different matrix elements for the $2N \times 2N$ matrix. ∎

Proposition 6.2. *Under the hypothesis of theorem 6.1 the condition $0 \le \omega \le 2$ is* necessary *for stability.*

Proof: A necessary criterion for stability as specified in definition 6.1 is "spectral stability" meaning that the whole spectrum of $\mathbf{E}$ must lie inside the closed unit disk. According to lemma 6.3, $1-\omega$ is an eigenvalue of the collision block. If ω violates the given stability condition, i.e. $\omega \notin [0, 2]$, then the eigenvalue $1 - \omega$ leaves the unit disk. Therefore the powers of the collision block cannot be bounded with respect to any norm. Now the left part of the above inequality implies, that the powers of the evolution matrix become unbounded too. ∎

Remark 6.1. The quadratic coupling of the space and time step is also encountered in the case of explicit finite difference methods for parabolic equations like the diffusion equation. It reflects that parabolic equations combine a first order time derivative with a second order space derivative. However, it is remarkable that this coupling does not involve the diffusivity $\nu = (\frac{1}{\omega} - \frac{1}{2})$ in opposition to standard finite difference schemes.

6.2.2 The spectral limit set of the evolution matrices

Theorem 6.2. *Consider the lattice-Boltzmann algorithm (6.4) with equilibrium (6.6) over a grid of N nodes. The characteristic polynomial of the associated evolution matrix $\mathbf{E}$ factors into N quadratic polynomials and is given by*

$$\chi_{\mathbf{E}}(\lambda) \;=\; \prod_{m=0}^{N-1} \left[\lambda^2 \;-\; (\alpha w^m + \delta \overline{w}^m)\lambda \;+\; \alpha\delta - \beta\gamma \right] . \tag{6.12}$$

Proof: At the end of section 6.1 we have seen, that the shift matrices $\mathbf{L}$ and $\mathbf{R}$ are diagonalized by a similarity transformation[10] with the Fourier matrix $\mathbf{F}$. This nurtures the hope to simplify $\mathbf{E}$ significantly using the Fourier matrix again. We

[10]Transition to Fourier basis.

try the following ansatz:

$$
\begin{aligned}
\tilde{\mathbf{F}} \cdot \mathbf{E} \cdot \tilde{\mathbf{F}}^{-1} \;&:=\; \begin{pmatrix} \mathbf{F} & 0 \\ 0 & \mathbf{F} \end{pmatrix} \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix} \begin{pmatrix} \mathbf{F}^{-1} & 0 \\ 0 & \mathbf{F}^{-1} \end{pmatrix} \\
&=\; \begin{pmatrix} \alpha\mathbf{FL} & \beta\mathbf{FL} \\ \gamma\mathbf{FR} & \delta\mathbf{FR} \end{pmatrix} \begin{pmatrix} \mathbf{F}^{-1} & 0 \\ 0 & \mathbf{F}^{-1} \end{pmatrix} \\
&=\; \begin{pmatrix} \alpha\mathbf{FLF}^{-1} & \beta\mathbf{FLF}^{-1} \\ \gamma\mathbf{FRF}^{-1} & \delta\mathbf{FRF}^{-1} \end{pmatrix} \\
&=\; \begin{pmatrix} \alpha\mathbf{W} & \beta\mathbf{W} \\ \gamma\overline{\mathbf{W}} & \delta\overline{\mathbf{W}} \end{pmatrix} .
\end{aligned}
$$

For the last step we have exploited the relations of corollary 6.1. Indeed, the transformation leads to a simplification, as each of the four blocks is diagonal. Since the determinant and hence the characteristic polynomial are invariant under similarity transformations we obtain

$$
\chi_{\mathbf{E}}(\lambda) \;=\; \chi_{\tilde{\mathbf{F}}\mathbf{E}\tilde{\mathbf{F}}^{-1}}(\lambda) \;=\; \det\left[\begin{pmatrix} \alpha\mathbf{W} & \beta\mathbf{W} \\ \gamma\overline{\mathbf{W}} & \delta\overline{\mathbf{W}} \end{pmatrix} - \lambda \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix} \right] .
$$

Furthermore the determinant does not alter if lines or columns are permuted. It is reasonable to rearrange the matrix, such that it gets blockdiagonal with N different 2×2-blocks, as indicated:

$$
\left(\begin{array}{ccc|ccc}
\alpha w^0 & \cdot & \cdot & \beta w^0 & \cdot & \cdot \\
\cdot & \ddots & \cdot & \cdot & \ddots & \cdot \\
\cdot & \cdot & \alpha w^{N-1} & \cdot & \cdot & \beta w^{N-1} \\
\hline
\gamma \overline{w}^0 & \cdot & \cdot & \delta \overline{w}^0 & \cdot & \cdot \\
\cdot & \ddots & \cdot & \cdot & \ddots & \cdot \\
\cdot & \cdot & \gamma \overline{w}^{N-1} & \cdot & \cdot & \delta \overline{w}^{N-1}
\end{array} \right)
\;\longrightarrow\;
\begin{pmatrix}
\boxed{\begin{matrix} \alpha w^0 & \beta w^0 \\ \gamma \overline{w}^0 & \delta \overline{w}^0 \end{matrix}} & \cdot & \cdot \\
\cdot & \ddots & \cdot \\
\cdot & \cdot & \boxed{\begin{matrix} \alpha w^{N-1} & \beta w^{N-1} \\ \gamma \overline{w}^{N-1} & \delta \overline{w}^{N-1} \end{matrix}}
\end{pmatrix}
$$

Since the determinant of a blockdiagonal matrix is equal to the product of the determinants associated to each block, we obtain

$$
\chi_{\mathbf{E}}(\lambda) \;=\; \prod_{m=0}^{N-1} \Big[(\alpha w^m - \lambda)(\delta \overline{w}^m - \lambda) - \beta w^m \gamma \overline{w}^m \Big].
$$

Expanding the product in rectangular braces and employing the properties (6.1) of the unit roots yields the asserted formula. ■

The explicit knowledge of the characteristic polynomial and rather more its decomposition into quadratic polynomials, whose zeros can be found analytically, permit us to study the spectrum in detail. Therefore the computation of the spectrum enables us to discover further features of the lattice-Boltzmann algorithm.

Obviously, the quadratic factors in (6.12) agree in their structure and differ only in the powers w^m of the unit root. For an analysis of these polynomials, it makes

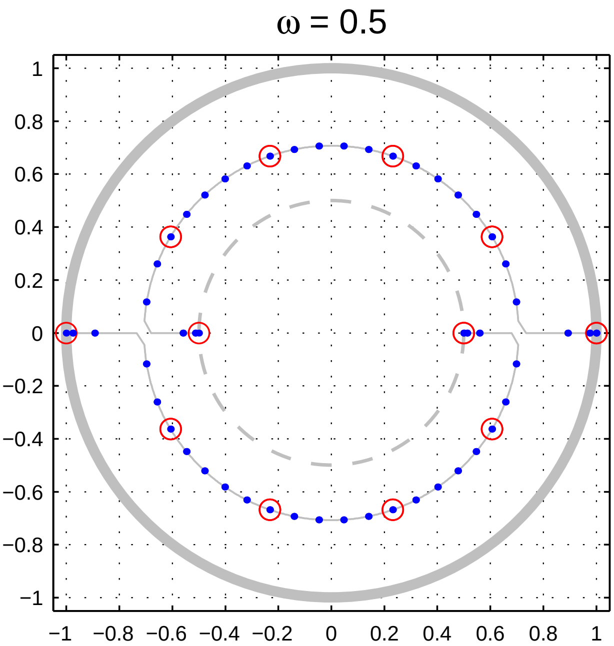


Figure 6.2: Plot of $\mathrm{spec}(E, 0.5, 0, 0, 10)$ (red circles) and $\mathrm{spec}(\mathbf{E}, 0.5, 0, 0, 50)$ (blue dots) illustrating the statement of corollary 6.2. The fine solid curve corresponds to the spectral limit set, that is always a line-like subset of the closed area between the circles of radius 1 (thick) and $|1-\omega|$ (dashed) if $0 \le \omega \le 2$.

sense to replace w^m by an arbitrary number $\mathrm{e}^{\mathrm{i}\phi}$ ($\phi \in [0, 2\pi)$) on the unit circle. So we define

$$\begin{aligned}\chi_\phi(\lambda) \;&:=\; \lambda^2 \;-\; (\alpha \mathrm{e}^{\mathrm{i}\phi} + \delta \mathrm{e}^{-\mathrm{i}\phi})\lambda \;+\; \alpha\delta - \beta\gamma \\ &=\; \lambda^2 \;+\; \left[(\omega + ch^2 - 2)\cos(\phi) + \mathrm{i}\omega a h \sin(\phi)\right]\lambda \;+\; (1-\omega)(1-ch^2).\end{aligned}$$

As the macroscopic quantities of the equilibrium, a and c, occur only with h and h^2 respectively, it will be advantageous to substitute them by new variables $r := ah$ and $p := ch^2$. Then the quadratic polynomial χ_ϕ becomes

$$\chi_\phi(\lambda) = \lambda^2 \;+\; \left[(\omega + p - 2)\cos(\phi) + \mathrm{i}\omega r \sin(\phi)\right]\lambda \;+\; (1-\omega)(1-p). \tag{6.13}$$

The parameter r has a special meaning: remembering that $h^2 = \Delta t$ corresponds to the time step, while h is just the spatial grid constant Δx, we see that the grid speed is given by $\Delta x / \Delta t = h^{-1}$. Thus $r = ah = a/h^{-1}$ can be considered as the *ratio* of the advection velocity and the grid speed. In allusion to the theory of finite differences, this quotient is referred to as **CFL number**.
It should be kept in mind that r and p shrink to 0 under grid refinement, if a and c are held fixed.

Let us abbreviate by $\mathrm{spec}(\mathbf{E}, \omega, r, p, N) \subset \mathbb{C}$ the spectrum of $\mathbf{E}$ associated with given values for ω, r and p and a grid of N nodes.

Corollary 6.2. *The spectrum of* $\mathbf{E}$ *admits the subsequent representation:*

$$\operatorname{spec}(\mathbf{E},\omega,r,p,N) = \big\{ z \in \mathbb{C} |\ \exists\, m \in \{0,\dots,N-1\}\ \ s.th.\ \ \chi_{\frac{2\pi m}{N}}(z) = 0\ \big\}.$$

In particular we have for $k \in \mathbb{N}$*:* $\operatorname{spec}(\mathbf{E},\omega,r,p,N) \subset \operatorname{spec}(\mathbf{E},\omega,r,p,kN)$

Proof: The first part is an immediate consequence of theorem (6.2) since $\chi_{\mathbf{E}}(z) = \prod_{m=0}^{N-1} \chi_{\frac{2\pi m}{N}}(z)$. The inclusion relation follows from the simple fact that $\{\frac{2\pi m}{N} |\ m = 0, 1, ..., N-1\} \subset \{\frac{2\pi m}{kN} |\ m = 0, 1, ..., kN-1\}$. ■

In the sequel we want to deal with the closed union of all eigenvalues pertaining to some $\mathbf{E}$ for fixed ω, r and p but variable grid size:

$$\begin{aligned} \mathfrak{S}(\omega,r,p) &:= \overline{\bigcup_{N\in\mathbb{N}} \operatorname{spec}(\mathbf{E},\omega,r,p,N)} \\ &\equiv \overline{\big\{ z \in \mathbb{C}\,|\,\exists N \in \mathbb{N}\ \text{ s.th. }\ z \in \operatorname{spec}(\mathbf{E},\omega,r,p,N) \big\}}, \end{aligned}$$

We shall call this set the *spectral limit set*, which is remarkably not a messy amorphous cloud of points in the complex plane. Instead, $\mathfrak{S}(\omega,r,p)$ reveals a clear structure, which might be foreshadowed by the inclusion property of corollary (6.2). In fact, all eigenvalues are located on a set that can be described by a few smooth curves. This becomes even more comprehensible by proposition 6.3, characterizing the spectral limit set as the zero level set of the parameterized polynomial χ_ϕ. For the proof the subsequent lemma is useful.

Lemma 6.6. *Let* X *and* Y *be two metric spaces, where* X *is assumed to be* compact*. Denote the projection onto* Y *by* p_2*, i.e.* $p_2 : X \times Y \to Y$ *with* $p_2\big((x,y)\big) := y$*. Then*

$$A \subset X \times Y \ \textit{closed} \qquad \Rightarrow \qquad p_2(A) \subset Y \ \textit{closed.}$$

Proof: It is to be shown that $y \in \overline{p_2(A)}$ already implies $y \in p_2(A)$. As y is supposed to be in the closure of $p_2(A)$, there exists a sequence $(y_n)_n \subset p_2(A)$ with $y_n \xrightarrow{n\to\infty} y$. For each y_n exists an $x_n \in X$ with $(x_n, y_n) \in A$. Since X is compact, the sequence $(x_n)_n$ has a convergent subsequence $(x_{n_k})_k$ with $x_{n_k} \xrightarrow{k\to\infty} x \in X$. As every subsequence of a convergent sequence is convergent, we have $y_{n_k} \xrightarrow{k\to\infty} y$. Hence $A \ni (x_{n_k}, y_{n_k}) \xrightarrow{k\to\infty} (x,y)$ in the product topology of $X \times Y$. So we can conclude $(x,y) \in A$ because A is assumed to be closed in the product topology. But this finally implies that y belongs to the image of A, i.e. $y \in P_2(A)$ as $p_2\big((x,y)\big) = y$. ■

The compactness of X is an essential hypothesis[11]; in general the canonical projections of a closed set in a product space need not to be closed.

Proposition 6.3. *The* spectral limit set *is alternatively given by*

$$\mathfrak{S}(\omega,r,p) = \big\{ z \in \mathbb{C} |\ \exists \phi \in [0, 2\pi)\ \ s.th.\ \ \chi_\phi(z) = 0 \big\} =: \mathcal{Z}.$$

[11] The statement of the lemma may be wrong if the hypothesis is not given. Example: $A := \{(x,y)|\ y \geq 1/x \ \text{ with } \ x, y \in (0,\infty)\}$ is closed in $\mathbb{R}\times\mathbb{R}$ but $p_1(A) = p_2(A) = (0,\infty)$ is not closed in $\mathbb{R}$.

Proof: The assertion is verified along with three steps:

i) $\mathcal{Z}$ is closed.

ii) $\forall N \in \mathbb{N}:\ \mathrm{spec}(\mathbf{E}, \omega, r, p, N) \subset \mathcal{Z}$.

iii) $\bigcup_{N\in\mathbb{N}} \mathrm{spec}(\mathbf{E}, \omega, r, p, N)$ is dense in $\mathcal{Z}$.

Since the function $\phi \mapsto \chi_\phi(z)$ is 2π-periodic, the set $\mathcal{Z}$ is not changed[12] if we permit $\phi \in [0, 2\pi]$ instead of $\phi \in [0, 2\pi)$. Let us introduce the mapping $f : [0, 2\pi] \times \mathbb{C} \to \mathbb{C}$ defined by $f(\phi, z) = \chi_\phi(z)$. Clearly, f is continuous with respect to the standard product topology in $[0, 2\pi] \times \mathbb{C}$. Therefore $f^{-1}(0)$ is closed in $[0, 2\pi] \times \mathbb{C}$. If $p_2 : [0, 2\pi] \times \mathbb{C} \to \mathbb{C}$ denotes the canonical projection, lemma 6.6 entails the closedness of the set $p_2\big(f^{-1}(0)\big) = \mathcal{Z}$, as $[0, 2\pi]$ is compact.
Obviously, we have $\mathrm{spec}(\mathbf{E}, \omega, r, p, N) \subset \mathcal{Z}$ for every $N \in \mathbb{N}$. So $\mathcal{Z}$ contains the union of all spectra. As $\mathcal{Z}$ is closed, it comprises also the closure of the union, i.e. $\mathfrak{S}(\omega, r, p)$, because the closure of a set is the smallest closed superset.
Let $\tilde{z} \in \mathcal{Z}$. Then there exists $\tilde{\phi} \in [0, 2\pi]$ with $\chi_{\tilde{\phi}}(\tilde{z}) = 0$. Since $U := \bigcup_{N\in\mathbb{N}}\{\frac{2\pi m}{N} | m = 0, 1, \dots, N-1\}$ is dense in $[0, 2\pi]$ there is a sequence $(\phi_n)_n$ in U converging to $\tilde{\phi}$. Now, the zeros of a polynomial depend continuously[13] on its coefficients. Therefore it exists a sequence $(z_n)_n$ with

$$\chi_{\phi_n}(z_n) = 0 \qquad \text{and} \qquad z_n \xrightarrow{n\to\infty} \tilde{z},$$

since the coefficients of the polynomials χ_{ϕ_n} converge to those of the polynomial $\chi_{\tilde{\phi}}$. Of course, we have $(z_n)_n \subset \mathfrak{S}(\omega, r, p)$. This shows that arbitrary elements of $\mathcal{Z}$ are limits of sequences in $\mathfrak{S}(\omega, r, p)$, which proves the denseness. ■

Geometric shape of $\mathfrak{S}(\omega, 0, 0)$ and multiplicity of eigenvalues

In the case of pure diffusion ($r = 0, p = 0$) the spectral limit set becomes particularly simple; the next proposition delivers a geometric description which is to be compared with figure 6.3.

Proposition 6.4. *If $r = 0$ and $p = 0$, the spectral limit set is given by*

i) $0 \le \omega \le 1 : \quad \mathfrak{S}(\omega, 0, 0) = [-1, \omega - 1] \cup [1 - \omega, 1] \cup \sqrt{1-\omega}\, S_1(0)$

ii) $1 \le \omega \le 2 : \quad \mathfrak{S}(\omega, 0, 0) = [-1, 1 - \omega] \cup [\omega - 1, 1]$

The statement is a consequence of the following observations. In the diffusive case the polynomial χ_ϕ reduces to

$$\chi_\phi(\lambda) = \lambda^2 \ + \ (\omega - 2)\cos(\phi)\lambda \ + \ (1 - \omega).$$

[12]Since χ_ϕ depends on ϕ via $e^{i\phi}$, it is possible to replace the angular parameter $\phi \in [0, 2\pi)$ by a complex-valued parameter belonging to the unit circle, which forms a compact metric space. Observe that there is no diffeomorphism (bijective, continuous and continuously invertible) between the unit circle and a compact interval.

[13]Here it is important to consider complex-valued roots (see also zero-counting formula). In contrast, the number (and position) of real roots may depend discontinuously on the coefficients.

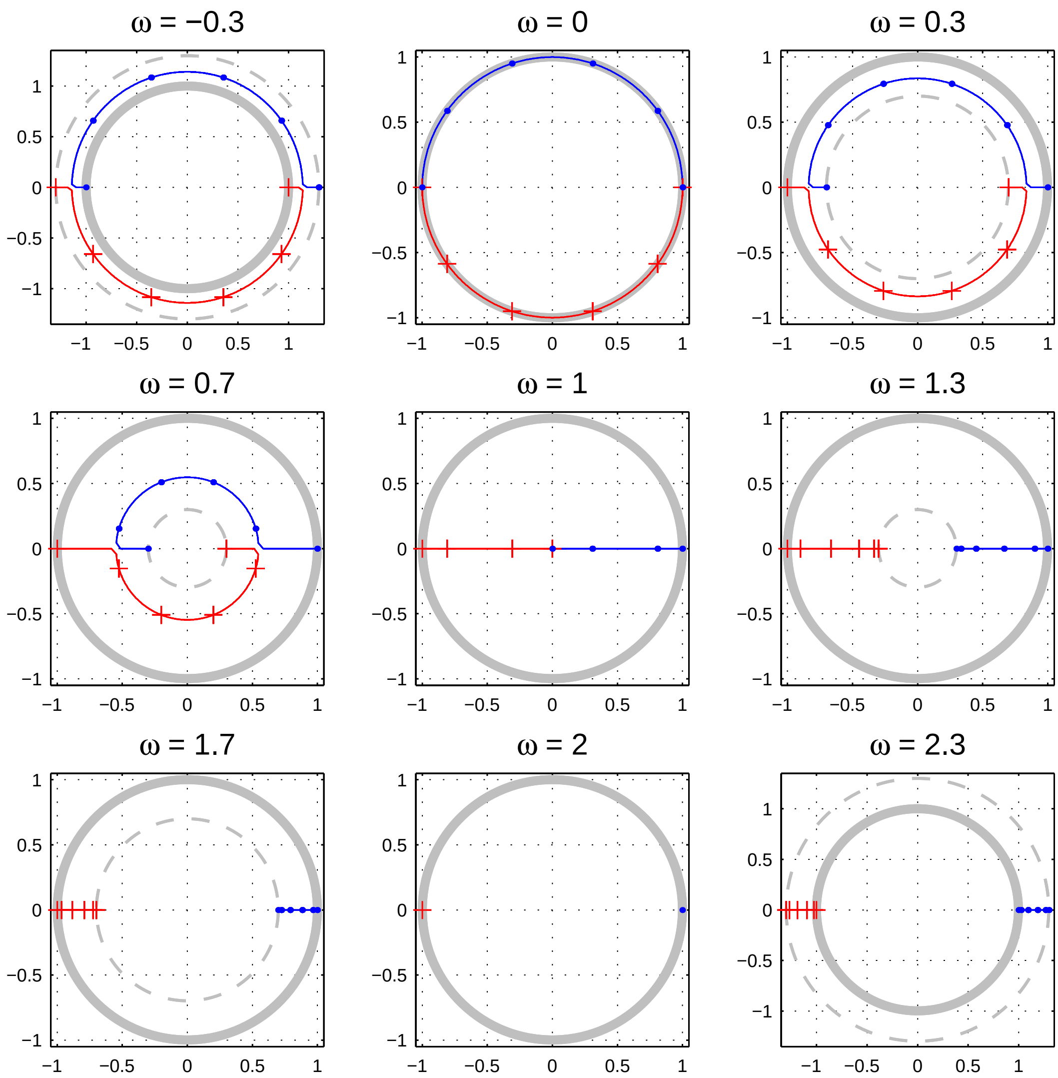


Figure 6.3: These plots illustrate the spectral limit set for the *purely diffusive* case, which is considered in proposition 6.4. The spectral limit set is plotted for values of ω both inside and outside of the stability interval $[0, 2]$ in order to demonstrate that the scheme may become unstable. It is striking that $\mathfrak{S}(\omega, 0, 0)$ always lies in the closed region between the unit circle (thick gray line) and the circle of radius $|1 - \omega|$ (dashed gray line). The dots and crosses indicate the eigenvalues for $N = 10$. Furthermore they help to distinguish both of the branches: λ_- (blue line with dots) and λ_+ (red line with crosses). Square roots of negative radicands are evaluated taking the principal value. Observe that the so obtained branches do not necessarily agree with the natural branches (especially if $\omega \in [0, 1)$) discussed in the next paragraph.

Using the standard formula for solving quadratic equations we get

$$\lambda_{\pm}(\phi) \quad = \quad (1-\tfrac{1}{2}\omega)\cos(\phi) \pm \sqrt{\tfrac{1}{4}(\omega-2)^2\cos^2(\phi)+\omega-1}. \tag{6.14}$$

ad i) For $0<\omega<1$ the radicand can be negative depending on ϕ. More exactly, there is a $\phi_0=\phi_0(\omega)\in(0,\frac{\pi}{2})$ such that the radicand vanishes. As $\cos(\phi_0)$ must be positive we obtain

$$\tfrac{1}{4}(\omega-2)^2\cos^2(\phi_0)+\omega-1=0 \qquad \Leftrightarrow \qquad \cos(\phi_0)=\frac{2}{2-\omega}\sqrt{1-\omega}\,.$$

The radicand in (6.14) becomes non-negative if $\phi\in[0,\phi_0]\cup[\pi-\phi_0,\pi+\phi_0]\cup[2\pi-\phi_0,2\pi]$ and hence λ_- and λ_+ are real. For $\phi\in(\phi_0,\pi-\phi_0)\cup(\pi+\phi_0,2\pi-\phi_0)$ the radicand is negative and the root gets imaginary. In this case λ_- and λ_+ are complex-valued and lie on a circle with radius $\sqrt{1-\omega}$, because

$$\begin{aligned}|\lambda_{\pm}(\phi)|^2 &= \left|(1-\tfrac{1}{2}\omega)\cos(\phi)\pm \mathrm{i}\sqrt{1-\omega-\tfrac{1}{4}(\omega-2)^2\cos^2(\phi)}\right|^2\\ &= \left[(1-\tfrac{1}{2}\omega)^2\cos^2(\phi)\right]+\left[1-\omega-\tfrac{1}{4}(\omega-2)^2\cos^2(\phi)\right]=1-\omega.\end{aligned}$$

ϕ	0	ϕ_0	$\frac{\pi}{2}$	$\pi-\phi_0$	π	$\pi+\phi_0$	$\frac{3\pi}{2}$	$2\pi-\phi_0$
$\lambda_-(\phi)$	$1-\omega$	$\sqrt{1-\omega}$	$-\mathrm{i}\sqrt{1-\omega}$	$-\sqrt{1-\omega}$	-1	$-\sqrt{1-\omega}$	$\mathrm{i}\sqrt{1-\omega}$	$\sqrt{1-\omega}$
$\lambda_+(\phi)$	1	$\sqrt{1-\omega}$	$\mathrm{i}\sqrt{1-\omega}$	$-\sqrt{1-\omega}$	$\omega-1$	$-\sqrt{1-\omega}$	$\mathrm{i}\sqrt{1-\omega}$	$\sqrt{1-\omega}$
	$\in\mathbb{R}^+$			$\in\mathbb{R}^-$				$\in\mathbb{R}^+$
		$\in\sqrt{1-\omega}S_1(0)$ (circle)				$\in\sqrt{1-\omega}S_1(0)$ (circle)		

The table lists the values of λ_+ and λ_- for selected values of ϕ.

ad ii) In the case of $1\le\omega\le2$ the radicand is nonnegative no matter how ϕ is chosen. Therefore the spectral limit set is real.

ϕ	0	$\frac{\pi}{2}$	π	$\frac{3\pi}{2}$	
$\lambda_-(\phi)$	$1-\omega$	$-\sqrt{\omega-1}$	-1	$-\sqrt{\omega-1}$	$\in\mathbb{R}^-$
$\lambda_+(\phi)$	1	$\sqrt{\omega-1}$	$\omega-1$	$\sqrt{\omega-1}$	$\in\mathbb{R}^+$

Both of the tables unveil, that each of the parameterizations $\phi\mapsto\lambda_-(\phi)$ and $\phi\mapsto\lambda_+(\phi)$ runs twice through a curve: first in forward direction if $\phi\in[0,\pi)$ and then backwards for $\phi\in[\pi,2\pi)$. This insinuates that all eigenvalues appear with double algebraic multiplicity except of the starting point ($\phi=0$) and the returning point ($\phi=\pi$), which are only passed once. In the case of an odd number of grid nodes, the returning point is not hit, because there is no choice of $m\in\{0,1,\ldots,N-1\}$, such that $2\pi\frac{m}{N}$ is equal to π if N is not even. Let us summarize the reasoning:

Corollary 6.3. *In the purely diffusive case the eigenvalues occur with the subsequent algebraic multiplicities depending partly on the number of grid nodes N:*

eigenvalue	-1	1	$1-\omega$	$\omega-1$	other
N odd	not	simple	simple	not	two-fold
N even	simple	simple	simple	simple	two-fold

As the evolution matrix $\mathbf{E}$ reduces to the collision block (confer lemma 6.3) for a single node grid ($N = 1$), the starting points $\lambda_-(0) = 1-\omega$ and $\lambda_+(0) = 1$ turn out to be the eigenvalues of the collision block. Similarly, the spectrum of $\mathbf{E}$ referring to a grid with just two nodes ($N = 2$) is composed of the starting and returning points.

In order to verify that $\mathbf{E}$ is diagonalizable (and admits a basis of eigenvectors) it is sufficient to check that the 2×2 blocks

$$\begin{pmatrix} \alpha w^m & \beta w^m \\ \gamma \overline{w}^m & \delta \overline{w}^m \end{pmatrix},$$

which appeared in the proof of theorem 6.2, are diagonalizable. This would also entail that all eigenvalues of $\mathbf{E}$ must have maximal geometric multiplicity. It should be noted that the two-fold degeneracy of the eigenvalues originates from a property of the diffusion operator with periodic boundary conditions[14] (see appendix).

Further observations in the general case

The degeneracy is broken, as soon as the advection velocity a (and hence the CFL number r) does not vanish. Obviously, the symmetry between both directions is destroyed by selecting a direction of propagation for the advection process. Figure 6.5 and 6.6 display, how the two-fold eigenvalues of the purely diffusive case split into a couple of eigenvalues if a gets different from 0. Moreover, it becomes clearly visible, how the spectral limit set develops gradually into a warped curve, that cannot be described any more by elementary line segments and circles.

Generally, the spectral limit set is computed as zero-level set of (6.13), which yields

$$\begin{aligned} \lambda_\pm(\phi) := &-\tfrac{1}{2}\big[(\omega+p-2)\cos(\phi) + \mathrm{i}\omega r \sin(\phi)\big] \\ &\pm \sqrt{\tfrac{1}{4}\big[(\omega+p-2)\cos(\phi) + \mathrm{i}\omega r \sin(\phi)\big]^2 - (1-\omega)(1-p)}. \end{aligned} \tag{6.15}$$

$\phi \mapsto \lambda_-(\phi)$ and $\phi \mapsto \lambda_+(\phi)$ with $\phi \in [0, 2\pi)$ are regarded as parameterizations tracing the spectral limit set. The formula permits two direct observations:

- The spectral limit set does not change if the sign of r is switched[15]. This reflects the *isotropy* of the lattice-Boltzmann algorithm, which has *a priori* no preferred direction (left/right). Hence a change of sign makes the parameterization $\lambda_\pm$ just run in the reverse sense.

[14] Similarly to the lattice-Boltzmann evolution matrix, the diffusion operator (second order derivative) $-\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ admits two eigenfunctions on the space of periodic $\mathcal{C}^2$-functions for every eigenvalue different from zero (scaled *sine* and *cosine* function).

[15] Observe that r and $\sin(\phi)$ occur in (6.15) only together as product. The assertion results from the relations $-r\sin(\phi) = r\sin(-\phi)$ and $\cos(\phi) = \cos(-\phi)$.

- The radicand in (6.15) is not only 2π- but π-periodic[16]. So the spectral limit set corresponds geometrically to an ellipse (terms in first line of (6.15)) superposed by the "root" of a rotated ellipse (root term), which is traversed twice as ϕ goes from 0 to 2π.

Natural branches and evaluation of the complex-valued root. For $a \neq 0$, figure 6.5 and 6.6 show that the spectral limit set is composed of two closed curves, that we denote as *natural branches*. Evidently, each branch should be associated with one of the two parameterizations λ_- and λ_+. However, there is some caution demanded, since the square root of complex numbers is multi-valued. Evaluating the root by the principal value[17], the curves may appear with spurious discontinuities. In order to obtain the natural branches, as highlighted in figure 6.5 and 6.6, it is necessary to compute the square root in a continuous way when the spectral limit set is plotted by sampling the parameterizations λ_- and λ_+.
Otherwise, if the square-root of the ϕ-parameterized radicand is evaluated staying always on the same *leaf* of the underlying *Riemannian surface*, its value may perform erratic-looking leaps in the complex plane. This entails that λ_- and λ_+ may artificially jump between the natural branches, whenever the radicand oversteps the line of discontinuity associated with the square root.
The shortcoming is remedied if both square roots are calculated *a priori* in order to choose the value extending the previous values for λ_- (or respectively for λ_+) "continuously", i.e. with the least difference. Alternatively, one can try to slit the complex plane along a tilted ray which is then used as line of discontinuity instead of the positive or negative real axis.

Sink and source points. Moving along the spectral limit set, one may ask how the eigenvectors change with the eigenvalues.
Eigenvectors with constant populations belong to the so-called *sink points* $\lambda_+(0) = 1$ and $\lambda_-(0) = 1 - \omega$ being just the eigenvalues of the collision block. Each sink point is accompanied by a *source point*[18]. Both of them lie on the real line where the source point is mostly symmetrically situated to its corresponding sink point with regard to the imaginary axis (see exception in figure 6.6).
Eigenvectors of high spatial frequency (shortwave modes) refer to the eigenvalues near the source points, while those with low spatial frequency (longwave modes) belong to eigenvalues around the sink points. Hopping along a branch from eigenvalue to eigenvalue towards the sink point, the spatial oscillations of the associated eigenvectors decreases each time by 1 (see figure 6.4). When the number of grid points is increased, it seems as if new eigenvalues are generated near the source points where eigenvectors reveal higher spatial frequencies than existing before. The other eigen-

[16] It should be noticed, that the square of a linear combination of the sine and cosine function is π-periodic, since $\sin^2(x) = \frac{1}{2}(1 - \cos(2x))$ and $\cos^2(x) = 1 - \sin^2(x)$ are π-periodic as well as the mixed term $\cos(x)\sin(x) = \frac{1}{2}\sin(2x)$.

[17] Any complex number $z \in \mathbb{C}$ can be represented in the form $z = re^{\mathrm{i}\theta}$ with $|z| = r > 0$ and $-\pi < \theta \leq \pi$. The *principal value* of $\sqrt{z}$ is usually defined by $\sqrt{r}e^{\mathrm{i}\theta/2}$.

[18] The source points are the additional eigenvalues of the evolution matrix pertaining to a two-node grid. They occur as eigenvalues if the number of grid nodes is even.

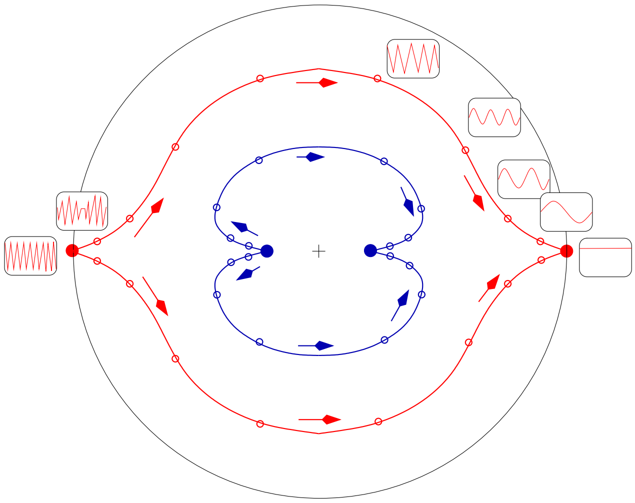


Figure 6.4: Schematic picture representing the spectral limit set in the complex plane for the underrelaxed case $\omega < 1$. The dots indicate the positions of eigenvalues that belong to an evolution matrix referring to a grid of $N = 18$ nodes. Red and blue coloring highlight the natural branches. Their associated sink and source points (being also eigenvalues) are marked by fat filled dots; the arrows are orientated in direction of the sink points. The small boxes impart an impression how the eigenvectors typically look like. As the components of the eigenvectors are complex-valued, the more or less oscillating lines inside the boxes can only display the real or imaginary part, which, furthermore, refers only to a single population. The other population looks quite similar. Moreover, the eigenvectors of the lower half branch and of the blue branch behave qualitatively in the same manner. Eigenvectors with the same spatial frequency distinguish in the ratio of the complex-valued amplitudes for each population (cf. computation of eigenvectors on page 257.)

values are apparently shifted in direction of the sink points[19]. The eigenvectors of **E** are computed at the end of the next subsection.

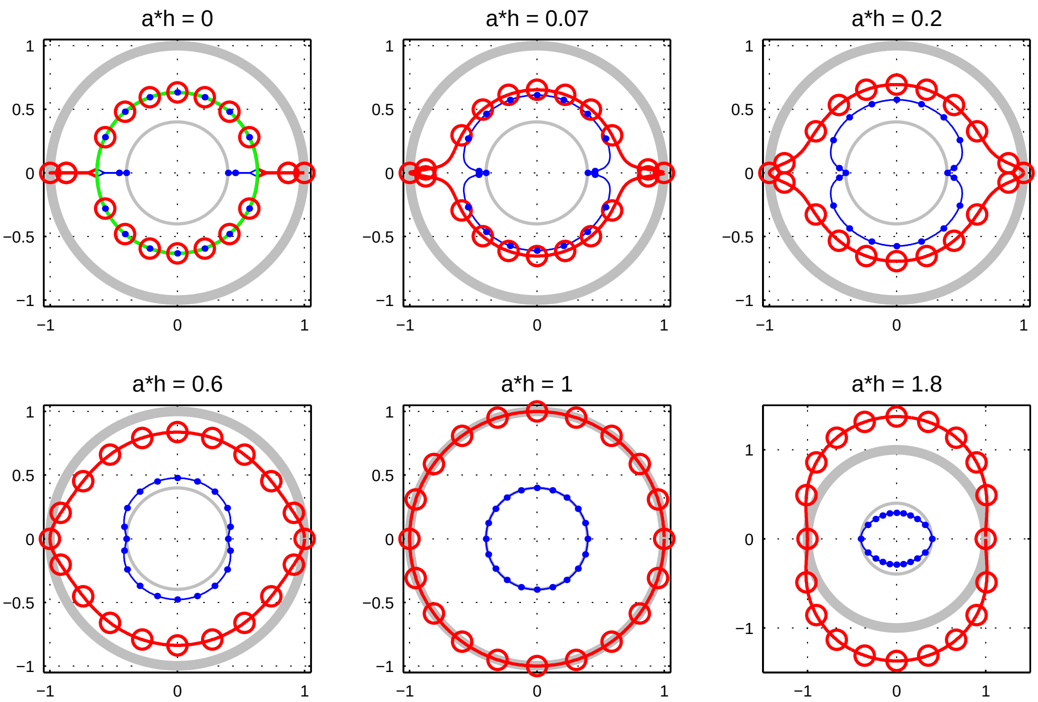


Figure 6.5: Sequence of spectral limit sets: underrelaxed case with $\omega = 0.6$. The radius of the small circle is $|1-\omega|$ (also in figure 6.6). For $a = 0$ the two branches coincide on a circle of radius $\sqrt{1-\omega}$. The markers indicate eigenvalues associated with a grid of 20 nodes. Dots mark the λ_--branch while circles indicate the λ_+-branch (also below).

"Spectral gallery". The following figures portray how the spectral limit set behaves if ω and $r = ah$ vary. In particular, one may perceive, that the eigenvalues are not uniformly (equidistantly) distributed on the curves for fixed N but sit more densely packed in some segments than in others.
Furthermore, we realize that the branches distinguished in figure 6.3 for $\omega = 0.3$ or $\omega = 0.7$ are different from the natural branches marked in figure 6.5 for $ah = 0$ (pure diffusion), because the complex root has been evaluated differently. In fact, for $r = 0$, the natural branches cover each other on the circular part of $\mathfrak{S}(\omega, 0, 0)$ (figure 6.5), where the dots fill just the centers of the little circles. As $r = ah$ starts moving away from 0 the branches deform continuously. So the inner branch associated with λ_- contracts around $(1-\omega)S_1(0)$, while the outer branch pertaining to λ_+ inflates towards the unit circle and beyond.
In the overrelaxed case (figure 6.6), the branches face each other symmetrically until a critical value of $|r|$ is reached provoking a sudden change of topology.
As long as $|r| \leq 1$ and $\omega \in [0, 2]$, the spectral limit set does not leave the closed area between the unit circle (bold gray) and the circle of radius $|1-\omega|$, whereas it does not intersect the open domain between the circles if $|r| > 1$. In particular the

[19]Note that our labeling, *sink* and *source* point, shall reflect these observations.

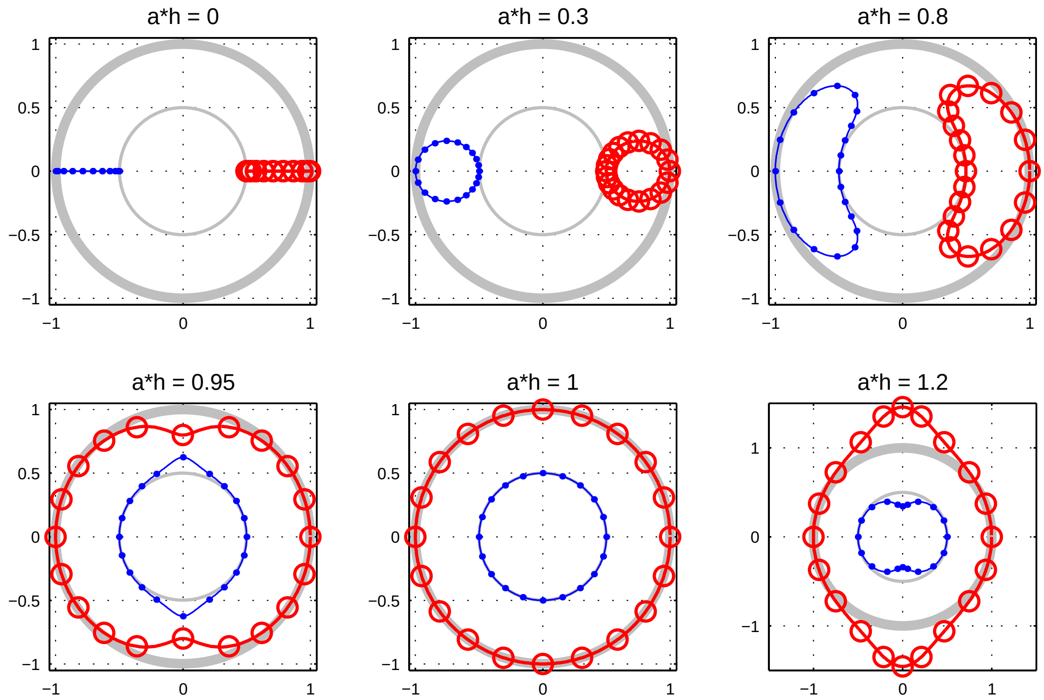


Figure 6.6: Sequence of spectral limit sets: overrelaxed case with $\omega = 1.5$. In the plots of the upper row, sink- and source-points belonging to the same branch are not symmetric with respect to the imaginary axis.

spectral limit set is not completely contained in the closed unit disk if $|r| > 1$ (see section 6.3.1).

6.2.3 Asymptotics and symmetry of eigenvalues

Macroscopic interpretation of eigenvalues near 1

By the asymptotic analysis in chapter 4, we have established a relation between the lattice-Boltzmann algorithm[20] (6.4) using the (6.6) and the following IBVP:

$$\left.\begin{array}{ll} \text{EQ:} & \partial_t u + a\partial_x u - \nu\partial_x^2 u + cu = 0 \quad \text{with} \quad \nu = \frac{1}{\omega} - \frac{1}{2}, \\ \text{BC:} & u(t,0) = u(t,L), \\ \text{IC :} & u(0,\cdot) = u_0. \end{array}\right\} \tag{6.16}$$

The spectral analysis opens the opportunity to understand this connection from another side. In particular, we want to investigate, how the eigenvalues of the lattice-Boltzmann evolution matrix $\mathbf{E}$ are related to (6.16). Thus, we should look for the evolution operator of the IBVP. Formally (6.16,EQ) can be written as ODE:

$$\tfrac{\mathrm{d}}{\mathrm{d}t}u + Au = 0 \quad \text{with} \quad A := -\nu\partial_x^2 + a\partial_x + c$$

To grasp the idea, let us proceed in a sloppy way ignoring technical details as the domain of definition of A etc.. In analogy to linear ODE systems in $\mathbb{R}^n$, the evolution

[20]The D1P2 algorithm considered here represents a special case of the D1P3 algorithm studied in chapter 4.

operator of (6.16) becomes the exponential[21] e^{-tA}. Following the analogy farther, we recognize that the eigenfunctions of the evolution operator correspond to the eigenfunctions of the operator A itself. Additionally we can infer

$$Av = \lambda v \qquad \Rightarrow \qquad e^{-tA} v = e^{-t\lambda} v \ .$$

So we have to find the eigenvalues of the operator A in order to get those of the evolution operator e^{-tA}. For this we take recourse to the appendix. It must be underlined, that the eigenvalues of A depend also on the function space the operator A is defined on. This function space usually incorporates the boundary conditions that may crucially influence the eigenvalues (see appendix for an example).
According to theorem A.2 we get for every $n \in \mathbb{Z}$ an eigenvalue

$$\lambda_n = 4\pi^2 n^2 \nu + 2i\pi n a + c \tag{6.17}$$

of the differential operator A applied to the space of (sufficiently smooth) *periodic* functions. Hence the eigenvalues of the evolution operator e^{-tA} are $e^{-\lambda_n t}, n \in \mathbb{Z}$.

Remark 6.2. Product structure of eigenfunctions. Actually, the eigenvalues and eigenfunctions of the evolution operator e^{-tA} are calculated by the method of *separation of variables* due to the following reason: Let $v = v(x)$ be an eigenfunction of e^{-tA} and let $u(t, x)$ be the solution of (6.16) with $u_0 = v$. The fact that v is an eigenfunction of the evolution operator implies, that v and $u(t, \cdot)$ are *linear dependent* as functions of the spatial variable x. Therefore there is a 'scalar' g depending on the time t, such that $u(t, x) = g(t)v(x)$. Thus it turns out, that the search for eigenfunctions of the evolution operator is equivalent with the quest for solutions of (6.16), that have the above product form.

Since the application of the lattice-Boltzmann evolution operator $\mathbf{E}$ corresponds to a time step $\Delta t = h^2$, $\mathbf{E}$ is to be compared with e^{-Ah^2}. We know that the lattice-Boltzmann algorithm generates a sequence of population grid functions ($h \downarrow 0$), whose mass moment converges to the solution of (6.16).
This circumstance should have a strong impact on the eigenvalues of $\mathbf{E}$ and e^{-Ah^2}. Figure 6.7 and 6.8 show some of the eigenvalues of $\mathbf{E}$ and e^{-Ah^2} near the common eigenvalue 1. Each figure contains actually several plots for a growing number N of grid nodes and a decreasing grid spacing h respectively. As a and not r remains fixed, the shown part of the spectral limit sets gets flatter when h is diminished. Similarly, the red curves interpolating the eigenvalues of e^{-Ah^2} become more tight for $h \downarrow 0$. However, the most striking feature is, that these curves fit the spectral limit sets near 1 the better the more h is reduced. Moreover, it is clearly visible that circles and dots seem to coincide the better the more they approach 1 and the smaller h is chosen. The observations are made more precise quantitatively by the following proposition.

Proposition 6.5. *Let λ_n and λ_+ as given in (6.17) and (6.15) respectively. Then*

[21] Note that this can be strictly defined in the framework of *strongly continuous semigroups*; but is of no relevance for the objective here.

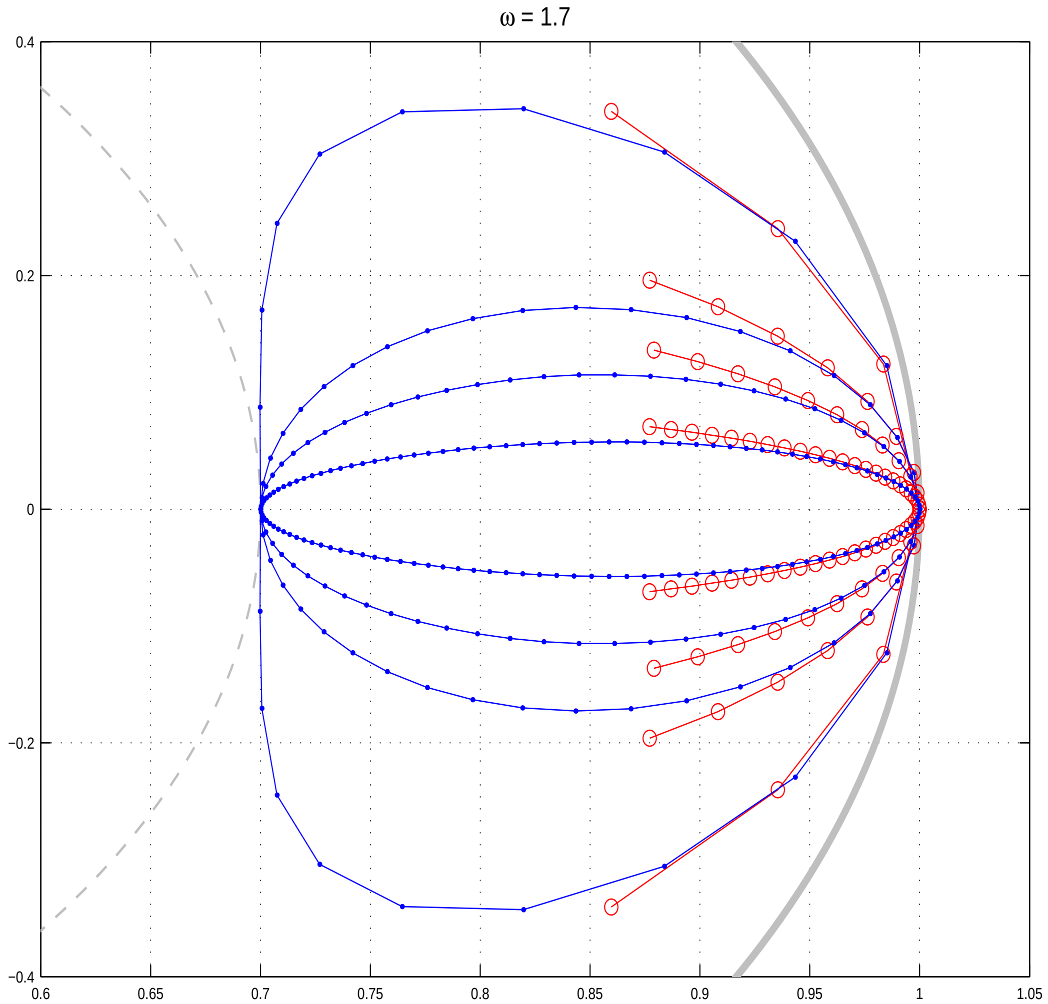


Figure 6.7: The figure shows qualitatively, how the eigenvalues of E (blue dots) pertaining to the right branch of the spectral limit set approximate those of e^{-Ah^2} (red circles) in the overrelaxed advective-diffusive case. Several curves are plotted corresponding to varying values of $h = \frac{L}{N}$. Observe that here the advection velocity a (and *not* the CFL-number r) is kept constant.

the following asymptotic similarities hold true for $h \in \mathcal{H}_L$:

$$\begin{aligned} &i) && \left|\mathrm{e}^{-\lambda_n h^2} - 1\right| = \mathrm{O}(h^2) \\ &ii) && \left|\lambda_+(2\pi n h) - 1\right| = \mathrm{O}(h^2) \\ &iii) && \left|\lambda_+(2\pi n h) - \mathrm{e}^{-\lambda_n h^2}\right| = \mathrm{O}(h^4) \end{aligned}$$

The relations, in particular ii) and iii), have been verified performing the necessary Taylor expansions using the computer algebra package MAPLE.

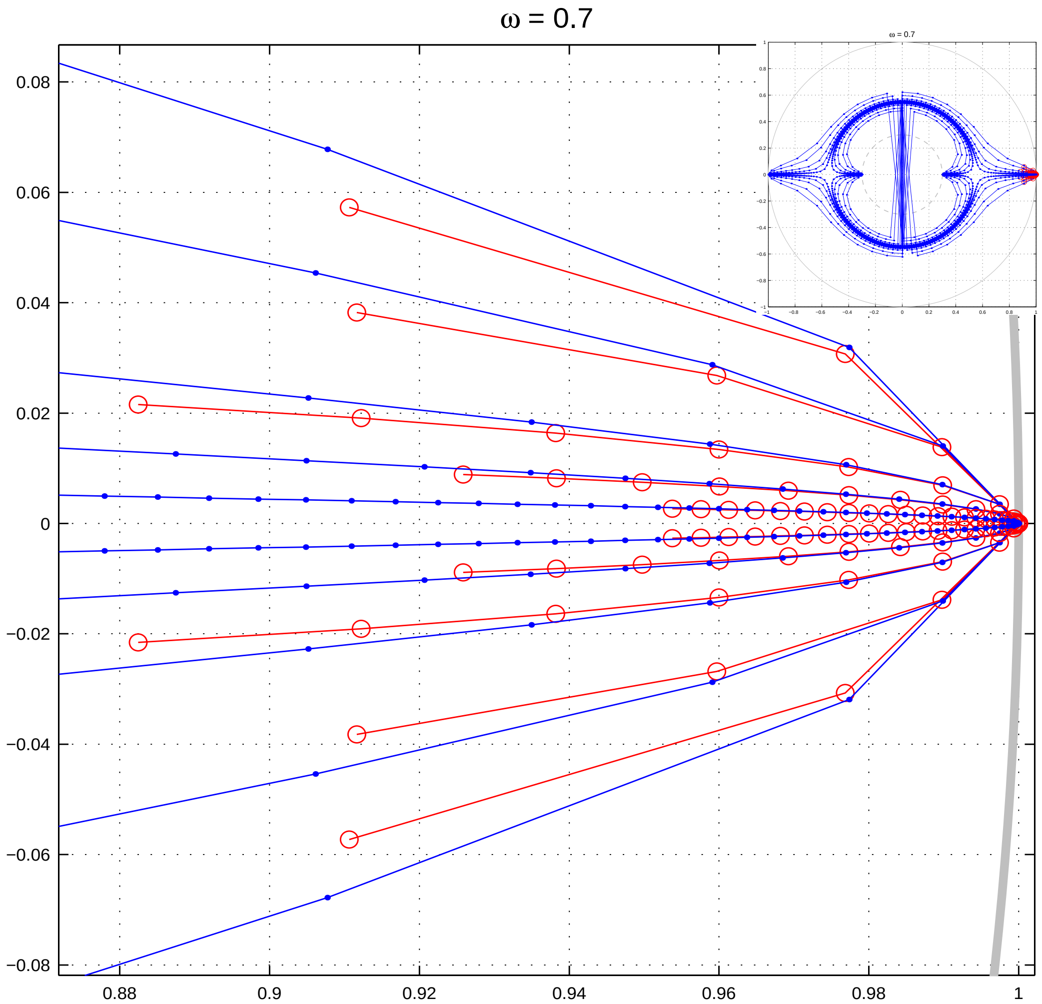


Figure 6.8: Analogously to figure 6.7, this plot refers to the *underrelaxed* advective-diffusive case. The small plot in the right upper corner should help for 'orientation' to see better which detail of the spectra is plotted. Notice how the eigenvalues gather near 1 if N is increased.

Spectral symmetry properties

Figure 6.3, 6.5 and 6.6 reveal that the spectral limit set appears invariant by reflections at the x- and y-axis. Actually, the symmetry with respect to the x-axis is trivial. Since the evolution matrix $\mathbf{E}$ has only real entries, its characteristic polynomial must have real coefficients as well. Therefore, with any zero z the complex conjugated number $\overline{z}$ is a zero too. This explains the invariance of the spectrum under complex conjugation and thus under reflection at the x-axis.

The y-symmetry is due to a less obvious reason. It is useful to observe, that axial symmetry with regard to two perpendicular axes entails the point-symmetry with respect to the intersection point. Conversely, the symmetry relative to a line and a point lying on it, requires also the symmetry with repsect to an orthogonal axis through this point. Therefore the pretended y-symmetry might be a mere consequence of the x-symmetry in 'cooperation' with the point-symmetry relative to the origin. Before we discuss this issue in more detail, the following general consideration in an abstract setup will be quite useful for our concrete purpose.

Abstract setting. For this we consider an endomorphism $A : V \to V$ of a finite dimensional $\mathbb{C}$-vector space V. We would like to establish a connection between symmetry properties of its spectrum and its mapping properties. The spectrum exhibits a rotational symmetry, if there exists a complex number σ with $\sigma{\cdot}\mathrm{spec}(A) \subset \mathrm{spec}(A)$. For instance, if $\sigma = -1$, then the spectrum is symmetric with respect to the reflection at the origin, whereas it is invariant under rotations of 90° for $\pm\sigma = i$. How to define symmetric mapping properties, is less obvious. It turns out that a sort of *generalized commutator relation* like (6.18) comes into play, if we construct another endomorphism, mapping eigenvectors on each other belonging to eigenvalues associated with one another by the symmetry.

Proposition 6.6. *Let $A : V \to V$ be a* diagonalizable *endomorphism of a finite dimensional $\mathbb{C}$-vector space V. Assume that there is $\sigma \in \mathbb{C}$, such that the spectrum of A is fulfills the condition*

$$\sigma \cdot \mathrm{spec}(A) \subset \mathrm{spec}(A).$$

Then there exists a nontrivial endomorphism $B : V \to V$ satisfying the equation

$$AB - \sigma BA = 0. \tag{6.18}$$

Proof: By hypothesis, A is diagonalizable, such that V can be represented as the direct sum of the eigenspaces associated with the (different) eigenvalues, i.e. $V = \oplus_j \mathrm{eig}(A, \lambda_j)$. Therefore it is possible to define B on each of the eigenspaces separately. This can be done in such a way that $B\big(\mathrm{eig}(A, \lambda_j)\big) \subset \mathrm{eig}(A, \sigma\lambda_j)$. For this one has just to choose a basis in $\mathrm{eig}(A, \lambda_j)$ and $\mathrm{eig}(A, \sigma\lambda_j)$. Then $B\big|_{\mathrm{eig}(A,\lambda_j)}$ can be defined by mapping the chosen basis vectors of $\mathrm{eig}(A, \lambda_j)$ on at least one of the basis vectors of $\mathrm{eig}(A, \sigma\lambda_j)$.
Consider now the equation $Av = \lambda v$ for some $\lambda \in \mathrm{spec}(A)$ and $v \in \mathrm{eig}(A, \lambda)$. Multiplying it from the left with σB yields $\sigma BAv = \sigma\lambda Bv$. On the other hand we must have $ABv = \sigma\lambda Bv$ since $Bv \in \mathrm{eig}(A, \sigma\lambda)$ by construction of B and choice of v. Comparing provides $ABv = \sigma BAv$. As this equation can be derived for any $v \in V$ belonging to one of the eigenspaces, it is valid for all elements of at least one basis of V. Hence we even have $AB = \sigma BA$. ∎

The next proposition contains the rather converse statement.

Proposition 6.7. *Let $A : V \to V$ be a* diagonalizable*, nontrivial endomorphism of a finite dimensional $\mathbb{C}$-vector space V. The endomorphism $B : V \to V$ and the number $\sigma \in \mathbb{C} \setminus \{0\}$ shall satisfy*

$$AB - \sigma BA = 0.$$

Suppose there is a basis $\{v_i\}$ of eigenvectors of A, such that $B^k v_i \neq 0$ for $k \in \mathbb{N}$ and $i \in \{1, ..., \dim V\}$. Then

$$\text{i) } \exists n \in \mathbb{N}: \ \sigma^n = 1, \qquad \text{and} \qquad \text{ii) } \sigma \cdot \operatorname{spec}(A) = \operatorname{spec}(A).$$

A sufficient condition for $B^k v_i \neq 0$ with $i \in \{1, ..., \dim V\}$ and $k \in \mathbb{N}$ is, that B has full rank.

Proof: Let us first verify the inclusion relation

$$\sigma \cdot \operatorname{spec}(A) \subset \operatorname{spec}(A). \tag{6.19}$$

For this we have to show that $\sigma\lambda \in \operatorname{spec}(A)$ whenever $\lambda \in \operatorname{spec}(A)$. As λ is an eigenvalue of A, there is $i \in \{1, ..., \dim V\}$ with $Av_i = \lambda v_i$. Multiplying this equation with σB from the left yields $\sigma BAv_i = \sigma\lambda Bv_i$. By using the hypothesis in the form $AB = \sigma BA$ we obtain $ABv_i = \sigma\lambda Bv_i$. Since $Bv_i \neq 0$ is assumed, we realize by definition that $\sigma\lambda$ is an eigenvalue of A. Hence $\sigma\lambda \in \operatorname{spec}(A)$.

ad i) *(reductio ad absurdum)* As A is by hypothesis not the zero-mapping, there is an eigenvalue $\lambda \neq 0$. Similarly to the proof of (6.19), it can be shown by induction that $AB^k v_i = \sigma^k \lambda B^k v_i$ for some $i \in \{1, \ldots, \dim V\}$ and *any* $k \in \mathbb{N}_0$. Thanks to the hypothesis $B^k v_i \neq 0$ we are allowed to conclude that $\sigma^k \lambda$ is an eigenvalue of A.

Now, let us assume that there is no $n \in \mathbb{N}$ with $\sigma^n = 1$. Then $\sigma^m \lambda \neq \sigma^\ell \lambda$ for any pair $m > \ell$ with $\ell, m \in \mathbb{N}_0$. Otherwise, if we had $\sigma^m \lambda = \sigma^\ell \lambda$, then by division of $\sigma^\ell \lambda \neq 0$ we would obtain the equation $\sigma^{m-\ell} = 1$. So we would have found $n = m - \ell \in \mathbb{N}$ with $\sigma^n = 1$, which is in contradiction to the assumption above. Therefore the assumption implies, that there are infinitely many eigenvalues $\lambda, \sigma\lambda, \sigma^2\lambda, ...$ of A. But this conflicts with the finiteness of $\dim V$ being an upper bound for the number of different eigenvalues.

ad ii) From (6.19) we can conclude that $\sigma^{k+1} \cdot \operatorname{spec}(A) \subset \sigma^k \cdot \operatorname{spec}(A)$ for arbitrary $k \in \mathbb{N}_0$, since

$$\begin{aligned}
\sigma^{k+1} \cdot \operatorname{spec}(A) &\equiv \left\{\sigma^{k+1}\mu \,\middle|\, \mu \in \operatorname{spec}(A)\right\} \\
&= \left\{\sigma^{k}(\sigma\mu) \,\middle|\, \mu \in \operatorname{spec}(A)\right\} \\
&= \left\{\sigma^{k}\varkappa \,\middle|\, \varkappa \in \sigma \cdot \operatorname{spec}(A)\right\} \\
&\overset{(6.19)}{\subset} \left\{\sigma^{k}\varkappa \,\middle|\, \varkappa \in \operatorname{spec}(A)\right\} \equiv \sigma^k \cdot \operatorname{spec}(A) \; .
\end{aligned}$$

Having by i) an exponent $n \in \mathbb{N}$ with $\sigma^n = 1$, we end up with the chain of inclusions

$$\begin{aligned}
\operatorname{spec}(A) &= 1 \cdot \operatorname{spec}(A) = \sigma^n \cdot \operatorname{spec}(A) \\
&\subset \sigma^{n-1} \cdot \operatorname{spec}(A) \subset \ldots \subset \sigma \cdot \operatorname{spec}(A),
\end{aligned}$$

which gives the converse inclusion to (6.19). Thus the equality of the sets follows together with (6.19). ■

As a complement of the preceding proposition we get:

Corollary 6.4. *If m denotes the smallest integer with $\sigma^m = 1$, then $m \leq \dim V$.*

Proof: If $\lambda \neq 0$ denotes an eigenvalue of the endomorphism A, then

$$\lambda,\ \sigma\lambda,\ \ldots,\ \sigma^{m-1}\lambda$$

are m different eigenvalues. The assertion follows from the fact, that there are maximally $\dim V$ different eigenvalues. ∎

Let us close this excursion by a final proposition dealing with a special case, that admits a short proof of its own.

Proposition 6.8. *Under the hypothesis of proposition 6.7 assume additionally that A and B are isomorphisms (invertible). Then $\sigma^{\dim V} = 1$.*

Proof: Since A and B are invertible, their determinants are different from 0. From the hypothesis in proposition 6.7 we get $AB = \sigma BA$. Applying the determinant on both sides and employing the multilinearity as well as the product rule, yields

$$\det(A)\det(B) = \sigma^{\dim V}\det(B)\det(A).$$

Dividing by $\det(A)\det(B) \neq 0$ leads to the assertion. ∎

Application. The shift matrices $\mathbf{L}$ and $\mathbf{R}$ (of dimension N) possess a highly symmetric spectrum, since the unit roots correspond to the N vertices of a regular polygon being rotationally symmetric with respect to the origin. If N is even, i.e. $N = 2K$ for some $K \in \mathbb{N}$, the polygon is in particular invariant under reflection at the origin. This special symmetry is inherited by the evolution operator, as we will see below.

According to lemma 6.2 the eigenvalues of $\mathbf{L}$ are $\mathrm{e}^{\mathrm{i}\pi m/K}$, $m \in \{0, ..., N-1\}$ with the eigenvectors $(\mathrm{e}^{\mathrm{i}\pi\ell m/K})_\ell$. By reflection under the origin the n^{th} eigenvalue $\mathrm{e}^{\mathrm{i}\pi n/K}$ is associated to the $(n+K)^{\text{th}}$ eigenvalue $\mathrm{e}^{\mathrm{i}\pi(n+K)/K} = \mathrm{e}^{\mathrm{i}\pi}\mathrm{e}^{\mathrm{i}\pi n/K} = -\mathrm{e}^{\mathrm{i}\pi n/K}$ for $n \in \{0, ..., K-1\}$. Considering the eigenvector of the $(n+K)^{\text{th}}$ eigenvalue, $(\mathrm{e}^{\mathrm{i}\pi\ell(n+K)/K})_\ell = \big((-1)^\ell \mathrm{e}^{\mathrm{i}\pi n/K}\big)_\ell$ yields the observation:

> The eigenvectors pertaining to the n^{th} and $(n+K)^{\text{th}}$ eigenvalue can be chosen, such that they agree in their even components and differ in their odd components only by the factor -1.

Thus, the matrix representation of the endomorphism $\mathbf{B}$ mentioned in proposition 6.6 and 6.7 is a diagonal matrix with the alternating sequence of 1 and -1 on its diagonal. By abuse of notation, let us denote this matrix also by $\mathbf{B}$. So by proposition (6.6) we expect $\mathbf{L}$ and $\mathbf{B}$ to anticommute, since $\sigma = -1$ due to the invariance of spec($\mathbf{L}$) under reflection at the origin (for N even!). It can be easily checked that

$$\mathbf{LB} + \mathbf{BL} = 0 \qquad \text{and likewise} \qquad \mathbf{RB} + \mathbf{BR} = 0. \tag{6.20}$$

From this we obtain for the evolution matrix

$$\begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix}\begin{pmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{B} \end{pmatrix} + \begin{pmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{B} \end{pmatrix}\begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix} = \begin{pmatrix} \alpha(\mathbf{LB}+\mathbf{BL}) & \beta(\mathbf{LB}+\mathbf{BL}) \\ \gamma(\mathbf{RB}+\mathbf{BR}) & \delta(\mathbf{RB}+\mathbf{BR}) \end{pmatrix}$$

which is equal to the zero matrix because of (6.20). Applying now proposition 6.7 proves the next proposition:

Proposition 6.9. *If the number N of grid nodes is even, then the spectrum of the evolution matrix $\mathbf{E}$ pertaining to the lattice-Boltzmann scheme (6.4) is invariant under reflections at the origin.*

So the observed symmetries of the spectrum can be verified without computing the eigenvalues explicitly!
Since the eigenvalues of the evolution matrices associated to an even number of grid nodes lie dense in the spectral limit set, i.e.

$$\mathfrak{S}(\omega, r, p) = \overline{\bigcup_{N\in\mathbb{N}} \operatorname{spec}(\mathbf{E}, \omega, r, p, 2N)}$$

we can eventually infer:

Corollary 6.5. *The spectral limit set $\mathfrak{S}(\omega, r, p)$ is invariant under reflections at the origin.*

Together with the x-symmetry this entails the y-symmetry of the spectral limit set mentioned at the beginning of this paragraph.

Eigenvectors of the evolution operator

The spectral analysis is completed by the computation of the eigenvectors pertaining to the evolution matrix $\mathbf{E}$ of the lattice-Boltzmann scheme 6.4. It seems nearly self-evident that the eigenvectors of $\mathbf{E}$ are somehow related to the eigenvectors of the shift matrices $\mathbf{L}$ and $\mathbf{R}$.
Let us denote by $\mathbf{v}_m$ the unique eigenvector of $\mathbf{L}$ pertaining to the eigenvalue w^m with $m \in \{0, ..., N-1\}$. Then the following holds true according to lemma 6.2:

$$\mathbf{v}_m = (w^{\ell m})_{\ell\in\mathbb{Z}_N}\,, \quad \mathbf{L}\mathbf{v}_m = w^m\mathbf{v}_m \quad \text{and} \quad \mathbf{R}\mathbf{v}_m = w^{-m}\mathbf{v}_m \equiv \overline{w}^m\mathbf{v}_m\,.$$

For the eigenvectors of $\mathbf{E}$ associated to the eigenvalues $\lambda_\pm(2\pi m/N)$ we make the ansatz $\begin{pmatrix} \mathbf{v}_m \\ \mathrm{a}_\pm^{[m]}\mathbf{v}_m \end{pmatrix}$. The unknown amplitude $\mathrm{a}_\pm^{[m]} \in \mathbb{C}$ must be determined in such a way that the equation

$$\begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix}\begin{pmatrix} \mathbf{v}_m \\ \mathrm{a}_\pm^{[m]}\mathbf{v}_m \end{pmatrix} \stackrel{!}{=} \lambda_\pm(2\pi m/N)\begin{pmatrix} \mathbf{v}_m \\ \mathrm{a}_\pm^{[m]}\mathbf{v}_m \end{pmatrix} \tag{6.21}$$

is satisfied. Exploiting the properties of $\mathbf{v}_m$ we obtain for the left hand side

$$\begin{aligned} \begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix}\begin{pmatrix} \mathbf{v}_m \\ \mathrm{a}_\pm^{[m]}\mathbf{v}_m \end{pmatrix} &= \begin{pmatrix} \alpha\mathbf{L}\mathbf{v}_m + \mathrm{a}_\pm^{[m]}\beta\mathbf{L}\mathbf{v}_m \\ \gamma\mathbf{R}\mathbf{v}_m + \mathrm{a}_\pm^{[m]}\delta\mathbf{R}\mathbf{v}_m \end{pmatrix} \\ &= \begin{pmatrix} \alpha w^m\mathbf{v}_m + \mathrm{a}_\pm^{[m]}\beta w^m\mathbf{v}_m \\ \gamma\overline{w}^m\mathbf{v}_m + \mathrm{a}_\pm^{[m]}\delta\overline{w}^m\mathbf{v}_m \end{pmatrix}\,. \end{aligned}$$

Comparison with (6.21) results to:

$$\alpha w^m + \mathrm{a}_\pm^{[m]}\beta w^m = \lambda_\pm(2\pi m/N) \qquad \Rightarrow \quad \mathrm{a}_\pm^{[m]} = \frac{\lambda_\pm(2\pi m/N) - \alpha w^m}{\beta w^m},$$

$$\gamma \overline{w}^m + \mathrm{a}_\pm^{[m]}\delta \overline{w}^m = \lambda_\pm(2\pi m/N)\mathrm{a}_\pm^{[m]} \quad \Rightarrow \quad \mathrm{a}_\pm^{[m]} = \frac{\gamma \overline{w}^m}{\lambda_\pm(2\pi m/N) - \delta \overline{w}^m}.$$

At first sight, it seems as if two different values for $\mathrm{a}_\pm^{[m]}$ come out. However, equating both expressions yields under the assumption that the denominators do not vanish:

$$\lambda_\pm^2(2\pi m/N) - (\alpha w^m + \delta \overline{w}^m)\lambda_\pm(2\pi m/N) + \alpha\delta - \beta\gamma = 0.$$

We know however that $\lambda_\pm(2\pi m/N)$ satisfies this equation by its definition. Therefore both expressions for $\mathrm{a}_\pm^{[m]}$ are equal. Moreover this shows that our ansatz works. Let us summarize the result by the following theorem:

Theorem 6.3. *For $0 \le m \le N-1$, let us denote by $\begin{pmatrix} \mathbf{g}_+^{[m]} \\ \mathbf{d}_+^{[m]} \end{pmatrix}$ and $\begin{pmatrix} \mathbf{g}_-^{[m]} \\ \mathbf{d}_-^{[m]} \end{pmatrix}$ the population vectors corresponding to the eigenvectors of the evolution operator with regard to $\lambda_+(2\pi m/N)$ and $\lambda_-(2\pi m/N)$, i.e.:*

$$\begin{pmatrix} \alpha\mathbf{L} & \beta\mathbf{L} \\ \gamma\mathbf{R} & \delta\mathbf{R} \end{pmatrix} \begin{pmatrix} \mathbf{g}_\pm^{[m]} \\ \mathbf{d}_\pm^{[m]} \end{pmatrix} = \lambda_\pm(2\pi m/N) \begin{pmatrix} \mathbf{g}_\pm^{[m]} \\ \mathbf{d}_\pm^{[m]} \end{pmatrix}.$$

Their ℓ^{th} component associated with the ℓ^{th} grid node ($\ell \in \mathbb{Z}_N$) is given by

$$\mathbf{g}_+^{[m]}(\ell) = \exp(2\pi\mathrm{i}\, m\ell/N), \qquad \mathbf{g}_-^{[m]}(\ell) = \exp(2\pi\mathrm{i}\, m\ell/N),$$
$$\mathbf{d}_+^{[m]}(\ell) = \mathrm{a}_+^{[m]}\exp(2\pi\mathrm{i}\, m\ell/N), \qquad \mathbf{d}_-^{[m]}(\ell) = \mathrm{a}_-^{[m]}\exp(2\pi\mathrm{i}\, m\ell/N),$$

where the amplitude of the right-moving populations stored in $\mathbf{d}_+$ and $\mathbf{d}_-$ is

$$\mathrm{a}_+^{[m]} := \frac{\lambda_+(2\pi m/N)\mathrm{e}^{-2\pi\mathrm{i}m/N} - \alpha}{\beta}, \qquad \mathrm{a}_-^{[m]} := \frac{\lambda_-(2\pi m/N)\mathrm{e}^{-2\pi\mathrm{i}m/N} - \alpha}{\beta}.$$

6.3 LB advection scheme with periodic boundary conditions

The D1P2 lattice-Boltzmann algorithm (6.4) for the advection-diffusion equation presented at the beginning of the previous section can be used to approximate also the solution of the pure advection equation

$$\partial_t v + a\partial_x v = 0 \tag{6.22}$$

with periodic boundary conditions if it is performed with two minor changes:

- The equilibrium (6.6) is replaced by

$$\mathcal{E}_s(U) := \tfrac{1}{2}(1 + as)U. \tag{6.23}$$

  Here the reaction term $-\frac{1}{\omega}ch^2$ is discarded; furthermore and most important is the advective term as occurring now without the factor h.

- Each iteration is identified with a time increment $\Delta t = \Delta x = h$.

The resulting algorithm is identical to the D1P2 scheme which is studied in chapter 5 and where the connection between the lattice-Boltzmann algorithm and the advection equation (6.22) is established by a consistency analysis.

Subsection 6.3.1 and 6.3.2 are based on theorem 6.2. Due to the structural equality of both algorithms (purely advective $\leftrightarrow$ advective-diffusive with $c = 0$) the calculation of eigenvalues can be transfered directly. In the sequel, $\mathbf{E}$ denotes the evolution matrix of the advective lattice-Boltzmann scheme to be considered here. Without repeating the computations, we set in analogy to subsection 6.2.2

$$\begin{aligned}\chi(\lambda,\phi) &:= \lambda^2 - (\alpha e^{i\phi} + \delta e^{-i\phi})\lambda_m + \alpha\delta - \beta\gamma \\ &= \lambda^2 + \big[(\omega-2)\cos(\phi) + i\omega a \sin(\phi)\big]\lambda + 1 - \omega,\end{aligned} \tag{6.24}$$

where the coefficients α, β, γ and δ of the collision block are now given by

$$\begin{aligned}\alpha &:= 1 - \tfrac{1}{2}\omega - \tfrac{1}{2}\omega a, \qquad & \beta &:= \tfrac{1}{2}\omega - \tfrac{1}{2}\omega a, \\ \gamma &:= \tfrac{1}{2}\omega + \tfrac{1}{2}\omega a, \qquad & \delta &:= 1 - \tfrac{1}{2}\omega + \tfrac{1}{2}\omega a.\end{aligned} \tag{6.25}$$

Furthermore (6.15) becomes:

$$\begin{aligned}\lambda_{\pm}(\phi) = -\tfrac{1}{2}\big[(\omega-2)\cos(\phi) + i\omega a\sin(\phi)\big] \\ \pm\sqrt{\tfrac{1}{4}\big[(\omega-2)\cos(\phi) + i\omega a\sin(\phi)\big]^2 - (1-\omega)}.\end{aligned} \tag{6.26}$$

The zero-level set of $\chi(\lambda,\phi)$ yields again the *spectral limit set*

$$\mathfrak{S}(a,\omega) := \big\{\lambda \in \mathbb{C} \,\big|\, \exists\phi \in [0,2\pi) \text{ with } \chi(\lambda,\phi) = 0\big\}, \tag{6.27}$$

which looks exactly like those depicted in figure 6.5 and 6.6 (replace ah by a!).

Thanks to the tight interlocking of both algorithms, we do not only benefit here from subsection 6.2.2 but, conversely, the results of subsection 6.3.1 and 6.3.2 (theorem 6.4 and 6.6) extend *mutatis mutandis* to the advective-diffusive algorithm of the previous section. In contrast, the proof of the ℓ_∞-stability theorem is not generally extendible to the advection scheme, as it essentially exploits the positivity of the collision block coefficients. This can be reached for sufficiently small h if the advective-diffusive equilibrium (6.6) is chosen, which differs from (6.23) by the additional scaling factor h in the advective part.

What have multiscale expansions to do with stability?

Let us briefly motivate the topic of subsection 6.3.3: The general convergence theory (Lax equivalence theorem) requires consistency and *stability*. Whereas the consistency analysis involves mainly elementary techniques like Taylor expansion, that may be applied in a quasi mechanical manner, the inspection of stability often represents a real challenge. However, since checking consistency alone is not sufficient[22]

[22]There are schemes, like the naive or the downstream discretization of the advection equation, lacking stability although being consistent. Unfortunately, in many situations stability may be deteriorated or even forfeited by increasing the order of consistency.

for completely assessing a numeric algorithm, it would be desirable to dispose of an handy, theoretical tool being able to provide indications about the stability behavior (or even to predict it) without implementing and testing the algorithm. This could contribute to systemize and thus to support the construction and improvement of numerical schemes.
Subsection 6.3.3 is an attempt to fathom whether it is possible to examine stability by asymptotic expansions just as in the case of consistency. Concretely, the subsection is inspired by the following background-question: Can multiscale expansions be used to find precise stability criteria for finite difference schemes? For the specific example of the advection equation discretized with the upwind method, a positive answer has been given in [37]. However this is not universally true as demonstrated here by the advective D1P2 lattice-Boltzmann algorithm, for which a twoscale expansion indicates stability, while the method is actually unstable.

6.3.1 The CFL-condition and stability

The discretization of (hyperbolic) PDEs necessitates the observation of the CFL-condition being indispensable for convergence. It requires the domain of dependence of the discretization method to contain the domain of dependence of the PDE (see also [45]). For a three-point stencil difference scheme solving the one-dimensional advection equation the CFL-condition translates into

$$\left|a\tfrac{\Delta t}{\Delta x}\right| \leq 1. \tag{6.28}$$

The inequality admits an easy interpretation, namely that the grid speed[23] $\Delta x/\Delta t$ must not be inferior to the advection speed $|a|$. This means that the scheme must be able to propagate disturbances as fast as the equation in order to converge.
It is remarkable that the CFL-condition is equivalent with stability in the case of standard finite-difference schemes like Upwind, Lax-Friederich or Lax-Wendroff. For the advective D1P2 lattice-Boltzmann algorithm, a similar equivalence holds as shown in theorem 6.6. Before we turn to this, we verify that $|a| \leq 1$ is necessary for stability (note that (6.28) turns into $|a| \leq 1$ for $\Delta t = \Delta x = h$).

Theorem 6.4. *The spectral limit set pertaining to the evolution matrix of the advective D1P2 lattice-Boltzmann scheme is entirely contained in the closed unit disk, i.e.* $\mathfrak{S}(\omega, a) \subset \overline{D_1(0)}$*, if and only if* $0 \leq \omega \leq 2$ *and* $-1 \leq a \leq 1$ *or* $\omega = 0$*.*

If $\mathfrak{S}(\omega, a) \not\subset \overline{D_1(0)}$ then for all but finitely many grids the corresponding evolution matrix has an eigenvalue whose distance from $\overline{D_1(0)}$ must be greater than a constant minimal distance (eigenvalues fill $\mathfrak{S}(\omega, a)$ densely). Therefore the algorithm cannot be stable neither according to definition 6.1 nor to definition 6.2.

As the radicand in (6.26) is generally complex-valued, it is not immediately obvious how to estimate $|\lambda_\pm(\phi)|$ in order to get conditions on the parameters a, ω assuring that the spectral limit set and hence the eigenvalues of $\mathbf{E}$ stay inside the closed unit disk.

[23] Determines how fast a disturbance is propagated by the scheme.

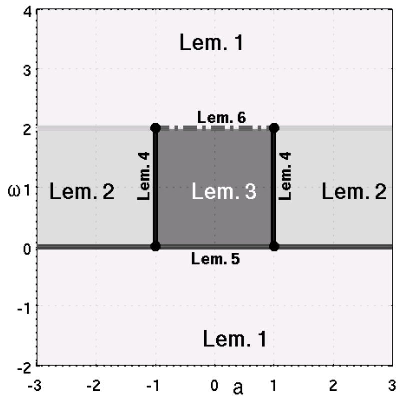


Figure 6.9: The stability domain of $\mathbf{E}$ in the $a\omega$ plane. The gray shading and line types visualizes the case differentiation to prove theorem 6.4. As the collision matrix corresponds to the identity if ω is set to 0, the whole line $\omega = 0$ belongs – apart from the closed rectangle $[-1, 1] \times [0, 2]$ – to the stability domain too. Note that "Lem. 1" refers to lemma 6.7, "Lem. 2" to lemma 6.8 and so on.

Thus, the proof of theorem 6.4 is furnished by chopping up the assertion among several lemmas. For this the $a\omega$-plane is decomposed as depicted in figure 6.9. The verification of the main lemma 6.9, which demarcates essentially the stability domain where $\rho(\mathbf{E}) \leq 1$, is not done by a straight forward computation but employs the theorem of *Rouché* from complex analysis.

Theorem 6.5. ***(Rouché's Theorem)*** *Let Γ be a closed path null-homologous in $G \subset \mathbb{C}$ and assume that Γ has an interior. Let f, g be holomorphic (analytic) on G, and*

$$|f(\zeta) - g(\zeta)| < |f(\zeta)| \qquad \text{for all } \zeta \text{ on } \Gamma.$$

Then f and g have the same number of zeros in the interior of Γ.

A proof is presented in [44], page 181.

In opposition, lemma 6.7 and 6.8 characterize the area of instability, while each of the lemmas 6.10, 6.11 and 6.12 deals with parts of the boundary of the stability domain.

Lemma 6.7. *Suppose $\omega < 0$ or $\omega > 2$. Then $\mathfrak{S}(\omega, a) \not\subset \overline{D_1(0)}$.*

Proof: The constant term of the quadratic polynomial $\chi(\phi, \lambda)$ is $1 - \omega$. According to *Vieta*'s rule the zeros of $\lambda \mapsto \chi(\phi, \lambda)$ satisfy $\lambda_-(\phi) \cdot \lambda_+(\phi) = 1 - \omega$. Now the hypothesis entails $|1 - \omega| > 1$ which implies that the modulus of at least one zero is greater than 1. Then this zero lies outside the closed unit disk. ■

Lemma 6.8. *Let $0 < \omega \leq 2$ and $|a| > 1$. Then $\mathfrak{S}(\omega, a) \not\subset \overline{D_1(0)}$.*

Proof: Assume $a > 1$. The assertion is proved by showing $|\lambda_+(\frac{\pi}{2})| > 1$. Setting $\phi = \frac{\pi}{2}$ in (6.26) yields:

$$\lambda_+(\tfrac{\pi}{2}) = \tfrac{1}{2}\omega a\mathrm{i} + \sqrt{\omega - 1 - \tfrac{1}{4}\omega^2 a^2}.$$

A small computation confirms that the root is imaginary since the radicand is negative for arbitrary ω and $|a| > 1$:

$$\omega - 1 - \tfrac{1}{4}\omega^2 a^2 = -(1 - \tfrac{1}{2}\omega)^2 - \tfrac{1}{4}\omega^2 \overbrace{(a^2 - 1)}^{>0} \;<\; 0.$$

Hence $\lambda_+(\frac{\pi}{2}) = \mathrm{i}\Big(\frac{1}{2}\omega a + \sqrt{1 + \frac{1}{4}\omega^2 a^2 - \omega}\,\Big)$. Finally, this leads to the estimate

$$\begin{aligned} |\lambda_+(\tfrac{\pi}{2})| &= \tfrac{1}{2}\omega a + \sqrt{1 + \tfrac{1}{4}\omega^2 a^2 - \omega} \\ &> \tfrac{1}{2}\omega + \sqrt{1 + \tfrac{1}{4}\omega^2 - \omega} \;=\; \tfrac{1}{2}\omega + (1 - \tfrac{1}{2}\omega) \;=\; 1, \end{aligned}$$

where we have used the hypothesis $0 < \omega \leq 2$ and $a > 1$. Similarly one can show that $\lambda_-(\frac{\pi}{2})$ is located outside the unit circle if $a < -1$. ■

Lemma 6.9. *If* $0 < \omega < 2$ *and* $|a| < 1$ *then* $\mathfrak{S}(\omega, a) \subset \overline{D_1(0)}$.

Proof: We have to show that, under the hypothesis above, the roots of (6.24) lie inside the closed unit disk independently of $\phi \in [0, 2\pi)$. For this, we distinguish two cases. The first one with $\phi \in \{0, \pi\}$ can be treated directly:

$$\chi(\lambda, 0) = \lambda^2 \;+\; (\omega - 2)\lambda \;+\; (1 - \omega), \quad \chi(\lambda, \pi) = \lambda^2 \;+\; (2 - \omega)\lambda \;+\; (1 - \omega)$$

Both polynomials do not depend on a. It is easily seen that the zeros of $\chi(\cdot, 0)$ are 1 and $1 - \omega$, while the zeros of $\chi(\cdot, \pi)$ are -1 and $\omega - 1$. So, they are located inside the closed unit disk if $0 < \omega < 2$.
The second case concerns the general situation $\phi \in [0, 2\pi) \setminus \{0, \pi\} = (0, \pi) \cup (\pi, 2\pi)$, which is handled by using the theorem of *Rouché* (theorem 6.5). To apply it, we need a function of comparison ψ, that must have as many zeros inside the *open* unit disk as we claim that χ has. So ψ is required to have two zeros. Moreover it must satisfy the following estimate on the unit circle:

$$|\chi(z) - \psi(z)| < |\psi(z)| \qquad \text{for all } z \in \partial D_1(0).$$

We choose for ψ the polynomial χ itself *but* setting $\omega = 1$:

$$\psi(\lambda, \phi) := \lambda^2 + \big[-\cos(\phi) + \mathrm{i}a\sin(\phi)\big]\lambda$$

The zeros of ψ are 0 and $\cos(\phi) - \mathrm{i}a\sin(\phi)$. If $\phi \in [0, 2\pi) \setminus \{0, \pi\}$ then $|\cos(\phi)| < 1$ and $|\sin(\phi)| > 0$. This implies that both zeros of ψ are located inside the *open* unit disk assuming the hypothesis $|a| < 1$. Annotation: Geometrically the second zero describes an ellipse for $\phi \in [0, 2\pi)$ parameterized in (counter)clockwise sense if a is positive(negative). Its major semiaxes is 1 along the real axis while its minor semiaxis in imaginary direction has length $|a|$. Evidently, the ellipse is completely contained in the closed unit disk if and only if $|a| \leq 1$.
Let us now verify the required estimate. We start with the right hand side and replace z by $\mathrm{e}^{\mathrm{i}\theta} \in \partial D_1(0)$ with $\theta \in [0, 2\pi)$. Drawing the factor $(\omega - 1)\mathrm{e}^{\mathrm{i}\theta}$ out of

the modulus, using $|e^{i\theta}| = 1$ and observing that the absolute value does not change under complex conjugation, we get:

$$\begin{aligned}
&\left|\chi(\mathrm{e}^{\mathrm{i}\theta},\phi) - \psi(\mathrm{e}^{\mathrm{i}\theta},\phi)\right| = \\
&= \left|(\omega-2)\cos(\phi)\mathrm{e}^{\mathrm{i}\theta} + \cos(\phi)\mathrm{e}^{\mathrm{i}\theta} + \mathrm{i}a(\omega-1)\sin(\phi)\mathrm{e}^{\mathrm{i}\theta} + (1-\omega)\right| \\
&= |\omega-1|\left|\cos(\phi)\mathrm{e}^{\mathrm{i}\theta} + \mathrm{i}a\sin(\phi)\mathrm{e}^{\mathrm{i}\theta} - 1\right| \\
&= |\omega-1||\mathrm{e}^{\mathrm{i}\theta}|\left|\cos(\phi) + \mathrm{i}a\sin(\phi) - \mathrm{e}^{-\mathrm{i}\theta}\right| \\
&= |\omega-1|\left|\overline{\cos(\phi) + \mathrm{i}a\sin(\phi) - \mathrm{e}^{-\mathrm{i}\theta}}\right| \\
&= |\omega-1|\left|\cos(\phi) - \mathrm{i}a\sin(\phi) - \mathrm{e}^{\mathrm{i}\theta}\right|.
\end{aligned}$$

For the left hand side we find by factoring out $-\mathrm{e}^{\mathrm{i}\theta}$:

$$\begin{aligned}
|\psi(\mathrm{e}^{\mathrm{i}\theta},\phi)| &= \left|\mathrm{e}^{2\mathrm{i}\theta} + \big(-\cos(\phi) + \mathrm{i}a\sin(\phi)\big)\mathrm{e}^{\mathrm{i}\theta}\right| \\
&= |-\mathrm{e}^{\mathrm{i}\theta}|\left|-\mathrm{e}^{\mathrm{i}\theta} - \big(-\cos(\phi) + \mathrm{i}a\sin(\phi)\big)\right| \\
&= \left|\cos(\phi) - \mathrm{i}a\sin(\phi) - \mathrm{e}^{\mathrm{i}\theta}\right|.
\end{aligned}$$

So we can conclude:

$$\begin{aligned}
\left|\chi(\mathrm{e}^{\mathrm{i}\theta},\phi) - \psi(\mathrm{e}^{\mathrm{i}\theta},\phi)\right| &= |\omega-1|\left|\cos(\phi) - \mathrm{i}a\sin(\phi) - \mathrm{e}^{\mathrm{i}\theta}\right| \\
&< \left|\cos(\phi) - \mathrm{i}a\sin(\phi) - \mathrm{e}^{\mathrm{i}\theta}\right| = |\psi(\mathrm{e}^{\mathrm{i}\theta},\phi)|.
\end{aligned}$$

The strict inequality holds, indeed, because the assumption $\omega \in (0,2)$ implies $|1-\omega| < 1$. Furthermore $\left|\cos(\phi) - \mathrm{i}a\sin(\phi) - \mathrm{e}^{\mathrm{i}\theta}\right|$ cannot be zero, since $\mathrm{e}^{\mathrm{i}\theta}$ is on the unit circle whereas $\cos(\phi) - \mathrm{i}a\sin(\phi)$ lies in its interior as justified above. Thus, all assumptions of *Rouché's* theorem are fulfilled which gives the assertion. ■

Lemma 6.10. *Assume* $0 \le \omega \le 2$ *and* $|a| = 1$. *Then* $\mathfrak{S}(\omega, a) \subset \overline{D_1(0)}$.

Proof: Setting in (6.24) $a = 1$, it has to be shown that the zeros of the polynomial

$$\chi(\lambda,\phi) = \lambda^2 + \big[(\omega-2)\cos(\phi) + \mathrm{i}\omega\sin(\phi)\big]\lambda + (1-\omega)$$

lie in the closed unit disk. It is easy to check that $\mathrm{e}^{-\mathrm{i}\phi}$ is a root of χ. By *Vieta*'s rule the product of the zeros must be equal to the coefficient of the constant term. Hence the other zero must be given by $(1-\omega)e^{\mathrm{i}\phi}$. So both of the zeros are located inside the closed unit disk under the hypothesis. The other case $a = -1$ is treated likewise. ■

Lemma 6.11. *Suppose* $\omega = 0$. *Then* $\mathfrak{S}(\omega, a) \subset \partial D_1(0) \subset \overline{D_1(0)}$.

Proof: In this case the collision matrix becomes the identity. Hence the evolution matrix is just the transport matrix. So the spectrum lies on the unit circle what is also obtained by formula (6.26) reducing to

$$\lambda_{\pm}(\phi) = \cos(\phi) \pm \sqrt{\cos^2(\phi) - 1} = \cos(\phi) \pm \mathrm{i}\big|\sin(\phi)\big| = \mathrm{e}^{\pm \mathrm{i}\phi}.$$

■

Lemma 6.12. *Suppose* $\omega = 2$ *and* $|a| \leq 1$. *Then* $\mathfrak{S}(\omega, a) \subset \partial D_1(0) \subset \overline{D_1(0)}$.

Proof: In this case (6.26) simplifies to $\lambda_{\pm}(\phi) = -\mathrm{i}a\sin(\phi) \pm \sqrt{1 - a^2\sin^2(\phi)}$. As long as $|a| \leq 1$ the root is real, and we get $|\lambda_{\pm}(\phi)|^2 = a^2\sin^2(\phi) + 1 - a^2\sin^2(\phi) = 1$. ■

6.3.2 Stability in the ℓ_2-norm

Theorem 6.4 already gives the necessary fundamental condition for stability[24]. If $\mathbf{E}$ is diagonalizable[25] (see appendix, corollary 6.6) and $\rho(\mathbf{E}) \leq 1$ the powers of the evolution matrix remain bounded. In this section we show that the bound is independent of the underlying grid which defines the size of the evolution matrix. It must be noted that this statement depends substantially[26] on the norm. An illustration for this is provided by figure 6.10.

Prior to the statement of the final theorem let us consider two specific cases where stability with respect to other norms than $\|\cdot\|_2$ is easily shown. For $a = 0$ (what does not make much sense regarding the advection equation) and $0 \leq \omega \leq 2$ one can argue as in the proof of propostion 6.1.
Summing the entries of the collision matrix column-wise it can be shown that $\|\mathbf{E}\|_1 \leq 1$ if ω and a are chosen such that the four coefficients α, β, γ and δ are non-negative (e.g. $0 \leq \omega \leq 1$ and $-1 \leq a \leq 1$).

Theorem 6.6. *The advective D1P2 lattice-Boltzmann scheme is* stable *in the sense of definition 6.1 with respect to the* ℓ_2*-norm, if and only if* $0 \leq \omega \leq 2$ *and* $-1 \leq a \leq 1$ *or if* $\omega = 0$.

[24] Sometimes the necessity of the stability condition $\rho(\mathbf{E}) \leq 1$ is confounded with sufficiency. The following circumstance might explain this common mistake: In simulations, instabilities become mostly perceivable by sudden blow-ups of the numerical solution. This usually happens, if the iteration operator has eigenvalues outside the closed unit disk. If the initial condition has even a very small component with respect to the 'maximal' eigenvector, for instance by rounding errors, then this component is strongly amplified and gives rise to the blow-up after a certain number of iterations. Indeed, this effect is excluded if $\rho(\mathbf{E}) \leq 1$ (and $\mathbf{E}$ diagonalizable). As the scheme can only be tested on a few (i.e. finite) number of refined grids, it seems justified from a practical perspective to take the above condition as sufficient. From the theoretical point of view, the scheme has to be considered on an infinite sequence of successively refined grids in order to decide wether the scheme is stable and convergent.

[25] If $\mathbf{E}$ were not diagonalizable, i.e. if the Jordan blocks were not simple, then $\|\mathbf{E}^k\|$ increased polynomially what is still less than exponential growth for $\rho(\mathbf{E}) > 1$.

[26] It is true that in finite dimensional vector spaces ($\mathbb{R}^n$) all norms are equivalent. But the coefficients occurring in mutual estimates depend on the dimension of the space. Therefore stability with respect to one norm need not be equivalent with stability in another norm.

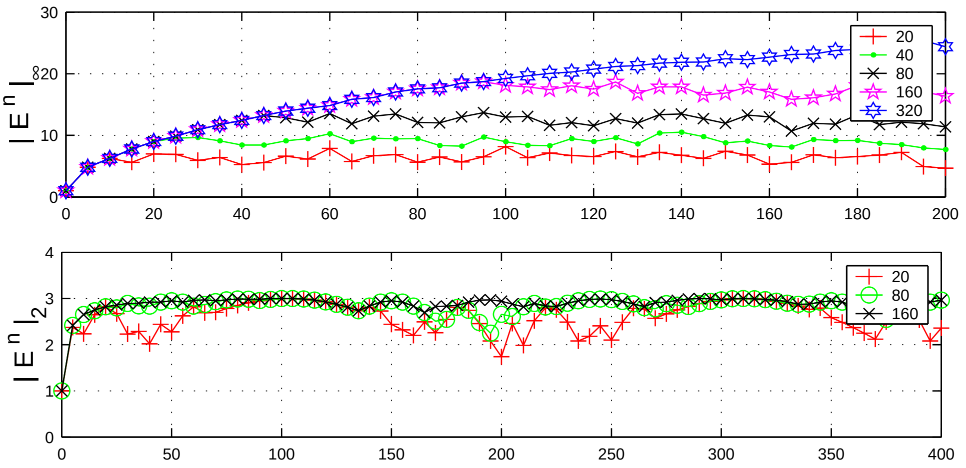


Figure 6.10: The upper diagram visualizes the maximum-norm $\|\mathbf{E}^n\|_\infty$ over the iteration index $n \in \{0, \ldots, 200\}$. For clarity only every fifth value is plotted, which lets the curves appear a little bit smoother. The curves refer to grids of $20, 40, 80, 160$ and 320 nodes with $\omega = 2$ and $a = 0.8$. Analogously, the lower diagram has to be interpreted, except that it refers to $\|\mathbf{E}^n\|_2$. Evidently, as far as the two-norm $\|\cdot\|_2$ is concerned, one realizes a common bound for the powers of the evolution matrices associated to different grids; regarding the $\|\cdot\|_\infty$-norm this bound seems to be lacking.

Before we start with the proof, let us mention that the matrix $\|\cdot\|_2$-norm is induced by the vectorial $\|\cdot\|_2$-norm.

$$\|M\|_2^2 := \sup_{\|x\|=1} \|Mx\|_2^2 = \sup_{\|x\|=1} \langle Mx, Mx\rangle = \sup_{\|x\|=1} \langle MM^*x, x\rangle = \rho(MM^*)$$

From this ensues that the (matricial) $\|\cdot\|_2$-norm keeps invariant under unitary similarity transformations as unitary matrices generate the isometries of the scalar product.

Proof: Defining

$$B(\phi) := \begin{pmatrix} \alpha e^{i\phi} & \beta e^{i\phi} \\ \gamma e^{-i\phi} & \delta e^{-i\phi} \end{pmatrix} \tag{6.29}$$

it can be shown in analogy to the proof of theorem 6.2 that the evolution matrix $\mathbf{E}$ is similar to the block-diagonal matrix

$$\tilde{\mathbf{E}} := \begin{pmatrix} B(0) & & & \\ & B\left(\frac{2\pi}{N}\right) & & \\ & & \ddots & \\ & & & B\left(\frac{2\pi(N-1)}{N}\right) \end{pmatrix} \tag{6.30}$$

by a unitary transformation. Moreover all powers of $\mathbf{E}$ are unitarily similar to those of $\tilde{\mathbf{E}}$ too. This implies $\|\mathbf{E}^n\|_2 = \|\tilde{\mathbf{E}}^n\|_2$. Evaluating the norm by means of lemma 6.13, we obtain

$$\|\mathbf{E}^n\|_2 = \|\tilde{\mathbf{E}}^n\|_2 = \max_{m\in\{0,\ldots,N-1\}} \|B(\tfrac{2\pi m}{N})^n\|_2 \le \sup_{\phi\in[0,2\pi]} \|B(\phi)^n\|_2. \tag{6.31}$$

Thanks to the estimate at right, we get rid of the high dimensional matrix $\tilde{\mathbf{E}}$ and the problem is formulated in terms of the family of 2×2 block-matrices $B(\phi)$. To finish the proof we have to bound the right hand side; for this two cases $a = 0$ and $a \neq 0$ are distinguished.
Case 1: $a = 0$
In this situation we directly calculate $\|B(\phi)\|_2^2 = \rho\big(B(\phi)B(\phi)^*\big)$. The computation

$$\begin{aligned}\det\big(B(\phi)B(\phi)^* - \lambda I\big) &= \det\begin{pmatrix} \alpha^2+\beta^2-\lambda & (\alpha\gamma+\beta\delta)\mathrm{e}^{\mathrm{i}\phi} \\ (\alpha\gamma+\beta\delta)\mathrm{e}^{-\mathrm{i}\phi} & \gamma^2+\delta^2-\lambda \end{pmatrix} \\ &= \lambda^2 - (\alpha^2+\beta^2+\gamma^2+\delta^2)\lambda + (\alpha\delta-\beta\gamma)^2 \\ &= \lambda^2 - (2-2\omega+\omega^2)\lambda + (1-\omega)^2\end{aligned}$$

yields the eigenvalues 1 and $(1-\omega)^2$ pertaining to $B(\phi)B(\phi)^*$. Independently of ϕ we find $\|B(\phi)\|_2 = 1$ for $\omega \in [0, 2]$. Hence it follows with the standard estimate $\sup_{\phi\in[0,2\pi]} \|B(\phi)^n\|_2 \leq \sup_{\phi\in[0,2\pi]} \|B(\phi)\|_2^n \leq 1$.
Case 2: $a \neq 0$
Here the direct argument does not work although $\|B(\phi)\|_2$ is also independent of ϕ. Actually we have $\|B(\phi)\|_2 > 1$ whenever the advection velocity does not vanish. Therefore let us introduce the following sequence of positive functions for $n \in \mathbb{N}$:

$$f_n : [0, 2\pi] \to \mathbb{R}, \qquad f_n(\phi) := \|B^n(\phi)\|_2.$$

As $B(\phi)$ depends continuously on ϕ and due to the continuity of the matrix multiplication and the matricial $\|\cdot\|_2$-norm, the functions f_n are continuous too.
According to lemma 6.15 the matrices $B(\phi)$ with $\phi \in [0, 2\pi]$ are diagonalizable if $a \neq 0$. From theorem 6.4 follows furthermore: $\rho\big(B(\phi)\big) \leq 1$. Hence lemma 6.14 guarantees that arbitrary powers of $B(\phi)$ remain bounded.
This shows that the function sequence $(f_n)_{n\in\mathbb{N}}$ is *pointwise bounded*:

$$\exists C_\phi > 0 : \ \forall n \in \mathbb{N} : \ \ f_n(\phi) < C_\phi.$$

So we can quote the *principle of uniform boundedness* (theorem 6.7) ensuring that the sequence $(f_n)_{n\in\mathbb{N}}$ is even *locally bounded*[27]. Eventually, the compactness[28] of the domain of definition, which is the interval $[0, 2\pi]$, yields global boundedness. So the right hand side of (6.31) must be bounded. ■

The case $a = 0$ has been treated separately, because proving the diagonalizibility of $B(\phi)$ (lemma 6.15) needs some more effort. Observe that case 2 can be made more explicit, by writing $B(\phi)^n = W(\phi)D(\phi)^n W(\phi)^{-1}$. The non-zero elements of the diagonal matrix $D(\phi)$ are of modulus less or equal 1 because $\rho\big(B(\phi)\big) \leq 1$. Then $\|B(\phi)^n\| = \|W(\phi)\|\,\|W(\phi)^{-1}\|$, where the right hand side is independent of n and bounded by reason of continuity and compactness.

[27] $\forall\phi \in [0, 2\pi]$, $\exists\delta > 0$, $\exists C > 0$, $\forall n \in \mathbb{N}$, $\forall\psi \in [0, 2\pi]$ with $|\phi - \psi| < \delta : \ \ |f_n(\psi)| < C$

[28] The theorem of Heine-Borel states that subsets of *finite* dimensional vector spaces are compact, if and only if they are closed and bounded. Generally, compactness of a set S in a topological space is defined by the property that any open cover of S permits a finite cover. In the concrete case a open cover is generated, since any $\phi \in [0, 2\pi]$ has an open neighborhood where the f_n's are bounded by a common constant. Due to the compactness, there are finitely many of these neighborhoods which cover $[0, 2\pi]$. So global boundedness ensues.

Complements to the proof of theorem 6.6

Theorem 6.7. ***(Osgood's theorem – principle of uniform boundedness)*** *Let (M,d) be a complete metric space and $\mathcal{F} \subset \mathcal{C}(M,\mathbb{R})$. Suppose that, for each $x \in M$, there exists a constant $C_x > 0$ such that $\forall f \in \mathcal{F}: \ |f(x)| \leq C_x$. Then there exists a nonempty open ball $B \subset M$ and a constant $C > 0$ such that $\forall f \in \mathcal{F}$, $\forall x \in B$: $|f(x)| \leq C$.*

For the proof we refer to [58] page 184.

Lemma 6.13. *Let $M = \mathrm{diag}(B_1, ..., B_N)$ be a block-diagonal matrix. Then $\|M\|_2 = \max_k \|B_k\|_2$.*

Proof: The assertion is due to the fact that the $\|\cdot\|_2$-norm is connected to the eigenvalues of MM^*, which can be computed block-wise:

$$\begin{aligned}\|M\|_2^2 &= \rho(MM^*) \\ &= \max\big\{|\lambda| \,:\, \det(MM^* - \lambda) = 0\big\} \\ &= \max\big\{|\lambda| \,:\, \prod_k \det(B_k B_k^* - \lambda) = 0\big\} \\ &= \max_k \max\big\{|\lambda| \,:\, \det(B_k B_k^* - \lambda) = 0\big\} \\ &= \max_k \rho(B_k B_k^*) \ = \ \max_k \|B_k\|_2^2\end{aligned}$$

■

Observe that other induced matrix norms (e.g. $\|\cdot\|_1$ or $\|\cdot\|_\infty$) behave analogically.

Lemma 6.14. *Let M be a diagonalizable matrix and assume $\rho(M) \leq 1$. Then there is for any matrix norm $\|\cdot\|$ a constant $C > 0$ such that: $\forall n \in \mathbb{N}: \ \|M^n\| \leq C$.*

A proof of a slightly more general assertion is contained in [12], theorem 3.33 page 108.

Lemma 6.15. *The 2×2 matrix $B(\phi)$ defined in (6.29) is diagonalizable for $a \neq 0$.*

Proof: It is enough to show that the characteristic polynomial, i.e. $\chi(\lambda, \phi)$, has two different zeros, since this implies immediately the existence of two different eigenvectors forming an eigenbasis.
Case 1: $\phi \notin \{0, \pi\}$ and $\omega \neq 0$
Under this condition the coefficient of the linear term is complex-valued while the constant term $1 - \omega$ is real. From *Vieta*'s rule we can conclude that at least one zero is properly complex, say $z_1 = r_1 \mathrm{e}^{\mathrm{i}\theta}$ with $r_1 > 0$ and $\theta \in (-\pi, 0) \cup (0, \pi)$. Furthermore we get $z_2 = r_2 \mathrm{e}^{-\mathrm{i}\theta}$ with $r_2 > 0$ for $\omega \neq 1$ or $r_2 = 0$ for $\omega = 1$ such that the product is real, more exactly $z_1 z_2 = 1 - \omega$. So we may conclude $z_1 \neq z_2$.
Case 2: $\phi \in \{0, \pi\}$ or $\omega = 0$
In these cases the characteristic polynomial and thus the eigenvalues do not depend on a. An explicit computation yields two different eigenvalues 1 and $\omega + 1$ for $\omega \neq 0$ and $\phi = 0$. Similarly, -1 and $\omega - 1$ are eigenvalues if $\phi = \pi$ and $\omega \neq 0$. If only the second condition is satisfied, then the eigenvalues are $\mathrm{e}^{\mathrm{i}\phi}$ and $\mathrm{e}^{-\mathrm{i}\phi}$. Finally, $B(\phi)$ is automatically diagonal (positive/negative identity) if both conditions are fulfilled. ■

For $a = 0$ the discriminant[29] of the characteristic polynomial is strictly positive for $\omega > 1$ and implies the existence of two eigenvalues and eigenvectors. However, if $\omega \leq 1$, then ω and ϕ can be chosen such that the discriminant may vanish. In this case it is not clear *a priori* whether $B(\phi)$ has two linearly independent eigenvectors. The lemma implies immediately that the transformed evolution matrix $\tilde{\mathbf{E}}$ defined in (6.30) is diagonalizable as well. The similarity between $\tilde{\mathbf{E}}$ and $\mathbf{E}$ entails the subsequent corollary.

Corollary 6.6. *The evolution matrix* $\mathbf{E}$ *of the advective D1P2 lattice-Boltzmann algorithm is* diagonalizable *if* $a \neq 0$.

6.3.3 Multiscale expansion and stability

Let us briefly recall the results of subsection 5.2.2 concerning the twoscale expansion with a linear and quadratic time $t_1 = nh$ and $t_2 = nh^2$. The idea is to approximate the population function F of the lattice-Boltzmann algorithm in the following way:

$$\mathsf{F}(nh, x) \approx \mathsf{f}^{(0)}(nh, nh^2, x) + h\mathsf{f}^{(1)}(nh, nh^2, x) + \dots \quad . \tag{6.32}$$

The asymptotic order functions $\mathsf{f}^{(0)}, \mathsf{f}^{(1)}, \dots$ are defined such that the order of the residual (with respect to h) becomes as large as possible. In contrast to a regular expansion with a single time, the order functions do not come out in a cogent manner. Therefore additional conditions may be imposed simplifying the calculation as much as possible without loosing the flexibility of the ansatz. So the subsequent PDEs are extracted, which determine the evolution of the mass moment $u^{(0)}$ associated to $\mathsf{f}^{(0)} = \frac{1}{2}(1 + a\mathsf{s})u^{(0)}$.

$$\partial_{t_1} u^{(0)}(t_1, t_2, x) + a\partial_x u^{(0)}(t_1, t_2, x) = 0 \tag{6.33}$$

$$\partial_{t_2} u^{(0)}(t_1, t_2, x) - \mu\partial_x^2 u^{(0)}(t_1, t_2, x) = 0 \quad \text{with} \quad \mu = (\tfrac{1}{\omega} - \tfrac{1}{2})(1 - a^2) \tag{6.34}$$

The advection equation (6.33) is already found using the regular expansion. It describes the most conspicuous behavior of $u^{(0)}$ because t_1 evolves faster than t_2. In opposition, the diffusion equation accounts for a less striking process, that becomes only noticeable after an observation over rather many iterations.
The sign of diffusivity μ plays a crucial role. In case of $\mu \geq 0$ the solutions of the diffusion equation, being damped and smoothed, exhibit the typical decently behaving dynamics. The situation changes dramatically if μ attains negative values. This case is referred to as the *backward diffusion equation*; the associated initial value problem is known to be ill-posed[30].
Whenever the parameters a and ω are set such that μ becomes negative, the twoscale expansion signalizes via equation (6.34) that there might arise some trouble with the stability of the algorithm. So we are led to the assumption that the domain of

[29] $\frac{1}{4}\omega^2\cos^2(\phi) + (\omega - 1)\sin^2(\phi)$

[30] Formally the associated evolution should be inverse to that of the diffusion equation. Therefore its effect would be roughing instead of smoothing. However, the example of an Gaussian bell function, evolving in finite time into a δ-distribution-like peak, gives clear evidence that the associated IVP generally permits no solution in the standard function spaces. Thus the twoscale expansion breaks down for $\mu < 0$, which implies that it is no more adequate to approximate the population function F.

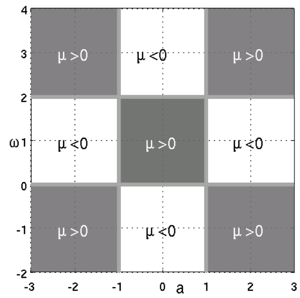


Figure 6.11: The stability domain S_μ as predicted by the twoscale expansion, suggesting that the four closed quadrants – each one sharing its corner with the central square – belong to the stability domain as well. Compare with figure 6.9.

stability is given by the subset $S_\mu := \{(a,\omega) \in \mathbb{R}^2 :\ \mu(a,\omega) = (\frac{1}{\omega} - \frac{1}{2})(1-a^2) \geq 0\}$ of the a-ω plane sketched in figure 6.11. Although this guess is not completely right, because it contains the actual domain of stability as a proper subset, it must be emphasized that it is only based on simple formal asymptotic computations and some elementary knowledge about the diffusion equation.
Still the question remains open, why exactly the analysis of the long-term behavior should contribute to understand stability properties. If the parameters a and ω are chosen outside of the actual stability domain, the algorithm blows up after a certain number of iterations that is almost independent of the grid spacing h. This insinuates rather that the instable behavior strikes very fast, more precisely in the discrete time scale. In the light of this observation the chosen twoscale ansatz with its additional slow time appears especially inappropriate, even contradictory.
Throwing a glance at the spectra in figure 6.12 and 6.13, reveals that the occurrence of instabilities must be attributed to the eigenvalues of maximal modulus located on the imaginary axis. Their associated eigenvectors are slightly excited because of rounding errors if not due to the initialization itself. By application of the evolution matrix they undergo the strongest amplification such that they quickly become disturbing and lead finally to the 'explosion'.
Since the spectral limit set depends only on a and ω, the position of the 'bad' eigenvalues, most responsible for instability, does practically not change under grid refinement. Furthermore, as these eigenvalues are located half-way between the source and sink points, the corresponding eigenvectors are strongly oscillatory with half of the maximal spatial frequency (see figure 6.12 right below). As the frequency grows with the number of grid nodes, the algorithm uncovers its discrete nature here, that eludes a description by smooth[31] functions. However, smooth functions are used to construct the asymptotic order functions $\mathsf{f}^{(0)}(t_1,t_2,x)$ etc.. Restricted to the grid, they are represented essentially as linear combinations of the 'smooth-

[31]One may object why, actually, it is reasonable to analyze discrete algorithms by means of smooth functions, if they can not fully mimic discrete behavior. There are two reasons: First, in many cases one is mostly interested in the properties of discrete algorithms that are not typically discrete. Second, one has generally a better feeling for smooth functions than for irregular, discrete functions.

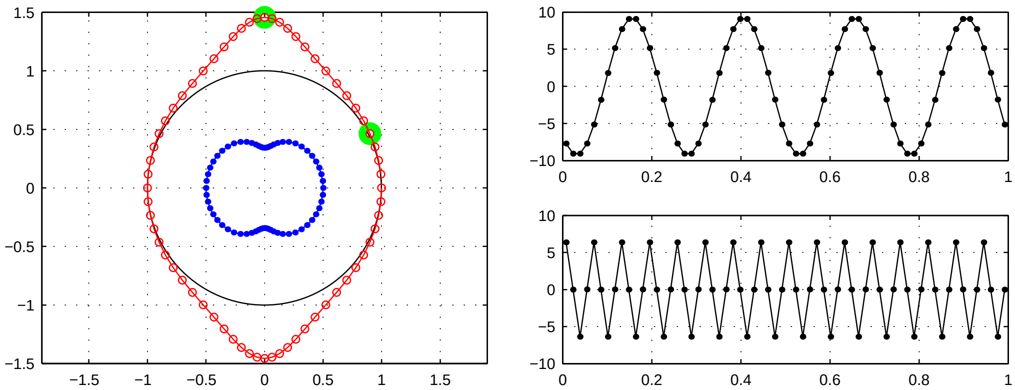


Figure 6.12: Left: Spectrum (grid with 64 nodes) and spectral limit set of the evolution operator for $\omega = 1.5$ and $a = 1.2$. Right: mass moments of the eigenvectors pertaining to the highlighted eigenvalues; the diagram shows the real part while the imaginary part vanishes. The sinusoidal curve (above) refers to the marked eigenvalue close to 1 on the circled branch (λ_+). This curve does not alter under grid refinement if the fourth eigenvalue from 1 (in counterclockwise orientation) is always considered. It just gets sampled more densely. A different situation is found for the eigenvalue closest to the markers position on the imaginary axis. The mass moment of the associated eigenvector looks like a zigzag whose oscillations grow if the grid is refined. The limit of such a sequence of eigenvectors cannot be described by a smooth function.

looking' eigenvectors (longwave modes, cf. figure 6.12 right above) pertaining to the eigenvalues near 1. It seems to be a trivial consequence that it is not possible to grasp the full behavior of the evolution matrix by just analyzing its action on a small subspace.

Nevertheless it is possible to gain certain information by this ansatz, if, figuratively speaking, the eigenvalues near 1 are drawn out of the closed unit disk $\overline{D_1(0)}$ by those which are farther outside already. This situation occurs if μ is negative (see figure 6.12). The eigenvectors pertaining to those eigenvalues are only faintly amplified by the evolution matrix so that their blow-up happens after comparatively many iterations. Moreover, the eigenvalues whose eigenvectors do not exceed a fixed maximal spatial frequency converge to 1 as h tends to 0. Therefore the amplification rate of the associated eigenvectors decreases under grid refinement and this entails that their blow-up takes place after more and more iterations. The quadratic time variable $t_2 = nh^2$ in (6.32) accounts for this phenomenon. What we actually sense by the twoscale expansion is not the early blow-up of the 'rough' eigenvectors but the blow-up of the 'smooth' ones occurring much later.

Figure 6.13 depicts a situation where the asymptotic approach is outwitted, which happens if $|a| > 1$ and $\omega \notin [0, 2]$ such that $\mu > 0$. Now the spectral branch referring to λ_+ and containing 1 lies within $\overline{D_1(0)}$. Instead the other branch has completely left $\overline{D_1(0)}$. Despite of μ being positive the algorithm cannot be stable. Why does the prediction based on the twoscale expansion miss this instability? The reason is simple: even those eigenvectors are very oscillating that belong to the eigenvalues being scarcely outside $\overline{D_1(0)}$ and close to 1 (confer 6.13 right below). Using only

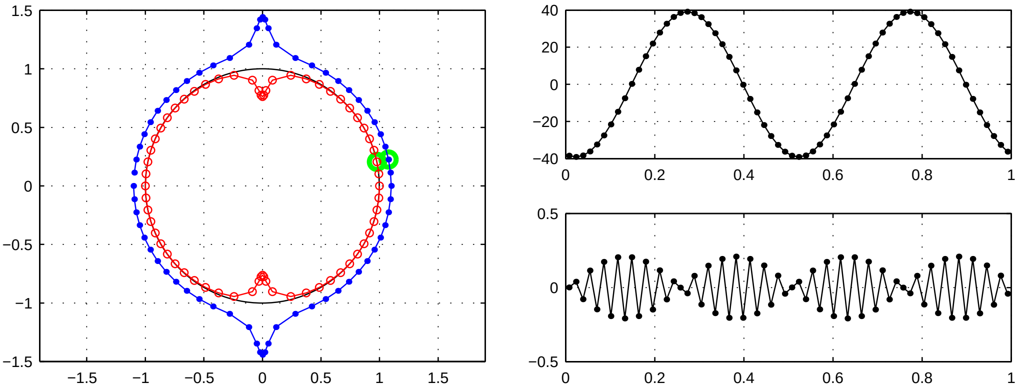


Figure 6.13: Another case with $\omega = 2.1$ and $a = 1.05$ in analogy to figure 6.12. The sinusoidal curve (above) refers to the eigenvalue of the interior, circled branch (λ_+) situated within $\overline{D_1(0)}$. The curve below with the beat-like bellies belongs to the marked eigenvalue on the exterior, dotted branch (λ_-) lying outside $\overline{D_1(0)}$. Again, under grid refinement the eigenvectors (referred to the eigenvalues at second position counted from the positive real axis in counterclockwise orientation) behave distinctly. The upper curve is filled up with dots but its shape is not affected. In contrast, the number of oscillations within each belly is enhanced in the lower curve although the bellies keep their form. The beats are generated by a small difference in the sampling frequency of the grid and the spatial frequency of an exponential function that is contained in the analytic expression of the eigenvector.

smooth ansatz functions, the expansion (6.32) is blind for them. If rounding errors were avoided and if the initialization involved only eigenvectors associated to the λ_+-branch, the algorithm would be stable, however.

Accordance of multiscale expansion and expansion of eigenvalues near 1

In order to elucidate the relation between the multiscale expansion and the spectrum of the evolution matrix we return to (6.26) which gives an analytic expression for the eigenvalues setting $\phi \in \{\frac{2\pi m}{N} \;:\; m \in 0, ..., N-1\}$. Substituting $1/N$ by h we recognize that the eigenvalues $\lambda_+(2\pi mh)$ converge to 1 for h tending to 0 and fixed m. Performing a Taylor expansion around $h = 0$

$$\begin{aligned}\lambda_+(2\pi mh) \;=\; 1 \;-\; 2\pi \mathrm{i} m a\, h \\ -\; 4\pi^2 m^2(\tfrac{1}{\omega} - \tfrac{1}{2})(1-a^2)\, h^2 \;-\; 2\pi^2 m^2 a^2\, h^2 \;+\; \mathrm{O}(h^3)\end{aligned} \tag{6.35}$$

gives an approximation of certain[32] eigenvalues near 1. The coefficients of first and second order turn out to be the negative eigenvalues[33] of the advection operator $a\partial_x$ and the diffusion operator $-\mu\partial_x^2 = -(\frac{1}{\omega} - \frac{1}{2})(1-a^2)\partial_x^2$ (on $[0,1]$ with periodic boundary conditions) appearing in (6.33) and (6.34). The additional term in the second order can be made clear by looking at the eigenvalues of the evolution operators related to the advection and diffusion

[32]Bear in mind, that there might be other eigenvalues close to 1 (e.g. on the branch associated with λ_-) that are not approximated by the above expansion.

[33]It is quickly checked that $x \mapsto \mathrm{e}^{2\pi m \mathrm{i} x}$ with $m \in \mathbb{Z}$ represents a basis of eigenfunctions respecting periodic boundary conditions.

equation and associated to a time interval Δt.

$$\begin{array}{llll} \text{advection:} & e^{-2\pi i m a \Delta t} & = & 1 - 2\pi i m a \Delta t - 2\pi^2 m^2 a^2 \Delta t^2 + O(\Delta t^3) \\ \text{diffusion:} & e^{-4\pi^2 m^2 \mu \Delta t} & = & 1 - 4\pi^2 m^2 \mu \Delta t + O(\Delta t^2) \end{array} \tag{6.36}$$

Apparently, (6.35) is obtained by setting $\Delta t = h$ in $(6.36)_1$ and $\Delta t = h^2$ in $(6.36)_2$, multiplying both truncated series and keeping only terms of order less than $O(h^3)$. So the expansion of the eigenvalues (6.35) corroborates the results of the multiscale expansion:

I) The effect of the lattice-Boltzmann evolution matrix (on a certain subspace) corresponds to an intertwining of an advection and diffusion process where the first one is dominant.

II) There is an inherent separation of time scales: the advection takes place in the linear (fast) time scale while the diffusion evolves in the quadratic (slow) time scale.

6.4 LB diffusion scheme with bounce-back type boundary conditions

Let us now return to the parabolicly scaled D1P2 lattice-Boltzmann scheme that is to be endowed with non-periodic boundary conditions. Concretely we have the following in mind:

- Proper bounce-back condition: this amounts to the *Robin*[34] condition for the target equation. In the absence of advection, Robin and Neumann conditions coincide[35] with each other.

- Sign inverting bounce-back condition: If the sign of the corresponding collision products is flipped before they undergo the bounce-back operation, the homogeneous *Dirichlet* boundary condition for the mass moment is obtained.

In order to see that these kinetic conditions are really connected with the indicated macroscopic boundary conditions for the target problem, a consistency analysis has to be performed. This, however, is not done in the scope of the thesis, instead, we refer to section 3.5, where analogous conditions are considered for the D1P3 lattice-Boltzmann equation.

The objective of this section is to compute the spectra of the modified evolution matrices resulting from the two non-periodic boundary conditions. Following the idea employed in the case of periodic boundary conditions, we first try to diagonalize the transport matrix. The analytic accessibility of the spectrum of the evolution matrix is finally owed to the circumstance that the collision matrix admits a simple[36] representation in the eigenbasis of the transport matrix. So the characteristic

[34] Also called *natural* or *no-flux* condition in the homogeneous case.

[35] The heat flux corresponds to the negative temperature gradient. Hence the no-flux condition at the boundary implies the vanishing of the normal derivative, i.e. the Neumann condition. In one space dimension we see that the flux $au - \nu\partial_x u$ gets proportional to the spatial derivative of u if the advection velocity a vanishes.

[36] The special structure of the collision matrix as nodal operation must be respected which implies the existence of invariant subspaces.

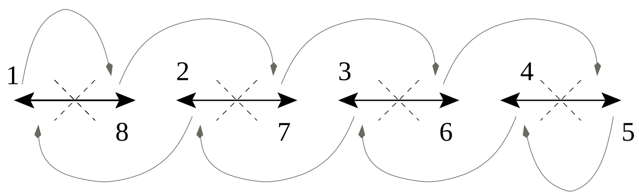


Figure 6.14: In the case of non-periodic boundary conditions the leftmost and rightmost node do not have neighbor nodes on both sides. Therefore each of the two boundary nodes has a population (4 and 8) that can not be updated in the usual way. In contrast, the collision products of index 1 and 5 cannot be sent away during the transport step like their counterparts in other nodes. To make a virtue of necessity, these collision products are used to supply the otherwise empty populations of index 8 and 4 respectively (bounce-back). One may imagine that these collision products are reflected by a wall and thus return to their emitting node with opposite velocity. In the case of bounce-back with sign flipping the numerical value of the collision products is multiplied by -1.

polynomial of the evolution matrix eventually decomposes into quadratic[37] factor polynomials.

The knowledge of the eigenvalues provides some insight how spectral properties of the target problem devolve upon the discretization. In subsection 6.4.2 the eigenvectors of the evolution matrix referring to bounce-back boundary conditions are calculated exemplarily. These computations may be the starting point for future investigations[38] concerning the consistency of boundary conditions and the convergence of the lattice-Boltzmann algorithms.

Some notational changes. In the case of non-periodic boundary conditions we use $\{1, \dots, N\}$ to index the nodes in standard manner from left to right. Furthermore the populations are indexed as depicted in figure 6.14, which differs from figure 6.1 (periodic case) since now one species is counted from left to right while the other is counted in backwards direction. In this way, however, the transport matrix becomes particularly simple, e.g. a shift-matrix.
From now on w stands for the unit root of order $2N$: $w := \mathrm{e}^{i\pi/N}$. In opposition to the previous convention, the shift matrix $\mathbf{L}$ and the Fourier matrix $\mathbf{F}$ are taken as $2N \times 2N$ matrices.
We consider the case of absent advection; the reason for this restriction is revealed below. Hence we take the equilibrium $\mathcal{E}_s(U) = \frac{1}{2}(1 - \frac{1}{\omega}ch^2)U$ instead of (6.6). Then the coefficients of the collision block become

$$\alpha = \delta = 1 - \tfrac{1}{2}\omega - \tfrac{1}{2}ch^2, \qquad \beta = \gamma = \tfrac{1}{2}\omega - \tfrac{1}{2}ch^2. \tag{6.37}$$

[37]The factor polynomials come out as a parameterized family sharing the same structure and distinguishing only in the value of a parameter. Observe that their degree seems to be determined by the size of the collision block.

[38]Similarly to the Fourier approach in section 2.2, this offers an alternative contrasting the approach via asymptotic expansions carried out in chapter 4.

6.4.1 Evolution matrices and their spectra

Bounce-back inducing Robin conditions

Let us start with bounce-back boundary conditions. In this case the lattice-Boltzmann equation is described by the following equations,

$$\left.\begin{array}{lll} \mathsf{F}_s(t_{n+1},x_j) &= \mathsf{C}_s(t_{n+1},x_{j-s}) & \text{(bulk algorithm)} \\ \mathsf{F}_{+1}(t_{n+1},x_1) &= \mathsf{C}_{-1}(t_{n+1},x_1) & \text{(bounce-back at left boundary)} \\ \mathsf{F}_{-1}(t_{n+1},x_N) &= \mathsf{C}_{+1}(t_{n+1},x_N) & \text{(bounce-back at right boundary)} \end{array}\right\} \tag{6.38}$$

where the index s refers to the D1P2 velocity set, i.e. $s \in \{-1,+1\}$, and the collision product is given by $\mathsf{C}_s(t_{n+1},x) := (1-\omega)\mathsf{F}_s(t_n,x) + \omega\mathcal{E}_s\big(U(t_n,x)\big)$.

The first equation summarizes the evolution step of the fundamental algorithm, which refers to all populations that can be supplied with collision products by the regular transport step. This particularly concerns the populations belonging to interior nodes of the grid (bulk nodes[39]) but also the "negative" and "positive" population of the left and right boundary node respectively. The second and third line regulate the update of the positive population $\mathsf{F}_{+1}(t_{n+1},x_1)$ at the leftmost node and the negative population $\mathsf{F}_{-1}(t_{n+1},x_N)$ at the rightmost node, where the standard update cannot be applied due to missing neighbor nodes.

If the populations are indexed as depicted in figure (6.14) the transport matrix $\mathbf{T}_R$ just corresponds to the leftshift matrix $\mathbf{L}$ of dimension $2N$. This circumstance turns out to be the key observation enabling the subsequent computations. For the four node system in figure (6.14) the transport and the collision matrix read

$$\mathbf{T}_R = \mathbf{L} = \left(\begin{array}{cccc|cccc} \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right) \qquad \mathbf{C} \;=\; \left(\begin{array}{cccc|cccc} \alpha & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \beta \\ \cdot & \alpha & \cdot & \cdot & \cdot & \cdot & \beta & \cdot \\ \cdot & \cdot & \alpha & \cdot & \cdot & \beta & \cdot & \cdot \\ \cdot & \cdot & \cdot & \alpha & \beta & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \beta & \alpha & \cdot & \cdot & \cdot \\ \cdot & \cdot & \beta & \cdot & \cdot & \alpha & \cdot & \cdot \\ \cdot & \beta & \cdot & \cdot & \cdot & \cdot & \alpha & \cdot \\ \beta & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \alpha \end{array}\right).$$

We notice that the transport matrix is irreducible and has no block-diagonal structure contrasting the situation of periodic boundary conditions. The collision matrix turns out to be a linear combination of the $2N \times 2N$ identity $\mathbf{I}$ and the "reflected" identity matrix $\mathbf{J}$, where the reflection refers to either the vertical or horizontal central axis forming together a cross that divides that matrix in four $N \times N$ blocks.

Let us therefore introduce the matrix $\mathbf{J}_M$, whose elements are defined by

$$\mathbf{J}_M(k,\ell) := \delta_{k,M+1-\ell} \qquad k,\ell \in \{1,\dots M\}.$$

[39]Observe that there are only bulk nodes in the case of periodic boundary conditions due to the choice of the index set $\mathbb{Z}/N\mathbb{Z}$.

In the sequel, we deal mostly with $\mathbf{J}_{2N}$ and $\mathbf{J}_N$; we omit the subscripts, whenever the dimension is clear by the context.

Hence the transport and collision matrix are generally given by

$$\mathbf{T}_R = \mathbf{L} \quad \text{and} \quad \mathbf{C} = \alpha \mathbf{I} + \beta \mathbf{J},$$

defining the evolution matrix[40] $\mathbf{E}_R = \mathbf{T}_R \mathbf{C}$ by their product.
Now the motivation for our restriction to switch off the advection parameter (i.e. setting $a = 0$) clarifies: Since the transport matrix has lost its 2×2 block structure (which it had in the case of periodicity), we expect that the interplay between collision and transport matrix only keeps analytically handable, if the collision matrix is easily expressed in terms of simply structured $2N \times 2N$ matrices. This is obtained by abandoning[41] the advection, such that transport and collision matrix appear on equal footing.

The main preparation for the computation of the eigenvalues consists in the following lemma.

Lemma 6.16. *The Fourier transform of the matrix $\mathbf{J}_M$ is given by*

$$\mathbf{F}_M \cdot \mathbf{J}_M \cdot \mathbf{F}_M^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \operatorname{diag}(w_M^1, ..., w_M^{M-1}) \mathbf{J}_{M-1} \end{pmatrix} = \left(\begin{array}{c|cccc} 1 & 0 & & \dots & 0 \\ \hline 0 & & & & w_M^1 \\ \vdots & & & \iddots & \\ 0 & w_M^{M-1} & & & \end{array} \right).$$

Proof: Using $w_M^M = 1$ and $\overline{w}_M = w_M^{-1}$, we compute the matrix element associated

---

[40] The subscript R stands for $\underline{\text{R}}$obin boundary conditions.
[41]

$$a = 0 \quad \Rightarrow \quad \alpha = \delta \wedge \beta = \gamma \quad \Rightarrow \quad \mathbf{C} = \begin{pmatrix} \alpha \mathbf{I} & \beta \mathbf{J} \\ \gamma \mathbf{J} & \delta \mathbf{I} \end{pmatrix} = \begin{pmatrix} \alpha \mathbf{I} & \beta \mathbf{J} \\ \beta \mathbf{J} & \alpha \mathbf{I} \end{pmatrix} = \alpha \mathbf{I} + \beta \mathbf{J}$$

to the indices $k, \ell \in \{1, ..., M\}$:

$$
\begin{aligned}
\left(\mathbf{F}_M \cdot \mathbf{J}_M \cdot F_M^{-1}\right)(k,\ell) &= \sum_{n=1}^{M}\sum_{m=1}^{M} \mathbf{F}_M(k,m)\,\mathbf{J}_M(m,n)\,\mathbf{F}_M^{-1}(n,\ell) \\
&= \sum_{n=1}^{M}\sum_{m=1}^{M} \frac{1}{\sqrt{M}}\overline{w}_M^{(k-1)(m-1)}\,\delta_{m,M+1-n}\,\frac{1}{\sqrt{M}}w_M^{(n-1)(\ell-1)} \\
&= \frac{1}{M}\sum_{n=1}^{M}\overline{w}_M^{(k-1)(M+1-n-1)}\,w_M^{(n-1)(\ell-1)} \\
&= \overline{w}_M^{(k-1)(M-1)}\,\frac{1}{M}\sum_{n=1}^{M} w_M^{(k-1)(n-1)}\,w_M^{(n-1)(\ell-1)} \\
&= w_M^{(k-1)(1-M)}\,\frac{1}{M}\sum_{n=1}^{M}\left(w_M^{(k-1)+(\ell-1)}\right)^{n-1} \\
&= w_M^{k-1}\frac{1}{M}\sum_{n=0}^{M-1}\left(w_M^{(k-1)+(\ell-1)}\right)^{n}
\end{aligned}
$$

For $w_M^{(k-1)+(\ell-1)} \neq 1$ we can apply the summation formula of the geometric series that yields 0:

$$
\begin{aligned}
\sum_{n=0}^{M-1}\left(w_M^{(k-1)+(\ell-1)}\right)^{n} &= \frac{1-\left(w_M^{(k-1)+(\ell-1)}\right)^{M}}{1-w_M^{(k-1)+(\ell-1)}} = \frac{1-\left(w_M^{M}\right)^{(k-1)+(\ell-1)}}{1-w_M^{(k-1)+(\ell-1)}} \\
&= \frac{1-1}{1-w_M^{(k-1)+(\ell-1)}} = 0
\end{aligned}
$$

For $k, \ell \in \{1, \ldots, M\}$ the case $w_M^{(k-1)+(\ell-1)} = 1$ occurs only if $(k-1)+(\ell-1) = 0$ or $(k-1)+(\ell-1) = M$. This is equal to the condition $k = 1$ and $\ell = 1$ or alternatively $\ell = M + 2 - k$ with $k > 1$. Thus the nonzero matrix elements are

$$
\left(\mathbf{F}_M \cdot \mathbf{J}_M \cdot \mathbf{F}_M^{-1}\right)(1,1) = 1, \qquad \left(\mathbf{F}_M \cdot \mathbf{J}_M \cdot \mathbf{F}_M^{-1}\right)(k, M+2-k) = w_M^{k-1} \quad k \geq 2,
$$

which proves the assertion. ■

Proposition 6.10. *The characteristic polynomial of the evolution matrix* $\mathbf{E}_R = \mathbf{T}_R\mathbf{C}$ *pertaining to the lattice-Boltzmann algorithm (6.38) with bounce-back boundary conditions is given by*

$$
\chi_R(\lambda) = (\alpha + \beta - \lambda)(\beta - \alpha - \lambda)\prod_{m=1}^{N-1}\left[\lambda^2 - 2\alpha\cos(\pi m/N)\lambda + \alpha^2 - \beta^2\right]. \tag{6.39}
$$

where the coefficients α, β *of the collision block are defined in* (6.37).

Proof: Fourier transformation of the evolution matrix $\mathbf{E}_R$ yields

$$
\begin{aligned}
\hat{\mathbf{E}}_R &= \mathbf{F}\,\mathbf{E}_R\,\mathbf{F}^{-1} = \mathbf{F}\,\mathbf{L}\left(\alpha\mathbf{I} + \beta\mathbf{J}\right)\mathbf{F}^{-1} \\
&= \underbrace{\mathbf{F}\,\mathbf{L}\mathbf{F}^{-1}}\left(\alpha\mathbf{I} + \beta\underbrace{\mathbf{F}\,\mathbf{J}\,\mathbf{F}^{-1}}\right) \\
&= \mathbf{W}\left(\alpha\mathbf{I} + \beta\mathbf{F}\,\mathbf{J}\,\mathbf{F}^{-1}\right)
\end{aligned}
$$

where we have applied corollary 6.1 to compute the left underbraced term. The right underbraced matrix product has been figured out in the proof of lemma 6.16. Thus we obtain

$$
\begin{aligned}
\mathbf{WFJF}^{-1} &= \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & w^1 & & \\ \vdots & & \ddots & \\ 0 & & & w^{2N-1} \end{array}\right) \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & w^1 \\ \vdots & & \iddots & \\ 0 & w^{2N-1} & & \end{array}\right) \\
&= \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & w^2 \\ \vdots & & \iddots & \\ 0 & w^{4N-2} & & \end{array}\right),
\end{aligned}
$$

which gives finally

$$
\hat{\mathbf{E}}_R = \left(\begin{array}{c|ccccccc}
\alpha+\beta & 0 & & & \dots & & & 0 \\ \hline
0 & \alpha w^1 & & & & & & \beta w^2 \\
 & & \ddots & & & & \iddots & \\
 & & & \alpha w^{N-1} & 0 & \beta w^{2N-2} & & \\
\vdots & & & 0 & -\alpha+\beta & 0 & & \\
 & & & \beta w^{2N+2} & 0 & \alpha w^{N+1} & & \\
 & & \iddots & & & & \ddots & \\
0 & \beta w^{4N-2} & & & & & & \alpha w^{2N-1}
\end{array}\right).
$$

Since the characteristic polynomial is invariant under similarity transformations like the discrete Fourier transformation, we find

$$
\begin{aligned}
\det(\mathbf{E}_R - \lambda\mathbf{I}) &= \det(\hat{\mathbf{E}}_R - \lambda\mathbf{I}) \\
&= (\alpha+\beta-\lambda)(\beta-\alpha-\lambda)\prod_{m=1}^{N-1} \det\begin{pmatrix} \alpha w^m - \lambda & \beta w^{2m} \\ \beta w^{4N-2m} & \alpha w^{2N-m} - \lambda \end{pmatrix} \\
&= (\alpha+\beta-\lambda)(\beta-\alpha-\lambda)\prod_{m=1}^{N-1} \left[(\alpha w^m - \lambda)(\alpha \overline{w}^m - \lambda) - \beta^2\right] \\
&= (\alpha+\beta-\lambda)(\beta-\alpha-\lambda)\prod_{m=1}^{N-1} \left[\lambda^2 - \alpha(w^m + \overline{w}^m)\lambda + \alpha^2 - \beta^2\right].
\end{aligned}
$$

Substituting $w = \mathrm{e}^{\mathrm{i}\pi/N}$ leads to (6.39). ■

Sign inverting bounce-back inducing Dirichlet conditions

If we modify (6.38) by writing the second and third equation with a minus sign on the right hand side

$$
\begin{aligned}
\mathsf{F}_{+1}(t_{n+1}, x_1) &= -\mathsf{C}_{-1}(t_{n+1}, x_1) \\
\mathsf{F}_{-1}(t_{n+1}, x_N) &= -\mathsf{C}_{+1}(t_{n+1}, x_N)
\end{aligned}
\tag{6.40}
$$

the lattice-Boltzmann algorithm realizes homogeneous Dirichlet boundary condition for the mass moment U. In the case of the four node system the transport matrix looks like

$$\mathbf{T}_D = \left(\begin{array}{cccc|cccc} \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right). \tag{6.41}$$

$\mathbf{T}_D$ is distinguished from the leftshift matrix $\mathbf{L} = \mathbf{T}_R$ only with respect to the matrix elements $\mathbf{T}_D(N, N+1)$ and $\mathbf{T}_D(2N, 1)$ that are equal to -1 instead of 1. So the sign of the "bounced" collision products is flipped.
Of course, we can not directly apply the results about the shift matrices to diagonalize $\mathbf{T}_D$. However, it turns out, that the two modified matrix elements have no crucial impact on the spectrum and the eigenvectors.

Lemma 6.17. *The characteristic polynomial of the $2N \times 2N$ transport matrix $\mathbf{T}_D$ is $\lambda \mapsto \lambda^{2N} - 1$. Hence the unit roots of $2N^{\text{th}}$ order, i.e. $1, w, w^2, \ldots, w^{2N-1}$ are the eigenvalues of $\mathbf{T}_D$.*

Let us consider the special case of $N = 3$, that demonstrates the general situation. Expanding the determinant by Leibniz' rule with respect to the first column, yields an upper and lower triangular matrix, whose determinants are given by the product of the diagonal elements. Thus the assertion becomes directly verified:

$$\begin{aligned} \det(\mathbf{T}_D - \lambda I) &= \begin{vmatrix} -\lambda & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & -\lambda & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & -\lambda & -1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & -\lambda & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & -\lambda & 1 \\ -1 & \cdot & \cdot & \cdot & \cdot & -\lambda \end{vmatrix} \\ &= -\lambda \begin{vmatrix} -\lambda & 1 & \cdot & \cdot & \cdot \\ \cdot & -\lambda & -1 & \cdot & \cdot \\ \cdot & \cdot & -\lambda & 1 & \cdot \\ \cdot & \cdot & \cdot & -\lambda & 1 \\ \cdot & \cdot & \cdot & \cdot & -\lambda \end{vmatrix} + (-1)\cdot(-1)^{6+1} \begin{vmatrix} 1 & \cdot & \cdot & \cdot & \cdot \\ -\lambda & 1 & \cdot & \cdot & \cdot \\ \cdot & -\lambda & -1 & \cdot & \cdot \\ \cdot & \cdot & -\lambda & 1 & \cdot \\ \cdot & \cdot & \cdot & -\lambda & 1 \end{vmatrix} \\ &= (-\lambda)\cdot(-\lambda)^{6-1} + (-1)^8 \cdot (-1) \;=\; \lambda^6 - 1. \end{aligned}$$

The eigenvector $\mathbf{U}_m$ associated to the eigenvalue w^m, $(m \in \{1, ..., 2N\})$ is easily computed by exploiting the recursive relation between its components provided by the equation $\mathbf{T}_D \cdot \mathbf{U}_m = w^m \mathbf{U}_m$. Setting the first component of the eigenvector equal to 1 yields

$$\mathbf{U}_m(k) = (-1)^{\left\lfloor \frac{k-1}{N} \right\rfloor} w^{m(k-1)}$$

for its k^{th} component. While the first half of the components agrees with the components of the corresponding eigenvector pertaining to $\mathbf{L}$, the second half differs by a minus sign.

We arrange the $2N$ eigenvectors of $\mathbf{T}_D$ in the columns[42] of a $2N \times 2N$ matrix denoted by $\mathbf{G}$:

$$\mathbf{G}(k,\ell) := \frac{(-1)^{\left\lfloor \frac{k-1}{N} \right\rfloor}}{\sqrt{2N}} w^{(k-1)(\ell-1)} \qquad k,\ell \in \{1,\dots,2N\},$$

Like $\mathbf{F}$, the matrix $\mathbf{G}$ is *unitary* too. In order to see this, we write the $2N \times 2N$ inverse Fourier matrix $\mathbf{F}^{-1} = \mathbf{F}^* = \overline{\mathbf{F}}$ as a 2×2 block matrix of four $N \times N$ blocks.

$$\overline{\mathbf{F}} = \mathbf{F}^* = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{pmatrix} \tag{6.42}$$

The symmetry of $\mathbf{F}$, i.e. $\mathbf{F} = \mathbf{F}^\top$, has been taken into account by setting for the upper right and lower left block $\mathbf{B}$ and $\mathbf{B}^\top$ respectively. Moreover the subblocks $\mathbf{A}$ and $\mathbf{D}$ are supposed to be symmetric too. Due to the above mentioned relation between the eigenvectors of $\mathbf{L}$ and $\mathbf{T}_D$ we get

$$\mathbf{G} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^\top & -\mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & -\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{pmatrix}.$$

Obviously $\mathbf{G}$ can be written as product of two unitary matrices. Since the set of unitary matrices with fixed dimension forms a group and is thus closed under matrix multiplication, this entails the unitarity of $\mathbf{G}$. Let us note this result in the next lemma followed by an alternative proof.

Lemma 6.18. *The matrix* $\mathbf{G}$ *is* unitary*:* $\quad \mathbf{G}\mathbf{G}^* = \mathbf{I} \quad \Leftrightarrow \quad \mathbf{G}^{-1} = \mathbf{G}^*$.

Proof:

$$\begin{aligned}
\sum_{\ell=1}^{2N} \mathbf{G}(k,\ell)\mathbf{G}^*(\ell,m) &= \frac{1}{2N}\sum_{\ell=1}^{2N}(-1)^{\left\lfloor \frac{k-1}{N} \right\rfloor} w^{(k-1)(\ell-1)}(-1)^{\left\lfloor \frac{m-1}{N} \right\rfloor}\overline{w}^{(m-1)(\ell-1)} \\
&= (-1)^{\left\lfloor \frac{k-1}{N} \right\rfloor}(-1)^{\left\lfloor \frac{m-1}{N} \right\rfloor}\frac{1}{2N}\sum_{\ell=0}^{2N-1} w^{(k-1)\ell} w^{-(m-1)\ell} \\
&= (-1)^{\left\lfloor \frac{k-1}{N} \right\rfloor}(-1)^{\left\lfloor \frac{m-1}{N} \right\rfloor}\frac{1}{2N}\sum_{\ell=0}^{2N-1} w^{(k-m)\ell} \\
&= (-1)^{\left\lfloor \frac{k-1}{N} \right\rfloor}(-1)^{\left\lfloor \frac{m-1}{N} \right\rfloor}\frac{1}{2N}\sum_{\ell=0}^{2N-1} \left(w^{k-m}\right)^\ell \\
&= (-1)^{\left\lfloor \frac{k-1}{N} \right\rfloor}(-1)^{\left\lfloor \frac{m-1}{N} \right\rfloor}\begin{cases} k = m: & \frac{1}{2N}\sum_{\ell=0}^{2N-1} 1 = 1 \\ k \neq m: & \frac{1}{2N}\frac{1-w^{2N(k-m)}}{1-w^{(k-m)}} = \frac{1}{2N}\frac{1-1}{1-w} = 0 \end{cases} \\
&= \delta_{k,m}
\end{aligned}$$

■

By construction, the matrix $\mathbf{G}$ diagonalizes the transport matrix $\mathbf{T}_D$, i.e.:

$$\mathbf{G}^{-1}\mathbf{T}_D\mathbf{G} = \mathbf{G}^*\mathbf{T}_D\mathbf{G} = \mathrm{diag}(1, w, \dots, w^{2N-1}) = \mathbf{W}.$$

Analogously to the previous paragraph, we need to compute the similarity transformation of $\mathbf{J}$ with respect to $\mathbf{G}$.

[42] Observe that the column-vectors of $\mathbf{F}^* = \overline{\mathbf{F}}$ are the eigenvectors of the shift matrix $\mathbf{L}$. Thus $\mathbf{G}$ plays the analogous role to $\mathbf{F}^*$.

Lemma 6.19. *The following equation holds true:* $\mathbf{G}^*\mathbf{J}\mathbf{G} = -\mathbf{F}\mathbf{J}\mathbf{F}^*$.

Proof: We use the block matrix representation of $\mathbf{F}^*$ in (6.42). Keeping the symmetry of $\mathbf{F}$ in mind, we get for the right product

$$\begin{aligned}
\mathbf{FJF}^* &= \begin{pmatrix} \overline{\mathbf{A}} & \overline{\mathbf{B}} \\ \overline{\mathbf{B}}^\top & \overline{\mathbf{D}} \end{pmatrix} \begin{pmatrix} 0 & \mathbf{J} \\ \mathbf{J} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{pmatrix} \\
&= \begin{pmatrix} \overline{\mathbf{B}}\mathbf{J} & \overline{\mathbf{A}}\mathbf{J} \\ \overline{\mathbf{D}}\mathbf{J} & \overline{\mathbf{B}}^\top\mathbf{J} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{D} \end{pmatrix} \\
&= \begin{pmatrix} \overline{\mathbf{B}}\mathbf{JA} + \overline{\mathbf{A}}\mathbf{JB}^\top & \overline{\mathbf{B}}\mathbf{JB} + \overline{\mathbf{A}}\mathbf{JD} \\ \overline{\mathbf{D}}\mathbf{JA} + \overline{\mathbf{B}}^\top\mathbf{JB}^\top & \overline{\mathbf{D}}\mathbf{JB} + \overline{\mathbf{B}}^\top\mathbf{JD} \end{pmatrix}
\end{aligned}$$

The left side becomes

$$\begin{aligned}
\mathbf{G}^*\mathbf{JG} &= \begin{pmatrix} \overline{\mathbf{A}} & -\overline{\mathbf{B}} \\ \overline{\mathbf{B}}^\top & -\overline{\mathbf{D}} \end{pmatrix} \begin{pmatrix} 0 & \mathbf{J} \\ \mathbf{J} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^\top & -\mathbf{D} \end{pmatrix} \\
&= \begin{pmatrix} -\overline{\mathbf{B}}\mathbf{J} & \overline{\mathbf{A}}\mathbf{J} \\ -\overline{\mathbf{D}}\mathbf{J} & \overline{\mathbf{B}}^\top\mathbf{J} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^\top & -\mathbf{D} \end{pmatrix} \\
&= \begin{pmatrix} -\overline{\mathbf{B}}\mathbf{JA} - \overline{\mathbf{A}}\mathbf{JB}^\top & -\overline{\mathbf{B}}\mathbf{JB} - \overline{\mathbf{A}}\mathbf{JD} \\ -\overline{\mathbf{D}}\mathbf{JA} - \overline{\mathbf{B}}^\top\mathbf{JB}^\top & -\overline{\mathbf{D}}\mathbf{JB} - \overline{\mathbf{B}}^\top\mathbf{JD} \end{pmatrix},
\end{aligned}$$

which is obviously the negative of $\mathbf{FJF}^*$. ■

From lemma 6.16 we obtain

$$\mathbf{G}^*\mathbf{JG} = \left(\begin{array}{c|cccc} -1 & 0 & & \dots & 0 \\ \hline 0 & & & & -w^1 \\ \vdots & & & \iddots & \\ 0 & -w^{2N-1} & & & \end{array}\right). \tag{6.43}$$

Now we are able to compute the characteristic polynomial of the evolution matrix.

Proposition 6.11. *The evolution matrix* $\mathbf{E}_D = \mathbf{T}_D\mathbf{C}$ *of the lattice-Boltzmann algorithm with sign reversing bounce-back boundary conditions has the characteristic polynomial*

$$\chi_D(\lambda) = (\alpha - \beta - \lambda)(-\alpha - \beta - \lambda) \prod_{m=1}^{N-1} \left[\lambda^2 - 2\alpha\cos(\pi m/N)\lambda + \alpha^2 - \beta^2\right] \tag{6.44}$$

with α, β *as given in* (6.37).

Proof: Taking the computation in the proof of proposition (6.10) as guidance, we set

$$\tilde{\mathbf{E}}_D := \mathbf{G}^*\mathbf{E}_D\mathbf{G} = \mathbf{W}\big(\alpha\mathbf{I} + \beta\mathbf{G}^*\mathbf{JG}\big).$$

Using (6.43) and the fact[43] that the product of a diagonal matrix with a "reflected" diagonal matrix results into a "reflected" diagonal matrix, $\tilde{\mathbf{E}}_D$ becomes explicitly

$$\tilde{\mathbf{E}}_D = \left(\begin{array}{c|ccccccc} \alpha-\beta & 0 & & & \dots & & & 0 \\ \hline 0 & \alpha w^1 & & & & & & -\beta w^2 \\ & & \ddots & & & & \iddots & \\ & & & \alpha w^{N-1} & & -\beta w^{2N-2} & & \\ \vdots & & & & -\alpha-\beta & & & \\ & & & -\beta w^{2N+2} & & \alpha w^{N+1} & & \\ & & \iddots & & & & \ddots & \\ 0 & -\beta w^{4N-2} & & & & & & \alpha w^{2N-1} \end{array}\right).$$

The X-shape structure of the $(2N-1)\times(2N-1)$ block permits us to read off the characteristic polynomial like in theorem 6.2 and proposition 6.10. ■

Comparison of eigenvalues

The computation of the spectra pertaining to the lattice-Boltzmann evolution matrix with different boundary conditions enables us to draw some parallels between the diffusion-reaction equation (A.1) (with $a = 0$) and the lattice-Boltzmann algorithm discretizing this equation.
From theorem A.1 and A.2 (see appendix), we know the eigenvalues of the BVP for the diffusion-reaction operator $(-\nu\partial_x^2 + c)$:

$$\pi^2 m^2 \nu + c \quad \text{with} \begin{cases} m \in \mathbb{Z} & \text{for periodic BCs} \\ m \in \mathbb{N}_0 & \text{for Dirichlet BCs} \\ m \in \mathbb{N} & \text{for Robin BCs} \end{cases} \tag{6.45}$$

So we can formulate the following two observations:

- The diffusion-reaction operator attains the same eigenvalues for all three boundary conditions (but not the same eigenfunctions!). Only exception represents c, which is **no** eigenvalue in the case of Dirichlet boundary conditions.
- In the case of periodic boundary conditions, the eigenvalues are twofold degenerated, as they are parameterized by $m \in \mathbb{Z}$ but depend only on m^2. Only exception is again the eigenvalue c obtained for $m = 0$.

An analogous statement refers to the eigenvalues of the evolution operators for the associated IBVPs. From section 6.2.3 we know already, that these eigenvalues are essentially related by the exponential function to the eigenvalues of the diffusion-reaction operator. In particular, if c vanishes, 0 is an eigenvalue of the BVP with either periodic or Robin boundary conditions. Its counterpart with respect to the

---

[43]

$$\begin{pmatrix} a & \cdot & \cdot \\ \cdot & b & \cdot \\ \cdot & \cdot & c \end{pmatrix} \begin{pmatrix} \cdot & \cdot & e \\ \cdot & f & \cdot \\ g & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \cdot & \cdot & ae \\ \cdot & bf & \cdot \\ cg & \cdot & \cdot \end{pmatrix}$$

evolution operator is 1, which implies the existence of conserved states.

It turns out that the evolution matrix of the lattice-Boltzmann algorithm shares a lot of these properties. To see this better, we write down once more the characteristic polynomials[44] that we have obtained for the different boundary conditions:

$$\begin{aligned}
\text{periodic BC:} \quad & \chi_P(\lambda) = \prod_{m=0}^{N-1} \left[\lambda^2 - 2\alpha\cos(\mathbf{2}\pi m/N)\lambda + \alpha^2 - \beta^2\right] \\
\text{ind. Robin BC:} \quad & \chi_R(\lambda) = (\lambda^2 - 2\beta\lambda + \beta^2 - \alpha^2) \prod_{m=1}^{N-1} \left[\lambda^2 - 2\alpha\cos(\pi m/N)\lambda + \alpha^2 - \beta^2\right] \\
\text{ind. Dirichlet BC:} \quad & \chi_D(\lambda) = (\lambda^2 + 2\beta\lambda + \beta^2 - \alpha^2) \prod_{m=1}^{N-1} \left[\lambda^2 - 2\alpha\cos(\pi m/N)\lambda + \alpha^2 - \beta^2\right]
\end{aligned}$$

The identity of the cosine, $\cos(2\pi m/N) = \cos(2\pi(N-m)/N) = \cos(2\pi - 2\pi m/N)$, effects, that almost all factor polynomials of χ_P occur twice in the product. So we obtain for even N

$$\begin{aligned}
\chi_P(\lambda) = (\lambda^2 - 2\alpha\lambda + \alpha^2 - \beta^2)(\lambda^2 + 2\alpha\lambda + \alpha^2 - \beta^2) \\
\times \prod_{\substack{1<m<N \\ m \text{ even}}} \left[\lambda^2 - 2\alpha\cos(\pi m/N) + \alpha^2 - \beta^2\right]^2,
\end{aligned}$$

while we get for odd N

$$\chi_P(\lambda) = (\lambda^2 - 2\alpha\lambda + \alpha^2 - \beta^2) \prod_{\substack{1<m<N \\ m \text{ even}}} \left[\lambda^2 - 2\alpha\cos(\pi m/N) + \alpha^2 - \beta^2\right]^2, \quad (6.46)$$

since $N/2$ is no integer in this case.
Evidently, all characteristic polynomials contain quadratic factor polynomials

$$\chi_m(\lambda) = \lambda^2 - 2\alpha\cos(\pi m/N)\lambda + \alpha^2 - \beta^2 \qquad \text{with } m \in \{1, \ldots, N-1\},$$

whose roots are given by

$$\lambda_m^\pm := \alpha\cos(\pi m/N) \pm \sqrt{\alpha^2\cos^2(\pi m/N) - \alpha^2 + \beta^2}. \quad (6.47)$$

In the case of periodic boundary conditions the χ_m's appear only if m is an even number of the above range; but they are squared in exchange.
Only by the quadratic polynomials in front of the product signs the characteristic polynomials of $\mathbf{E}_P, \mathbf{E}_R$ and $\mathbf{E}_D$ differ from each other. The associated zeros characterize the spectra as shown in table 6.1.

[44] Observe that we also set $a = 0$, i.e. $\alpha = \delta$ and $\beta = \gamma$, in the case of periodic boundary conditions.

kinetic BCs	induced macroscopic BCs	characterizing eigenvalues
bounce-back	Robin (natural)	$\alpha+\beta, \quad \beta-\alpha$
bb. with sign-flipping	Dirichlet	$-\alpha-\beta, \quad \alpha-\beta$
periodic, N arbitrary	periodic	$\alpha+\beta, \quad \alpha-\beta$
periodic, N even	"	$-\alpha-\beta, \quad \beta-\alpha$

Table 6.1: To which boundary condition the spectrum refers is best distinguished by the characterizing eigenvalues. Note: $\alpha+\beta = 1-ch^2$ and $\alpha-\beta = 1-\omega$. These values are well known as eigenvalues of the collision block.

The feature of the diffusion-reaction equation summarized in (6.45) is reflected by the lattice-Boltzmann algorithm in the following way:

- The spectra of the three evolution matrices obey the following mutual inclusions:

$$\begin{aligned} \operatorname{spec}(\mathbf{E}_R,\omega,0,p,N)\cup\{-\alpha-\beta,\ \alpha-\beta\} &= \operatorname{spec}(\mathbf{E}_P,\omega,0,p,\mathbf{2}N),\\ \operatorname{spec}(\mathbf{E}_D,\omega,0,p,N)\cup\{\ \alpha+\beta,\ \beta-\alpha\} &= \operatorname{spec}(\mathbf{E}_P,\omega,0,p,\mathbf{2}N),\\ \operatorname{spec}(\mathbf{E}_R,\omega,0,p,N)\cap\operatorname{spec}(\mathbf{E}_D,\omega,0,p,N) &= \{\lambda_m^{\pm}\mid 1\le m\le N-1\}. \end{aligned}$$

- With exception of the characterizing eigenvalues (including 1 in the purely diffusive case), all eigenvalues of $\mathbf{E}_P$ are twofold degenerated.

The relations above imply the equality of the spectral limit sets corresponding to the three boundary conditions. Hence we are allowed to conclude that the spectral limit set is contained in the closed unit disk if and only if $\omega\in[0,2]$ (compare with proposition 6.4). This entails that $\omega\in[0,2]$ is a *necessary stability condition* in the case of the presented boundary conditions. The *sufficiency* for stability (e.g. with respect to the infinity norm) can be proved literally like proposition 6.1.

Additionally, we get another consequence: Between two neighboring eigenvalues of $\mathbf{E}_P$ there lies a common eigenvalue of $\mathbf{E}_R$ and $\mathbf{E}_D$, provided the three matrices refer to the same number of grid nodes.

6.4.2 Computing eigenbases

For the computation of the eigenvectors it is crucial to observe that the transport matrices $\mathbf{T}_R, \mathbf{T}_D$ and the collision matrix $\mathbf{C}$ have common invariant subspaces. For this let us introduce the subspaces of

$$\begin{aligned} &\textit{positively repetitive}\text{ vectors}\quad \mathbb{C}^{2N}_{r+} := \left\{\mathbf{V}\in\mathbb{C}^{2N}:\ \exists\mathbf{v}\in\mathbb{C}^N\ \text{ s.th. }\ \mathbf{V}=\tbinom{\mathbf{v}}{\mathbf{v}}\right\},\\ &\textit{negatively repetitive}\text{ vectors}\quad \mathbb{C}^{2N}_{r-} := \left\{\mathbf{V}\in\mathbb{C}^{2N}:\ \exists\mathbf{v}\in\mathbb{C}^N\ \text{ s.th. }\ \mathbf{V}=\tbinom{\mathbf{v}}{-\mathbf{v}}\right\}. \end{aligned}$$

Since every vector $\binom{\mathbf{u}}{\mathbf{w}}\in\mathbb{C}^{2N}$ admits the decomposition

$$\begin{pmatrix}\mathbf{u}\\ \mathbf{w}\end{pmatrix} = \underbrace{\frac{1}{2}\begin{pmatrix}\mathbf{u}+\mathbf{w}\\ \mathbf{u}+\mathbf{w}\end{pmatrix}}_{\in\mathbb{C}^{2N}_{r+}} + \underbrace{\frac{1}{2}\begin{pmatrix}\mathbf{u}-\mathbf{w}\\ \mathbf{w}-\mathbf{u}\end{pmatrix}}_{\in\mathbb{C}^{2N}_{r-}}$$

with $\mathbf{u}, \mathbf{w} \in \mathbb{C}^N$, it becomes obvious that $\mathbb{C}^{2N}$ is the *direct, orthogonal* sum of these subspaces, i.e. $\mathbb{C}^{2N} = \mathbb{C}^{2N}_{r+} \oplus \mathbb{C}^{2N}_{r-}$.

It turns out, that these subspaces remain invariant under the application of $\mathbf{T}_R, \mathbf{T}_D$ and $\mathbf{C}$. Introducing the $N \times N$ matrix $\tilde{\mathbf{L}}$, which shall be equal to the $N \times N$ left-shift matrix with the exception of the matrix element $\tilde{\mathbf{L}}(N,1) = -1$, the following identities can be easily checked confirming immediately the invariance properties of the transport matrices:

$$\mathbf{T}_R \begin{pmatrix} \mathbf{v} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{L} & 0 \\ 0 & \mathbf{L} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{Lv} \\ \mathbf{Lv} \end{pmatrix} \qquad \mathbf{T}_R \begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}} & 0 \\ 0 & \tilde{\mathbf{L}} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}}\mathbf{v} \\ -\tilde{\mathbf{L}}\mathbf{v} \end{pmatrix}$$

$$\mathbf{T}_D \begin{pmatrix} \mathbf{v} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}} & 0 \\ 0 & \tilde{\mathbf{L}} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}}\mathbf{v} \\ \tilde{\mathbf{L}}\mathbf{v} \end{pmatrix} \qquad \mathbf{T}_D \begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{L} & 0 \\ 0 & \mathbf{L} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{Lv} \\ -\mathbf{Lv} \end{pmatrix}$$

The transport matrices operate on each of the subspaces like 2×2 diagonal block matrices whose $N \times N$ blocks are either given by $\mathbf{L}$ or by $\tilde{\mathbf{L}}$, depending on whether they are applied to a positively or negatively repetitive vector. The collision matrix reveals the same 2×2 block structure:

$$\mathbf{C} = \alpha \mathbf{I} + \beta \mathbf{J} = \begin{pmatrix} \alpha \mathbf{I} & \beta \mathbf{J} \\ \beta \mathbf{J} & \alpha \mathbf{I} \end{pmatrix}.$$

Notice that $\mathbf{I}$ and $\mathbf{J}$ are understood as $2N \times 2N$ matrices at the left hand side, while they represent $N \times N$ blocks at the right. Hence we get

$$\mathbf{C} \begin{pmatrix} \mathbf{v} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \alpha \mathbf{v} + \beta \mathbf{Jv} \\ \beta \mathbf{Jv} + \alpha \mathbf{v} \end{pmatrix} \in \mathbb{C}^{2N}_{r+} \quad \text{and} \quad \mathbf{C} \begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix} = \begin{pmatrix} \alpha \mathbf{v} - \beta \mathbf{Jv} \\ \beta \mathbf{Jv} - \alpha \mathbf{v} \end{pmatrix} \in \mathbb{C}^{2N}_{r-}$$

which proves the invariance of $\mathbb{C}^{2N}_{r+}$ and $\mathbb{C}^{2N}_{r-}$ also with respect to $\mathbf{C}$.

The decisive issue of these observations consists in reducing by half the dimension of the linear system to be solved for finding the eigenvectors. In the sequel we focus on the computation of the eigenvectors pertaining to $\mathbf{E}_R$; however the computation should be easily adapted to the case of $\mathbf{E}_D$.

Eigenvectors of $\mathbf{E}_R$ in $\mathbb{C}^{2N}_{r+}$. Let us start with the lattice-Boltzmann evolution matrix $\mathbf{E}_R$ inducing natural, Robin type boundary conditions. If λ is an eigenvalue, then the equation $\mathbf{E}_R \mathbf{V} = \lambda \mathbf{V}$ has a nontrivial solution, that we seek now in $\mathbb{C}^{2N}_{r+}$. Setting $\mathbf{V} = \binom{\mathbf{v}}{\mathbf{v}}$ and using the results from above we get

$$\mathbf{E}_R \mathbf{V} = \lambda \mathbf{V} \quad \Rightarrow \quad \begin{pmatrix} \mathbf{L} & 0 \\ 0 & \mathbf{L} \end{pmatrix} \begin{pmatrix} \alpha \mathbf{I} & \beta \mathbf{J} \\ \beta \mathbf{J} & \alpha \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \alpha \mathbf{Lv} + \beta \mathbf{LJv} \\ \beta \mathbf{LJv} + \alpha \mathbf{Lv} \end{pmatrix} = \begin{pmatrix} \lambda \mathbf{v} \\ \lambda \mathbf{v} \end{pmatrix}.$$

As an identical equation is produced for the upper and lower block, it is enough to consider the dimensionally halved equation from the upper block, i.e.

$$\mathbf{L}(\alpha \mathbf{I} + \beta \mathbf{J}) \mathbf{v} = \lambda \mathbf{v}.$$

Multiplied by $\mathbf{R} = \mathbf{L}^{-1}$, it transforms into

$$\alpha\mathbf{v} + \beta\mathbf{J}\mathbf{v} = \lambda\mathbf{R}\mathbf{v}.$$

Writing this equation line by line (componentwise), we discover how the linear system can be solved by a recursion:

$$\begin{array}{llllll}
\alpha\,\mathbf{v}(1) & + & \beta\,\mathbf{v}(N) & = & \lambda\,\mathbf{v}(N) & \\
\alpha\,\mathbf{v}(k) & + & \beta\,\mathbf{v}(N-k+1) & = & \lambda\,\mathbf{v}(k-1) & \text{for } 2 \le k \le N \\
\alpha\,\mathbf{v}(N-k) & + & \beta\,\mathbf{v}(k+1) & = & \lambda\,\mathbf{v}(N-k-1) & \text{for } 1 \le k \le N-1 \\
\alpha\,\mathbf{v}(N) & + & \beta\,\mathbf{v}(1) & = & \lambda\,\mathbf{v}(N-1) &
\end{array} \tag{6.48}$$

As $\mathbf{J}$ and $\mathbf{R}$ map the N^{th} component of $\mathbf{v}$ to the first component of the corresponding image vector, the first row in (6.48) relates only two components, $\mathbf{v}(1)$ and $\mathbf{v}(N)$, with each other. This enables us to determine $\mathbf{v}(N)$ in dependence of $\mathbf{v}(1)$. Under the assumption $\mathbf{v}(1) \neq 0$, we are allowed to pick up an arbitrary value for $\mathbf{v}(1)$ (besides 0), since eigenvectors are only determined up to a constant on account of the homogeneity of the equation $\mathbf{E}_R\mathbf{V} = \lambda\mathbf{V}$. Thus we obtain

$$\mathbf{v}(1) = 1 \qquad \Rightarrow \qquad \mathbf{v}(N) = \frac{\alpha}{\lambda - \beta}.$$

In order to compute a further component we have to jump to the lowest line of (6.48) expressing $\mathbf{v}(N-1)$ in terms of $\mathbf{v}(1)$ and $\mathbf{v}(N)$. The knowledge of $\mathbf{v}(N-1)$ can be used to calculate $\mathbf{v}(2)$ if we set $k = 2$ in the second equation of (6.48). Writing $\mathbf{v}(N-1)$ and $\mathbf{v}(2)$ in terms of $\mathbf{v}(1)$ and $\mathbf{v}(N)$ yields

$$\begin{pmatrix} \mathbf{v}(2) \\ \mathbf{v}(N-1) \end{pmatrix} = \begin{pmatrix} \alpha^{-1}\lambda - \alpha^{-1}\lambda^{-1}\beta^2 & -\beta\lambda^{-1} \\ \beta\lambda^{-1} & \alpha\lambda^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{v}(1) \\ \mathbf{v}(N) \end{pmatrix}.$$

This cross-over computation is continuable providing successively $\mathbf{v}(N-2), \mathbf{v}(3)$ then $\mathbf{v}(N-3)$, $\mathbf{v}(4)$ and so forth. Thereby $\mathbf{v}(k+1)$ and $\mathbf{v}(N-k-1)$ depend only directly from $\mathbf{v}(k)$ and $\mathbf{v}(N-k)$ just in the same manner as $\mathbf{v}(2), \mathbf{v}(N-1)$ depend on $\mathbf{v}(1), \mathbf{v}(N)$. Introducing the recursion matrix

$$R_\lambda := \begin{pmatrix} \frac{\lambda}{\alpha} - \frac{\beta^2}{\alpha\lambda} & -\frac{\beta}{\lambda} \\ \frac{\beta}{\lambda} & \frac{\alpha}{\lambda} \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

we obtain

$$\begin{pmatrix} \mathbf{v}(k+1) \\ \mathbf{v}(N-k-1) \end{pmatrix} = R_\lambda \begin{pmatrix} \mathbf{v}(k) \\ \mathbf{v}(N-k) \end{pmatrix} = R_\lambda^k \begin{pmatrix} \mathbf{v}(1) \\ \mathbf{v}(N) \end{pmatrix}. \tag{6.49}$$

After $\lceil N/2 \rceil$ iterations all components of $\mathbf{v}$ are known. Until now its seems, as if the equation $\mathbf{E}_R\mathbf{V} = \lambda\mathbf{V}$ has a nontrivial solution in $\mathbb{C}^{2N}_{r+}$ for any λ, because the recursion works apparently always independently of whether λ is an eigenvalue or not. Of course, it is clear from linear algebra (elementary theory of eigenvalue equation in finite dimensional vector spaces) that this is not possible.

$$\begin{array}{ll}
\mathbf{v}(1),\mathbf{v}(8) \xrightarrow{R_\lambda} \mathbf{v}(2),\mathbf{v}(7) & \mathbf{v}(1),\mathbf{v}(7) \xrightarrow{R_\lambda} \mathbf{v}(2),\mathbf{v}(6) \\
\mathbf{v}(2),\mathbf{v}(7) \xrightarrow{R_\lambda} \mathbf{v}(3),\mathbf{v}(6) & \mathbf{v}(2),\mathbf{v}(6) \xrightarrow{R_\lambda} \mathbf{v}(3),\mathbf{v}(5) \\
\mathbf{v}(3),\mathbf{v}(6) \xrightarrow{R_\lambda} \mathbf{v}(4),\mathbf{v}(5) & \mathbf{v}(3),\mathbf{v}(5) \xrightarrow{R_\lambda} \mathbf{v}(4),\mathbf{v}(4) \\
\mathbf{v}(4),\mathbf{v}(5) \xrightarrow{R_\lambda} \mathbf{v}(5),\mathbf{v}(4) & \mathbf{v}(4),\mathbf{v}(4) \xrightarrow{R_\lambda} \mathbf{v}(5),\mathbf{v}(3) \\
\mathbf{v}(5),\mathbf{v}(4) \xrightarrow{R_\lambda} \mathbf{v}(6),\mathbf{v}(3) & \mathbf{v}(5),\mathbf{v}(3) \xrightarrow{R_\lambda} \mathbf{v}(6),\mathbf{v}(2) \\
\mathbf{v}(6),\mathbf{v}(3) \xrightarrow{R_\lambda} \mathbf{v}(7),\mathbf{v}(2) & \mathbf{v}(6),\mathbf{v}(2) \xrightarrow{R_\lambda} \mathbf{v}(7),\mathbf{v}(1) \\
\mathbf{v}(7),\mathbf{v}(2) \xrightarrow{R_\lambda} \mathbf{v}(8),\mathbf{v}(1) &
\end{array}$$

Table 6.2: The recursion for $N = 8$ and $N = 7$. Performing the recursion more than $\lfloor N/2 \rfloor$ times, reproduces already known components in backwards order.

In order to see where the computation breaks down if λ is not an eigenvalue, let us continue the recursion beyond $\lceil N/2 \rceil$ steps. The equations in (6.48) imply, that the components of $\mathbf{v}$ should be reproduced in converse order, if they are well-defined. Table 6.4.2 exemplifies this kind of 'reproduction-property'. In the case of $N = 8$ we get for instance

$$R_\lambda \begin{pmatrix} \mathbf{v}(4) \\ \mathbf{v}(5) \end{pmatrix} = J \begin{pmatrix} \mathbf{v}(4) \\ \mathbf{v}(5) \end{pmatrix} = \begin{pmatrix} \mathbf{v}(5) \\ \mathbf{v}(4) \end{pmatrix}, \quad R_\lambda^3 \begin{pmatrix} \mathbf{v}(3) \\ \mathbf{v}(6) \end{pmatrix} = J \begin{pmatrix} \mathbf{v}(3) \\ \mathbf{v}(6) \end{pmatrix} = \begin{pmatrix} \mathbf{v}(6) \\ \mathbf{v}(3) \end{pmatrix}, \quad \ldots .$$

So we can infer a necessary condition for the reproduction-property to be fulfilled: For even N, the matrices

$$R_\lambda - J, \quad R_\lambda^3 - J, \quad R_\lambda^5 - J, \quad \ldots$$

must be singular, whereas the case of odd N necessitates the matrices

$$R_\lambda^2 - J, \quad R_\lambda^4 - J, \quad R_\lambda^6 - J, \quad \ldots$$

to possess a nontrivial nullspace. We will verify that this condition is satisfied indeed, surprisingly independently of λ.

Lemma 6.20. *For each $n \in \mathbb{N}_0$ there is an $x_n \in \mathbb{R}^2$ such that $R_\lambda^n x_n = J x_n$ or equivalently $(R_\lambda^n - J)x_n = 0$.*

Proof: Let us first discuss the case of n being odd i.e. $n = 2j - 1$ for some $j \in \mathbb{N}$. By checking that the determinant $\det(R_\lambda - J)$ vanishes, the assertion is verified for $n = 1$:

$$\det(R - J) = \begin{vmatrix} \frac{\lambda}{\alpha} - \frac{\beta^2}{\alpha\lambda} & -\frac{\beta}{\lambda} - 1 \\ \frac{\beta}{\lambda} - 1 & \frac{\alpha}{\lambda} \end{vmatrix} = 1 - \frac{\beta^2}{\lambda^2} - (-1)\left(\frac{\beta}{\lambda} + 1\right)\left(\frac{\beta}{\lambda} - 1\right) = 0$$

The general case is treated by observing the following commutation relation of R_λ and J: $JR_\lambda = R_\lambda^{-1} J$ or equivalently $JR_\lambda J = R_\lambda^{-1}$ as $J = J^{-1}$. This is explained by the fact that R_λ is the sum of a diagonal and an antisymmetric matrix (implying

$R_\lambda(1,2) = -R_\lambda(2,1)$). Moreover we have $\det(R_\lambda) = 1$. Indeed, we get for an arbitrary matrix with the mentioned properties:

$$\begin{aligned} J\begin{pmatrix} a & b \\ -b & d \end{pmatrix} J^{-1} &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ -b & d \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} b & a \\ d & -b \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} d & -b \\ b & a \end{pmatrix} = \begin{pmatrix} a & b \\ -b & d \end{pmatrix}^{-1}. \end{aligned}$$

Furthermore $\det(R_\lambda) = 1$ guarantees that R_λ and all its powers are invertible. Hence there is for every $2 \le j \in \mathbb{N}$ an x_{2j-1} with $x_1 = R_\lambda^{j-1} x_{2j-1}$. Exploiting the commutator relation also for the powers of R_λ we obtain the assertion for odd $n = 2j-1$.

$$\begin{aligned} R_\lambda x_1 = J x_1 \quad &\Rightarrow & R_\lambda R_\lambda^{j-1} x_{2j-1} &= J R_\lambda^{j-1} x_{2j-1} \\ &\Leftrightarrow & R_\lambda R_\lambda^{j-1} x_{2j-1} &= R_\lambda^{-j+1} J x_{2j-1} \\ &\Leftrightarrow & R_\lambda^{j-1} R_\lambda R_\lambda^{j-1} x_{2j-1} &= J x_{2j-1} \\ &\Leftrightarrow & R_\lambda^{2j-1} x_{2j-1} &= J x_{2j-1} \end{aligned}$$

Likewise the case of even $n = 2j$ with $j \in \mathbb{N}_0$ is verified. For $n = 0$ the kernel of $R_\lambda^0 - J = I - J$ is trivially given by the multiples of $x_0 = (1,1)^\top$. If we define the vector x_{2j} ($j \in \mathbb{N}$) by the equation $x_0 = R_\lambda^j x_{2j}$, we can proceed as above to show $R_\lambda^{2n} x_{2n} = J x_{2n}$. ∎

In particular, the reproduction property requires the matrix $R_\lambda^{N-1} - J$ to have a nontrivial kernel. Additionally, the nullspace of this matrix is supplied by the known values for $\mathbf{v}(1)$ and $\mathbf{v}(N)$. So, its kernel must correspond to the span of the vector $\begin{pmatrix} 1 \\ \frac{\alpha}{\lambda-\beta} \end{pmatrix}$. Hence:

$$R_\lambda^{N-1}\begin{pmatrix} 1 \\ \frac{\alpha}{\lambda-\beta} \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\lambda-\beta} \\ 1 \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} \frac{\lambda}{\alpha} - \frac{\beta^2}{\alpha\lambda} & -\frac{\beta}{\lambda} \\ \frac{\beta}{\lambda} & \frac{\alpha}{\lambda} \end{pmatrix}^{N-1}\begin{pmatrix} 1 \\ \frac{\alpha}{\lambda-\beta} \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\lambda-\beta} \\ 1 \end{pmatrix}. \tag{6.50}$$

It turns out that this condition is *not* any more satisfied for arbitrary λ. Therefore we have found the equation, which determines (at least some of) the eigenvalues of the lattice-Boltzmann evolution operator $\mathbf{E}_R$. As the above equation seems to be quite complicated, we abstain however from extracting the eigenvalues out of it the more so as we know them already. Instead, we simply show that nearly half of the eigenvalues will do the job.

Lemma 6.21. *Equation (6.50) holds if $\lambda = \lambda_m^\pm$ for* even $m \in \{1, \dots N-1\}$.

Recall, that λ_m^+ and λ_m^- as defined in (6.47) are eigenvalues of $\mathbf{E}_R$.

Proof: Let us first consider the characteristic polynomial of R_λ:

$$\mu \mapsto \det(R_\lambda - \mu I) = \begin{vmatrix} \frac{\lambda}{\alpha} - \frac{\beta^2}{\alpha\lambda} - \mu & -\frac{\beta}{\lambda} \\ \frac{\beta}{\lambda} & \frac{\alpha}{\lambda} - \mu \end{vmatrix} = \mu^2 - \left(\frac{\alpha}{\lambda} + \frac{\lambda}{\alpha} - \frac{\beta^2}{\alpha\lambda}\right)\mu + 1.$$

Multiplying the equation $\det(R_\lambda - \mu I) = 0$ with $-\alpha\mu^{-1}\lambda$ yields

$$\lambda^2 - \alpha(\mu + \mu^{-1})\lambda + \alpha^2 - \beta^2 = 0.$$

Comparing the left hand side with the polynomial χ_m in equation (6.46), replacing $2\cos(\pi m/N)$ by $w^m + w^{-m}$, we see that it is satisfied setting $\mu = w^m = e^{i\pi m/N}$ and $\lambda = \lambda_m^\pm$ with $1 \leq m \leq N-1$. In other words: If λ is equal to the eigenvalue $\lambda_m^\pm$ of the lattice-Boltzmann evolution matrix $\mathbf{E}_R$, then the eigenvalues of the recursion matrix $R_{\lambda_m^\pm}$ are w^m and w^{-m}. This has an important consequence: As $(w^m)^N = w^{mN} = (-1)^m$, we may conclude[45] that $R^N_{\lambda_m^\pm} = I$ if m is even and $R^N_{\lambda_m^\pm} = -I$ if m is odd.
Now we are able to verify (6.50) supposing m to be even. From $R^N_{\lambda_m^\pm} = I$ we obtain $R^{N-1}_{\lambda_m^\pm} R_{\lambda_m^\pm} = I$ and thus $R^{-1}_{\lambda_m^\pm} = R^{N-1}_{\lambda_m^\pm}$. So we have to check whether the equation $R^{-1}_{\lambda_m^\pm}\binom{1}{\alpha/(\lambda_m^\pm-\beta)} = \binom{\alpha/(\lambda_m^\pm-\beta)}{1}$ is satisfied or not. Fortunately, we get a positive result:

$$\begin{pmatrix} \frac{\lambda_m^\pm}{\alpha} - \frac{\beta^2}{\alpha\lambda_m^\pm} & -\frac{\beta}{\lambda_m^\pm} \\ \frac{\beta}{\lambda_m^\pm} & \frac{\alpha}{\lambda_m^\pm} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \frac{\alpha}{\lambda_m^\pm-\beta} \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\lambda_m^\pm} & \frac{\beta}{\lambda_m^\pm} \\ -\frac{\beta}{\lambda_m^\pm} & \frac{\lambda_m^\pm}{\alpha} - \frac{\beta^2}{\alpha\lambda_m^\pm} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\alpha}{\lambda_m^\pm-\beta} \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\lambda_m^\pm-\beta} \\ 1 \end{pmatrix}.$$

This proves the assertion. ■

Clearly, (6.50) can not hold for odd m, since we obtain then the negative inverse $-R^{-1}_{\lambda_m^\pm}$ for $R^{N-1}_{\lambda_m^\pm}$. Below, we will see that, if m is odd, the eigenvectors associated to $\lambda_m^\pm$ lie in the subspace of the negatively repetitive vectors.

Let us designate by $\mathbf{V}_m^\pm$ the (unique) eigenvector of $\mathbf{E}_R$ pertaining to the eigenvalue $\lambda_m^\pm$. So far we know, that, m being even, $\mathbf{V}_m^\pm$ has the structure $\mathbf{V}_m^\pm = \binom{\mathbf{v}_m^\pm}{\mathbf{v}_m^\pm}$ with $\mathbf{v}_m^\pm \in \mathbb{C}^N$. To find a direct and explicit representation for the components of $\mathbf{v}_m^\pm$, the reasoning above and the recursion relation (6.49) yield[46]

$$\begin{pmatrix} \mathbf{v}_m^\pm(k+1) \\ \mathbf{v}_m^\pm(N-k-1) \end{pmatrix} = R^k_{\lambda_m^\pm} \begin{pmatrix} 1 \\ \frac{\alpha}{\lambda_m^\pm-\beta} \end{pmatrix} = B_{\lambda_m^\pm} \begin{pmatrix} w^m & 0 \\ 0 & w^{-m} \end{pmatrix} B^{-1}_{\lambda_m^\pm} \begin{pmatrix} 1 \\ \frac{\alpha}{\lambda_m^\pm-\alpha} \end{pmatrix},$$

where $B_{\lambda_m^\pm}$ denotes the matrix whose columns contain the eigenvectors of $R_{\lambda_m^\pm}$. This motivates the ansatz

$$\mathbf{v}_m^\pm(k) = \kappa_1 w^{mk} + \kappa_2 w^{-mk} \tag{6.51}$$

with the unknowns κ_1, κ_2 depending on m and $\lambda_m^\pm$. The coefficients can be determined by the subsequent 2×2 system resulting from the 'initial condition' for $k=1$ and $k=N$

$$\begin{aligned} \mathbf{v}_m^\pm(1) &= 1 = \kappa_1 w^m + \kappa_2 w^{-m} \\ \mathbf{v}_m^\pm(N) &= \tfrac{\alpha}{\lambda_m^\pm-\beta} = \kappa_1(-1)^m + \kappa_2(-1)^m = \kappa_1 + \kappa_2 \end{aligned}$$

where we have employed our assumption of m being even. Solving for κ_1, κ_2 leads finally to the desired formula:

$$\mathbf{v}_m^\pm(k) = \frac{1}{w^m - w^{-m}} \left[\left(1 - \frac{\alpha w^{-m}}{\lambda_m^\pm - \beta}\right) w^{mk} - \left(1 - \frac{\alpha w^m}{\lambda_m^\pm - \beta}\right) w^{-mk} \right]. \tag{6.52}$$

[45] Observe that $w^m \neq w^{-m}$ holds for $m \in \{1, \ldots, N-1\}$. Therefore $R_{\lambda_m^\pm}$ is diagonalizable entailing either $R^N_{\lambda_m^\pm} = I$ or $R^N_{\lambda_m^\pm} = -I$.

[46] Here we use once again the diagonalizability of $R_{\lambda_m^\pm}$

Eigenvectors of $\mathbf{E}_R$ in $\mathbb{C}^{2N}_{r-}$. In order to address the eigenvectors in the subspace of negatively repetitive vectors we have basically to repeat the steps from the previous paragraph. Therefore we will shorten the presentation.
The ansatz of an eigenvector $\mathbf{V} = \binom{\mathbf{v}}{-\mathbf{v}} \in \mathbb{C}^{2N}_{r+}$ leads also to a reduction of the dimension:

$$\mathbf{E}_R\mathbf{V} = \lambda\mathbf{V} \quad \Rightarrow \quad \begin{pmatrix} \tilde{\mathbf{L}} & 0 \\ 0 & \tilde{\mathbf{L}} \end{pmatrix} \begin{pmatrix} \alpha\mathbf{I} & \beta\mathbf{J} \\ \beta\mathbf{J} & \alpha\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ -\mathbf{v} \end{pmatrix} = \begin{pmatrix} \alpha\tilde{\mathbf{L}}\mathbf{v} - \beta\tilde{\mathbf{L}}\mathbf{J}\mathbf{v} \\ -\beta\tilde{\mathbf{L}}\mathbf{J}\mathbf{v} + \alpha\tilde{\mathbf{L}}\mathbf{v} \end{pmatrix} = \begin{pmatrix} \lambda\mathbf{v} \\ -\lambda\mathbf{v} \end{pmatrix}.$$

Since we get equivalent equations in both of the blocks, it suffices to solve

$$\tilde{\mathbf{L}}(\alpha\mathbf{I} - \beta\mathbf{J})\mathbf{v} = \lambda\mathbf{v} \qquad \Leftrightarrow \qquad \alpha\mathbf{v} - \beta\mathbf{J}\mathbf{v} = \lambda\tilde{\mathbf{R}}\mathbf{v}.$$

In components the latter equation becomes:

$$\begin{array}{llllll} \alpha\,\mathbf{v}(1) & - & \beta\,\mathbf{v}(N) & = & -\lambda\,\mathbf{v}(N) & \\ \alpha\,\mathbf{v}(k) & - & \beta\,\mathbf{v}(N-k+1) & = & \lambda\,\mathbf{v}(k-1) & \text{for } 2 \le k \le N \\ \alpha\,\mathbf{v}(N-k) & - & \beta\,\mathbf{v}(k+1) & = & \lambda\,\mathbf{v}(N-k-1) & \text{for } 1 \le k \le N-1 \\ \alpha\,\mathbf{v}(N) & - & \beta\,\mathbf{v}(1) & = & \lambda\,\mathbf{v}(N-1) & \end{array} \tag{6.53}$$

Unlike equation (6.53) there are now minus signs in front of all β-terms and in front of the λ-term in the first line. Setting once more $\mathbf{v}(1) = 1$ this yields the same value for $\mathbf{v}(N)$ but with reverse sign:

$$\mathbf{v}(1) = 1 \qquad \Rightarrow \qquad \mathbf{v}(N) = \frac{\alpha}{\beta - \lambda}.$$

Similarly, the recursion matrix resulting from (6.53) gets only slightly modified with respect to the sign of its off-diagonal elements.

$$R_\lambda := \begin{pmatrix} \frac{\lambda}{\alpha} - \frac{\beta^2}{\alpha\lambda} & \frac{\beta}{\lambda} \\ -\frac{\beta}{\lambda} & \frac{\alpha}{\lambda} \end{pmatrix}$$

Also for this new recursion matrix, lemma (6.20) can be proved since the crucial commutation relation is not affected by the change of sign concerning the off-diagonal elements.
Analogously to the demonstration of lemma (6.21) one shows, that the eigenvalues of R_λ are w^m and w^{-m} if $\lambda = \lambda_m^\pm$. But in contrast to the previous situation we have now

$$-R^{-1}_{\lambda_m^\pm} \begin{pmatrix} 1 \\ \frac{\alpha}{\lambda_m^\pm - \beta} \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\lambda_m^\pm - \beta} \\ 1 \end{pmatrix},$$

since

$$-\begin{pmatrix} \frac{\lambda_m^\pm}{\alpha} - \frac{\beta^2}{\alpha\lambda_m^\pm} & \frac{\beta}{\lambda_m^\pm} \\ -\frac{\beta}{\lambda_m^\pm} & \frac{\alpha}{\lambda_m^\pm} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \frac{\alpha}{\beta - \lambda_m^\pm} \end{pmatrix} = -\begin{pmatrix} \frac{\alpha}{\lambda_m^\pm} & -\frac{\beta}{\lambda_m^\pm} \\ \frac{\beta}{\lambda_m^\pm} & \frac{\lambda_m^\pm}{\alpha} - \frac{\beta^2}{\alpha\lambda_m^\pm} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\alpha}{\beta - \lambda_m^\pm} \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\beta - \lambda_m^\pm} \\ 1 \end{pmatrix}.$$

This excludes the possibility of m being even, such that $R^{N-1}_{\lambda_m^\pm} = I$. Therefore the odd m's are favored this time, implying $R^{N-1}_{\lambda_m^\pm} = -I$.

An explicit expression for the eigenvectors can be found with the ansatz (6.51) employing however the modified 'initial condition'.

$$\begin{array}{rcccl} \mathbf{v}_m^{\pm}(1) & = & 1 & = & \kappa_1 w^m + \kappa_2 w^{-m} \\ \mathbf{v}_m^{\pm}(N) & = & \frac{\alpha}{\beta-\lambda_m^{\pm}} & = & \kappa_1(-1)^m + \kappa_2(-1)^m \end{array}$$

Taking into account that $(-1)^m = -1$, as m is supposed to be odd, and noticing that the left hand side of the second equation is just the negative of the corresponding term in (6.51), we realize exactly the same system as in (6.51). Therefore the formula (6.52) does not only hold for even $m \in \{1, \ldots, N-1\}$ but for all $m \in \{1, \ldots N-1\}$.

Eigenvectors of $\mathbf{E}_R$ pertaining to $\alpha+\beta$ and $\beta-\alpha$. We have already seen that the eigenvalues $\alpha+\beta$ and $\beta-\alpha$ play a special role. In fact, the eigenvector $\mathbf{V}_0^+$ pertaining to $\alpha+\beta$ can be guessed immediately. The eigenvalue equation (in $\mathbb{C}^{2N}$) for $\lambda = \alpha+\beta$

$$\mathbf{E}_R\mathbf{V} = (\alpha+\beta)\mathbf{V} \;\Leftrightarrow\; \mathbf{L}(\alpha\mathbf{I}+\beta\mathbf{J})\mathbf{V} = (\alpha+\beta)\mathbf{V} \;\Leftrightarrow\; \alpha\mathbf{V}+\beta\mathbf{J}\mathbf{V} = (\alpha+\beta)\mathbf{R}\mathbf{V}$$

is obviously satisfied by the *constant* vector, that is mapped onto itself by $\mathbf{R}$ and $\mathbf{J}$. Hence $\mathbf{V}_0^+(k) = 1, k \in \{1, ..., 2N\}$.

The eigenvector $\mathbf{V}_0^-$ associated to $\beta-\alpha$ needs a little bit more scrutiny. The fact that the eigenvalue $\beta-\alpha$ is just the difference of the coefficients appearing on the left hand side of the eigenvalue equation

$$\mathbf{L}(\alpha\mathbf{I}+\beta\mathbf{J})\mathbf{V} = (\beta-\alpha)\mathbf{V} \quad \Leftrightarrow \quad \alpha\mathbf{V}+\beta\mathbf{J}\mathbf{V} = (\beta-\alpha)\mathbf{R}\mathbf{V}$$

suggests, that up to a constant $\mathbf{V}_0^-(k) = (-1)^k$ for $k \in \{1, \ldots, 2N\}$. So $\mathbf{V}_0^-$ is the vector, whose components are alternatingly either 1 or -1.
Let us suppose that our ansatz is correct, then a shiftmatrix applied to$\mathbf{V}_0^-$ yields just its negative. Thus we get especially $\mathbf{R}\mathbf{V}_0^- = -\mathbf{V}_0^-$. Furthermore we have $\mathbf{J}\mathbf{V}_0^- = -\mathbf{V}_0^-$, since $\mathbf{V}_0^-$ is of dimension $2N$ being always even. Plugging now $\mathbf{V}_0^-$ into the right equation form above leads to

$$\alpha\mathbf{V}_0^- + \beta\mathbf{J}\mathbf{V}_0^- = (\beta-\alpha)\mathbf{R}\mathbf{V}_0^- \quad \Leftrightarrow \quad \alpha\mathbf{V}_0^- - \beta\mathbf{V}_0^- = -(\beta-\alpha)\mathbf{V}_0^-,$$

which verifies our ansatz as being true. Notice, that $\mathbf{V}_0^-$ is either positively or negatively repetitive, depending on whether N is even or odd respectively. This varying behavior is necessary in order to ensure that the subspaces of positive and negative parity, $\mathbb{C}_{r+}^{2N}$ and $\mathbb{C}_{r-}^{2N}$, contain the same number of eigenvectors. Otherwise this would provoke a contradiction to the equal dimension of the subspaces.

6.5 Towards the D1P3 scheme & Concluding remarks

Before finishing this chapter, we want shortly discuss the possibility to generalize the results of the preceding sections.

D1P3 lattice-Boltzmann algorithm

In particular one might ask whether the considerations in section 6.2 can be transferred to the D1P3 algorithm which contains the D1P2 model as a specific limit case.
For this, let us consider the D1P3 equilibrium given in table 1.1 which leads to the advection-diffusion equation as target equation.

$$\mathcal{E}(u;s) = \mathsf{w}(s)\big(1 + h\theta a s - h^2\tfrac{1}{\omega}c\big)\,u \quad \text{with } \mathsf{w}(s) = \begin{cases} \frac{1}{2\theta} & s = -1 \\ \frac{1}{2\theta} & s = +1 \\ \frac{\theta-1}{\theta} & s = \ \ 0 \end{cases} \tag{6.54}$$

As usual, h denotes the grid spacing. Here a parameter $\theta > 1$ occurs that is absent in the D1P2 equilibrium (6.6). Like the relaxation frequency ω, this weight parameter figures in the diffusivity[47].

In order to obtain a matrix formulation of the algorithm, rest populations must be additionally taken into account. We assume that they are also indexed from left to right (for instance from 9 to 12 in figure 6.1). Adopting this kind of indexation for an arbitrary grid with N nodes, the transport and collision matrix of the D1P3 algorithm become

$$\mathbf{T} = \begin{pmatrix} \mathbf{L} & 0 & 0 \\ 0 & \mathbf{R} & 0 \\ 0 & 0 & \mathbf{I} \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \alpha\,\mathbf{I} & \beta\,\mathbf{I} & \beta\,\mathbf{I} \\ \gamma\,\mathbf{I} & \delta\,\mathbf{I} & \gamma\,\mathbf{I} \\ \eta\,\mathbf{I} & \eta\,\mathbf{I} & \zeta\,\mathbf{I} \end{pmatrix},$$

where $\mathbf{L}, \mathbf{R}$ and $\mathbf{I}$ respectively denote the leftshift, rightshift and identity matrix of dimension N. The coefficients of the collision block are given by

$$\begin{aligned} \alpha &:= 1 - \omega + \beta & \beta &:= \tfrac{\omega}{2\theta}\big(1 - \theta a h - \tfrac{1}{\omega}ch^2\big), \\ \delta &:= 1 - \omega + \gamma & \gamma &:= \tfrac{\omega}{2\theta}\big(1 + \theta a h - \tfrac{1}{\omega}ch^2\big), \\ \zeta &:= 1 - \omega + \eta & \eta &:= \tfrac{(\theta-1)\omega}{\theta}\big(1 - \tfrac{1}{\omega}ch^2\big). \end{aligned}$$

Proceeding analogously to the proof of theorem 6.2, it is shown that the characteristic polynomial of the evolution matrix $\mathbf{E} = \mathbf{T}\cdot\mathbf{C}$ can be written as the following product of 3×3 determinants

$$\chi_{\mathbf{E}}(\lambda) = \prod_{m=0}^{N-1} \det \begin{pmatrix} \alpha w^m - \lambda & \beta w^m & \beta w^m \\ \gamma \overline{w}^m & \delta \overline{w}^m - \lambda & \gamma \overline{w}^m \\ \eta & \eta & \zeta - \lambda \end{pmatrix}.$$

Hence $\chi_{\mathbf{E}}$ factorizes in a product of N cubic polynomials. Thereby, we are led to the following family of cubic polynomials parametrized by $\phi \in [0, 2\pi)$

$$\begin{aligned} \chi(\lambda,\phi) := \lambda^3 &- \big(\alpha e^{i\phi} + \delta e^{-i\phi} + \zeta\big)\lambda^2 \\ &+ \big(\alpha\delta - \beta\gamma + (\alpha\zeta - \beta\eta)e^{i\phi} + (\delta\zeta - \gamma\eta)e^{-i\phi}\big)\lambda - (1-\omega)^2(1 - ch^2). \end{aligned}$$

[47] The algorithmic parameters ω and θ are connected to the diffusivity ν by $\nu = \frac{1}{\theta}(\frac{1}{\omega} - \frac{1}{2})$.

Plots of the resulting spectral limit sets are given in figure 6.15 and 6.17.

Furthermore let us note that the eigenvalues of the collision block are

$$1-\omega \quad \text{with} \quad \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad 1-ch^2 \quad \text{with} \quad \begin{pmatrix} -1+\theta ah+\tau ch^2 \\ 1+\theta ah+\tau ch^2 \\ 2(\theta-1)(\tau ch^2-1) \end{pmatrix}$$

spanning the associated eigenspaces (observe the parallel to lemma 6.3).

The limit $\theta \to 1$. We observe in figure 6.15 that the spectral limit set for $\theta = 1$ (plot left above) equals the spectral limit set of the corresponding D1P2 evolution matrix. Indeed, the evolution of the rest population decouples from the evolution of the other populations since $\theta \to 1$ implies $\eta \to 0$. Hence, for $\theta = 1$, the evolution matrix is *blockdiagonal*: the upper block of dimension $2N$ corresponds to the evolution matrix of the D1P2 algorithm, whereas the lower $N \times N$ block is just $(1-\omega)\mathbf{I}$. This explains the above observation[48].
If the rest populations are initialized by 0, the D1P3 algorithm turns out to be numerically equivalent to the D1P2 algorithm; otherwise the rest populations are quickly damped down because[49] they are multiplied by $1-\omega$ in each iteration.
If θ is chosen to be less than 1, the rest populations are associated with negative weights. Apart from appearing quite unphysical, this setting leads to instable behavior of the algorithm.

The limit $\theta \to \infty$. In this case, one could expect that the scheme changes into a standard discretization of the diffusion equation using only a single variable, due to the following reason: As θ determines how the total mass $U = \mathsf{F}_{-1} + \mathsf{F}_{+1} + \mathsf{F}_0$ is distributed among the three populations, the increase of θ means that mass is "sucked" from the left and right moving population to be pumped into the resting population until they end up with zero mass in the limit. Thereby, F_0 finally becomes identical to the mass moment which approximates the solution of the diffusion equation.
However, this expectation is not corroborated by figure 6.15 (see plot right below for comparatively large θ). If θ tends to ∞, the green and blue branch are tied up around the circle of radius $|1-\omega|$. Simultaneously, the red branch is constricted and eventually dwindles into a single point which is 1. So, the scheme decouples again in the limit $\theta \to \infty$. But now the populations $\mathsf{F}_{-1}, \mathsf{F}_{+1}$ are rapidly attenuated because the diagonal elements α, δ converge to $1-\omega$ while the off-diagonal elements β, γ vanish. Since those populations loose significance which are responsible for the exchange of information between the nodes, it is intuitively evident that the macroscopic dynamics of the D1P3 scheme is rather stagnant in the limit $\theta \to \infty$. Thus, the scheme turns into a "boring" algorithm. There are only two kinds of

[48] The determinant (and hence the characteristic polynomial) of a blockdiagonal matrix is given as product of the determinants (characteristic polynomials) pertaining to the diagonal blocks. Observe that the lower block of the evolution matrix admits only the eigenvalue $(1-\omega)$ with N-fold multiplicity (see green point in the plot).

[49] Provided that $\omega \in (0,2)$ such that $|1-\omega| < 1$.

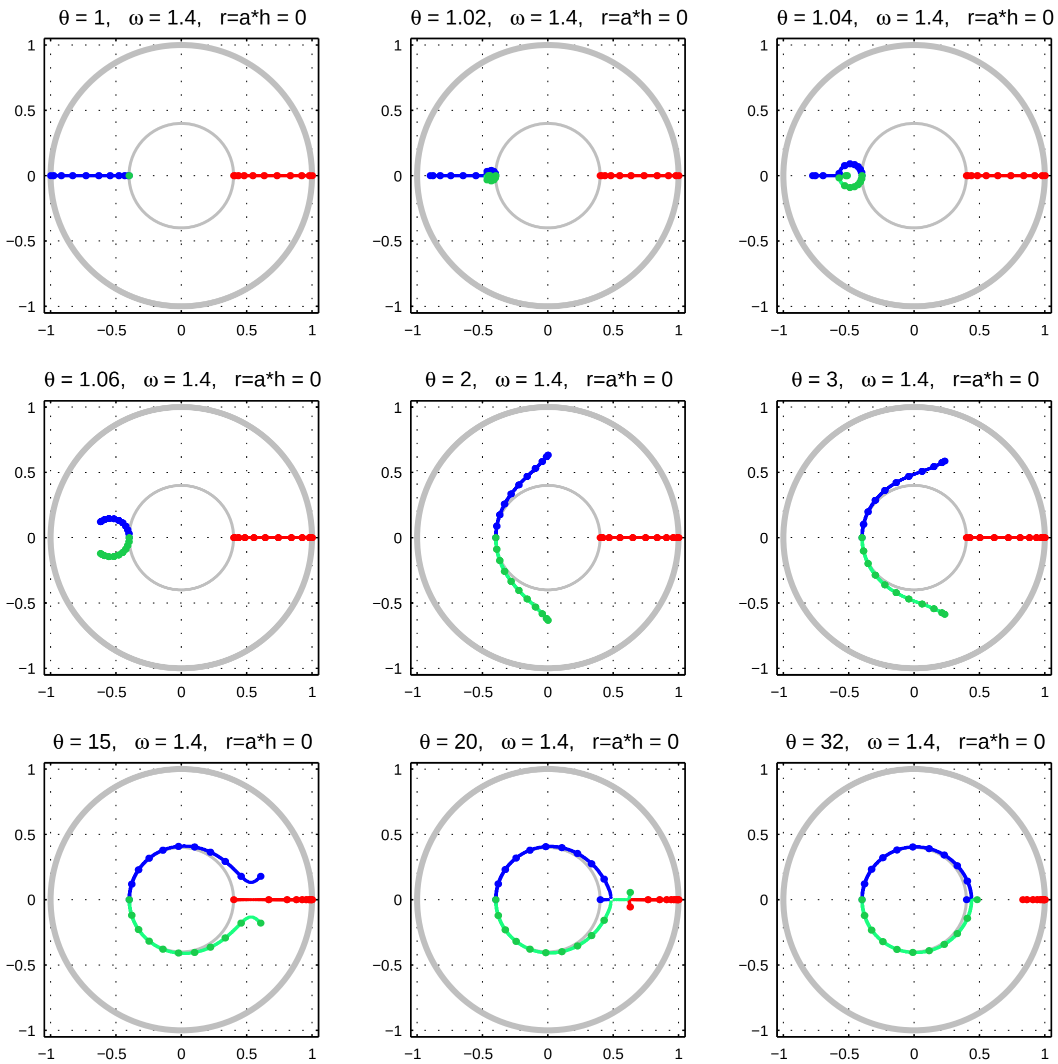


Figure 6.15: The plots illustrate how the spectral limit set deforms if the weight parameter θ is varied whereas $\omega = 1.4$ and $a = c = 0$ remain fixed. The colors come about as we have used *Cardano's formula* to solve the cubic factor polynomials the characteristic polynomial is composed of. So the colors mark the first, second and third solution obtained with Cardano's formula which seem to agree with the natural branches of the spectral limit set. For $\theta = 1$ the spectral limit set of the D1P2 algorithm is obtained (to be compared with figure 6.3). The plot next to it suggests that it might be advantageous to run the algorithm with θ slightly greater than 1. So the positive properties of the D1P2 scheme (strong interaction between the nodes) can be combined with the fact that the left branch of the spectrum stays away from the unit circle. This ensures that undesired, highly oscillating and unphysical modes being possibly excited by initialization and rounding errors are damped out.

eigenvectors: those which are quickly[50] damped to 0 because their eigenvalue lies on the circle with radius $|1-\omega|$, and those eigenvectors which are left invariant as the associated eigenvalue is 1.

Stability. As long as ω and θ are suitably selected, h can be chosen sufficiently small such that all six coefficients of the collision block are positive. Then stability with respect to the ℓ^∞-norm can be shown following the steps in the proof of theorem 6.1. However, employing this approach, we cannot handle the full range of $\omega \in [0,2]$ and $\theta \in (1,\infty)$ where a stable behavior is expected by numerical tests. This problem already arises in the purely diffusive case. To become more precise, let us consider the following collision block ($a=0, c=0$):

$$\begin{pmatrix} 1-\omega+\frac{1}{2\theta}\omega & \frac{1}{2\theta}\omega & \frac{1}{2\theta}\omega \\ \frac{1}{2\theta}\omega & 1-\omega+\frac{1}{2\theta}\omega & \frac{1}{2\theta}\omega \\ \frac{\theta-1}{\theta}\omega & \frac{\theta-1}{\theta}\omega & 1-\omega+\frac{\theta-1}{\theta}\omega \end{pmatrix}$$

- Obviously, the sum of the three coefficients in each column always yields 1. Therefore stability with respect to the ℓ^1-norm is immediately obtained if all coefficients are non-negative.
- Under the assumption of non-negative coefficients, even ℓ^∞-stability can be verified by adapting the proof of theorem 6.1.
- Negative diagonal coefficients of the collision block can only appear for $\omega > 1$. There are two cases to be distinguished:

  $$\begin{array}{llll} \frac{\theta-1}{\theta} < \frac{1}{2\theta} & \text{if } 1<\theta<\frac{3}{2} & \text{hence} & 1-\omega+\frac{1}{2\theta}\omega<0 \;\Rightarrow 1-\omega+\frac{\theta-1}{\theta}\omega<0, \\ \frac{\theta-1}{\theta} \geq \frac{1}{2\theta} & \text{if } \frac{3}{2}\leq\theta & \text{hence} & 1-\omega+\frac{\theta-1}{\theta}\omega<0 \;\Rightarrow 1-\omega+\frac{1}{2\theta}\omega<0. \end{array}$$

  What relation must hold between ω and θ such that at least one of the coefficients in the collision block gets negative? In the first case $\theta < \frac{3}{2}$ we find

  $$1-\omega+\frac{\theta-1}{\theta}\omega<0 \quad\Leftrightarrow\quad \theta<\omega,$$

  while the second case $\theta \geq \frac{3}{2}$ leads to the condition

  $$1-\omega+\frac{1}{2\theta}\omega<0 \quad\Leftrightarrow\quad \theta>\frac{1}{2}\frac{\omega}{\omega-1}.$$

  So, the above mentioned stability arguments fail if (ω,θ) belongs to the subarea of the $\omega\theta$-plane, which is marked in figure 6.16 (left) by the darker (pink) hachures. A possibility to rigorously demonstrate stability (with respect to the ℓ^2-norm) independently of the sign that is taken on by the coefficients of the collision block is to mimic the proof of theorem 6.6 in section 6.3. However, for this it must be shown that the spectral limit set lies completely inside the closed unit disk.

[50] "Quickly" means here in the *discrete time scale*, that only depends on the number of iterations but not on the grid spacing h. Therefore no macroscopic evolution can occur which would be independent of the underlying grid.

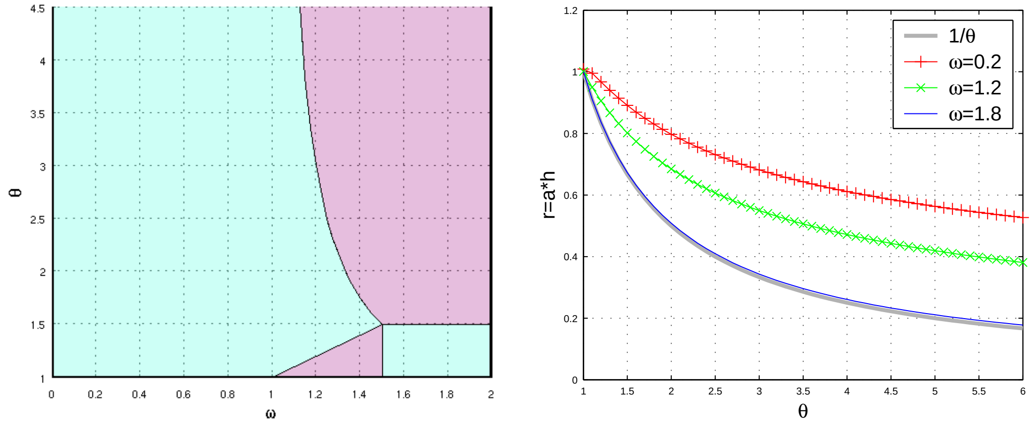


Figure 6.16: Left: From the eigenvalues of the collision block we conclude that $\omega \in [0,2]$ is necessary for stability. To avoid a negative weight for the rest population, $\theta \geq 1$ is additionally required. The diagram depicts for which choice of (ω, θ) stability is proved so far (greenish/bright subregion) and where it is only expected (reddish/darker subregion). Notice that the curved boundary between the two zones corresponds to the graph $(\omega, \theta = \frac{1}{2}\frac{\omega}{\omega-1})$. Right: The curves visualize the functional dependence between θ and the specific CFL-number, where the spectral limit set starts to cross the unit circle if the CFL-number is increased further. For each curve, ω is kept constant. The plot reveals a clear trend: the larger θ and ω are chosen, the more blatantly the CFL condition is violated, i.e. the smaller the maximal CFL-number can be chosen where instabilities do not yet occur. Furthermore it is recognized that $r(\theta) \to \frac{1}{\theta}$ as $\omega \to 2$. Only in the case of the D1P2 scheme $(\theta = 1)$ the CFL-condition is respected.

The CFL-condition. In section 6.3 we have considered the D1P2 algorithm in hyperbolic scaling, which gave rise to the pure advection equation as target problem. Here the issue of the CFL-condition[51] came up. It was proved that the inequalities

$$\begin{aligned} |a| &\leq 1 \quad \text{for hyperbolic scaling (grid speed } \tfrac{\Delta x}{\Delta t} = 1) \\ |ah| &\leq 1 \quad \text{for parabolic scaling (grid speed } \tfrac{\Delta x}{\Delta t} = \tfrac{1}{h}) \end{aligned}$$

together with the requirement $\omega \in [0,2]$ assure as necessary and sufficient condition, that the spectral limit set is situated inside the closed unit disk. Now the question raises, whether this condition can be generalized to the D1P3 algorithm. The plots in figure 6.17 give clear evidence that the condition is not sufficient to guarantee that the spectral limit set stays inside the closed unit disk.
However, in view of figure 6.16 it is conspicuous that stability seems to be provided whenever the CFL-number $r = ah$ is less than $1/\theta$. Moreover, this bound may be sharp and should be exact for $\omega = 2$. From a purely algebraic perspective this means that the advective term $h\theta as$ of the equilibrium 6.54 must be bounded by 1. How might this relation be interpreted?

[51] The advection speed must not be larger than the grid speed; in other words the ratio *advection speed / grid speed* must be less or at most equal than 1.

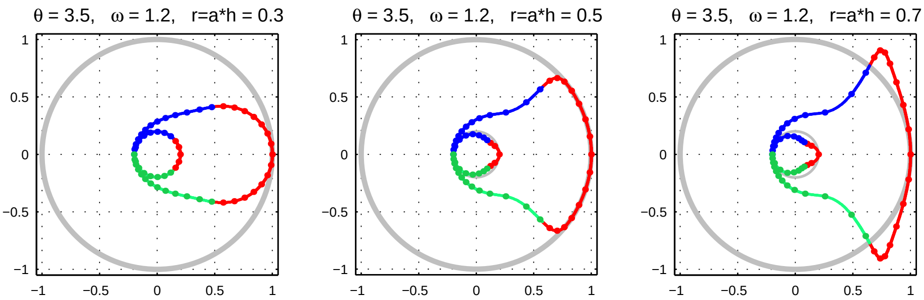


Figure 6.17: The spectral limit set for fixed $\omega = 1.2$ and $\theta = 3.5$ but variable CFL-number $r = ah$ (for the parabolic case). It is clearly seen that the spectrum lies partly outside the closed unit disk although $r = ah = 0.7$. (Annotation: The coloring is again due to the application of Cardano's formula – however it seems here less natural than in figure 6.16.)

At present, trying to answer this question, we reach an area of speculation. Besides the fact that θ enters the formula for the diffusivity ν, we remember having found another physical meaning for this parameter in section 2.1 on page 74. There it appeared as squared wave velocity (squared sound speed) in combination with a two moment conserving equilibrium. Possibly, there is some connection behind which escapes for the moment. It may be, that a multiscale expansion could help to clear up this issue by tracking down further dynamics hidden in the D1P3 lattice-Boltzmann algorithm.

To sum up this exposition, it became clear that the rest population can be employed to control the stability of the D1P2/D1P3 lattice-Boltzmann algorithm. Therefore it should be worthwile to investigate further, practically more relevant lattice-Boltzmann methods (D2P4/D2P5 schemes for diffusion and above all D2P8/D2P9 schemes for fluid flows) with respect to this point.

Bounce-back boundary conditions and advective equilibrium

Throughout our discussion concerning the stability of boundary conditions, the advective term of the equilibrium (6.6) has been disregarded. Therefore we briefly touch this topic here once again. It was mentioned in section 6.4 that the spectrum pertaining to the evolution matrix of the lattice-Boltzmann algorithm imitates certain properties of the spectrum which belongs to the evolution operators of the target equation. Figure 6.18 and 6.19 illustrate how the spectrum reacts if the advection parameter a gets different from 0.
Let us close the chapter with a stability statement:

Proposition 6.12. *The D1P2 advective-diffusive lattice-Boltzmann algorithm with (sign inverting) bounce-back boundary conditions defined in* (6.38) *and* (6.40) *is stable with respect to the* ℓ^1*-norm if* $0 < \omega < 2$ *(independently of* $a \in \mathbb{R}$ *and* $c \geq 0$*).*

Proof: Considering the coefficients of the collision block in (6.8), we see that α, β, γ and δ are positive for $\omega \in (0, 2)$ if h is chosen sufficiently small. Furthermore

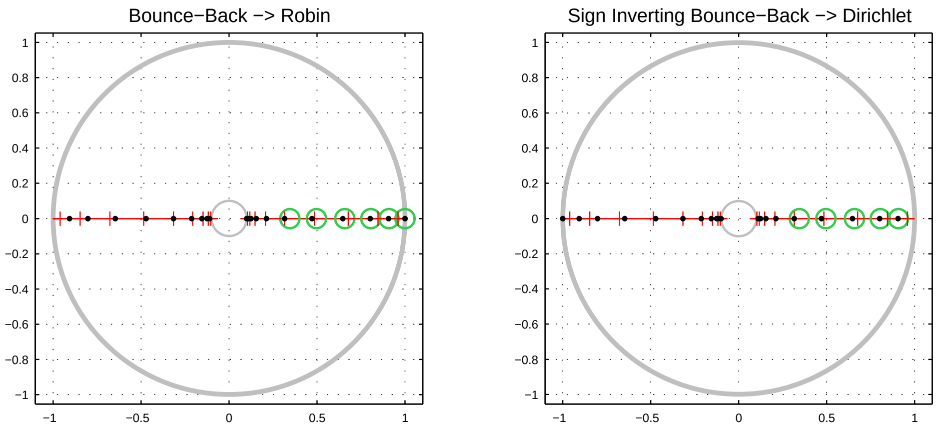


Figure 6.18: Overrelaxed case $\omega = 1.1$: The spectrum for the purely diffusive case is marked by vertical lines (red). If the advection is switched on (here $r = ah = 0.3$) the eigenvalues (dots) are shifted along the real line. In contrast to the case of periodic boundary conditions, the eigenvalues remain real. This property parallels the corresponding behavior of the target equation (compare the eigenvalues in theorem A.2 and theorem A.1). The green circles represent some eigenvalues pertaining to the evolution operator of the target problem. Especially near 1, they surround in their center a dot, which indicates how well they are approximated by the eigenvalues of the lattice-Boltzmann evolution operator. It should be observed that the spectra are not completely symmetric with respect to the imaginary axis (although the spectral limit set is). Furthermore we note that 1 belongs to the left but not to the right spectrum. This reflects the observation that homogeneous Robin boundary conditions admit a non-zero conserved state which does not exist in the case of homogeneous Dirichlet boundary conditions.

$\alpha+\gamma = 1$ and $\beta+\delta = 1$ which shows that the ℓ^1-norm of the collision matrix $\mathbf{C}$ is 1. Since the ℓ^1-norm of the transport matrices $\mathbf{T}_R$ and $\mathbf{T}_D$ is equal to 1 as well, we infer $\|\mathbf{E}_{R,D}\|_1 = \|\mathbf{T}_{R,D}\mathbf{C}\|_1 \leq \|\mathbf{T}_{R,D}\|_1 \|\mathbf{C}\|_1 = 1$ and hence $\|\mathbf{E}^n_{R,D}\|_1 \leq \|\mathbf{E}_{R,D}\|_1^n = 1$. ■

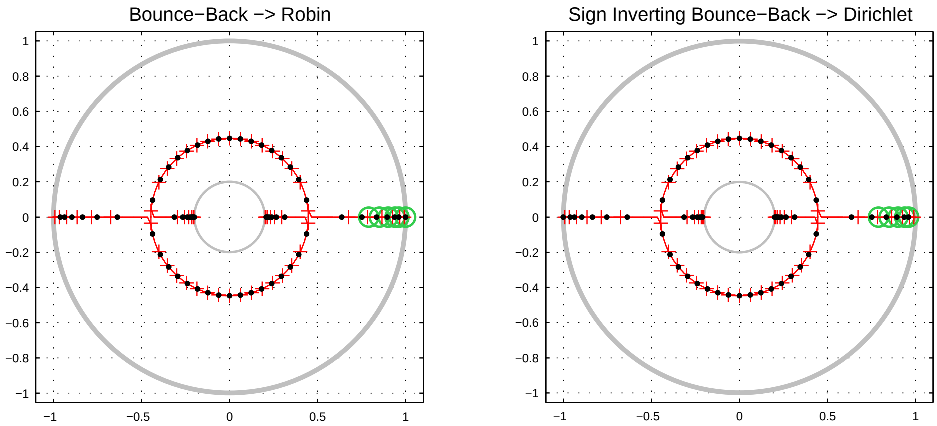


Figure 6.19: In analogy to figure 6.18, here the underrelaxed case for $\omega = 0.8$ is depicted. The spectra refer to the same CFL-number $r = 0.3$ but the advection velocity a is larger as the number of grid points has been increased to obtain more eigenvalues. Also here the eigenvalues of physical importance near 1 remain real-valued. Even the complex eigenvalues stay on the circle of radius $\sqrt{1-\omega}$ as far as r does not become too large. Therefore the spectral limit set does not alter if r may be varied over a certain range. Even though the algorithms referring to periodic or bounce-back type boundary conditions differ just in the treatment of two populations, we recognize a fundamental difference concerning the spectral properties of the evolution matrices.

Chapter 7

Asymptotic analysis of a numeric boundary layer

Initial layers are just a specific realization of a much broader class of phenomena known under the name of *boundary layers*[1]. These typically arise in connection with singular limits or numerical schemes, if the approximating problem admits additional degrees of freedom that may be chosen in non-conformity to the limit problem.
Lattice-Boltzmann methods are not only prone to initial layers as already seen in the previous chapters but they are also susceptible for spatial boundary layers. For instance, the analysis of the standard D2P9 algorithm by means of regular expansions shows that the bounce-back boundary condition implies homogeneous Dirichlet conditions in the leading zeroth order of the momentum moment. However, contradictory boundary conditions are already found for the following order. So the regular expansion breaks down in this case. To obtain an adequate description of the algorithm, the regular expansion must be complemented by further, irregular terms representing a boundary layer.

In order to become more acquainted with spatial boundary layers, we study this phenomenon in the context of a numerical scheme that is somewhat more simple than the lattice-Boltzmann model algorithm, as it involves only a single, time-independent variable. Concretely, we consider the five-point stencil discretization for the one-dimensional Poisson equation over a bounded interval. In contrast to the standard three-point stencil discretization, the five-point stencil cannot be combined with Dirichlet boundary conditions in a trivial, natural way. In order to apply the stencil in each of the neighbors associated with the two boundary nodes, several possibilities are imaginable, how to extend the solution beyond the boundary node using different orders of polynomial extrapolation. Alternatively, one-sided or restricted stencils could be used to define the scheme in both of the exceptional nodes – but this possibility is not to be considered here.

A priori, there are plausible arguments backing lower as well as higher extrapolation orders:

- One might argue in favor of a low extrapolation order: In spite of providing

[1]Interpret initial layers as boundary layers at the temporal boundary of the underlying time space domain.

rather bad stencils for the two exceptional nodes, the resulting error should become neglegible as the discretization is fourth order accurate in the vast majority of grid nodes. So even the nearest neighbor extrapolation seems to be sufficient.

- On the other hand, the standard argument (compare with proposition 1.9) to infer convergence from consistency and stability suggests that already a large residual in a single node may spoil the convergence order with respect to the maximum norm. Hence, the extrapolation order should be chosen such that the residual is at least of the same order in the exceptional as in the the bulk nodes. Thus, a higher extrapolation order should be preferred.

- Finally, it seems very reasonable to choose the extrapolation order equal to the consistency order of the regular five-point stencil in the bulk nodes.

It turns out that the third argument is right. So cubic interpolation, which generally is fourth order accurate, corresponds to the lowest order of extrapolation leading to the maximal convergence order of the scheme. Although being quite plausible, this result is remarkable from the viewpoint of the second argument. Actually, it turns out that the associated classical consistency order is only 2 at the exceptional nodes, which is less than the experimentally tested fourth order convergence. This shows that there is also some truth in the first argument. Even the nearest neighbor interpolation, which leads to a negative consistency order, results in a convergent scheme. Obviously, the scheme "tolerates" less accurate stencils at the exceptional nodes near the boundary.

Of course, the proof of these observations cannot be based on proposition 1.9 by just assuming stability (uniform boundedness of the inverse finite difference operator). In fact, another more subtle property of the finite difference operator and its inverse comes into play, that we denote as *damping property* (see for example chapter 4 of [57] and references therein, where similar phenomena are discussed).

The goal of the following analysis is to exemplify the following aspects:

- asymptotic expansions may contain discrete order functions beside smooth order functions.

- formal asymptotic analysis may not be sufficient to predict the convergence order even if stability is provided.

- what kind of information about the inverse finite difference operator might be necessary to prove actual convergence rates.

Section 7.1, 7.2 and 7.3 are mainly of preliminary character. Section 7.1 presents some prerequisites whereas section 7.2 and 7.3 introduce the model problem and the specific discretizations. While stability is proved in section 7.4, the decisive damping property is discussed in section 7.5, which is used to establish an asymptotic expansion of the numerical error in section 7.6. Numerical tests are provided in the last section.

7.1 Some remarks about interpolation and difference stencils

This section assembles in its first part some facts about interpolation – including extrapolation as a special case. The second part is devoted to difference stencils. Some material should be found in standard textbooks, while others might be more original and especially adapted to our needs. Although the intention was to develop the material only as far as it used in later sections, it might be also of interest for its own. Moreover, we indicate also some points that might spur to further investigations.

The problem of interpolation consists in extending a function f, that is only given on a finite set of *sampling nodes/points* $x_0, \ldots, x_n$. The extension is performed by certain *ansatz functions*, where the most popular ones are polynomials. The interpolating polynomial is defined in such a way, that it attains the provided values of the function f to be interpolated in the sampling nodes. This amounts to as many equations as sampling nodes, viz. $n+1$. The equations can be used to determine the $n+1$ coefficients of a polynomial of n'th order.

$$\begin{array}{l} a_n x_0^n + \ldots + a_0 \overset{!}{=} f(x_0) \\ a_n x_1^n + \ldots + a_0 \overset{!}{=} f(x_1) \\ \vdots \qquad\qquad \vdots \quad \vdots \\ a_n x_n^n + \ldots + a_0 \overset{!}{=} f(x_n) \end{array} \quad \Leftrightarrow \quad \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

The system matrix is known as the *Vandermonde* matrix $V = V(x_0, \ldots, x_n)$. In the case of pairwise distinct sampling nodes the Vandermonde matrix is invertible[2]; hence the interpolation polynomial $P(f|x_0, \ldots, x_n)$ is uniquely defined.

Often one is only interested in the interpolation value at some special position, called *interpolation point*, which we denote in the sequel by y. Observe that y is, strictly speaking, an extrapolation point, if $y \notin [\min\{x_0, \ldots, x_n\}, \max\{x_0, \ldots, x_n\}]$. As the coefficients of the interpolation polynomial depend linearly on $f(x_0), \ldots, f(x_n)$, the value $P(f|x_0, \ldots, x_n)(y)$ should also depend linearly on them via the *interpolation weights* $w_0, \ldots, w_n$:

$$P(f|x_0, \ldots, x_n)(y) = w_0 f(x_0) + \ldots + w_n f(x_n) \; .$$

Noticing that f is exactly recovered by the interpolation if it is a polynomial of order less or equal than n, we can set up the following system for the interpolation weights,

$$\begin{array}{l} w_0 x_0^0 + \ldots + w_n x_n^0 \overset{!}{=} y^0 \\ w_0 x_0^1 + \ldots + w_n x_n^1 \overset{!}{=} y^1 \\ \vdots \qquad\qquad \vdots \quad \vdots \\ w_0 x_0^n + \ldots + w_n x_n^1 \overset{!}{=} y^n \end{array} \quad \Leftrightarrow \quad \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \cdots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} 1 \\ y \\ \vdots \\ y^n \end{pmatrix}$$

[2] Its determinant is given by $\det V(x_0, \ldots, x_n) = \prod_{0 \le i < j \le n} (x_j - x_i)$, which is different from zero, if $x_0, \ldots, x_n$ are pairwise different.

where the transposed Vandermonde matrix is encountered.
The interpolation polynomial and the interpolation weights can be computed alternatively by means of the *Lagrange* polynomials defined for a set of sampling nodes by

$$L_i(t) := \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-x_j}{x_i-x_j}, \qquad \text{for } i \in \{0,...,n\}.$$

They are characterized by the property $L_i(x_j) = \delta_{ij}$. Therefore, the interpolating polynomial becomes

$$P(f|x_0,...,x_n)(t) = \sum_{i=1}^{n} L_i(t) f(x_i),$$

For $t = y$ we discover that the interpolation weights of y with respect to $x_0,...,x_n$ are simply given by $w_i = L_i(y)$.

In numerical schemes for solving (partial) differential equations, interpolation is frequently used to deal with irregular nodes that do not possess the full number of neighbor nodes (for instance at boundaries or interfaces). This gives rise to the following question: Do the interpolation weights depend on the very special position of the sampling nodes or are they already completely determined by the configuration (ratios of distances/ angles)? The subsequent theorem yields an affirmative answer for the latter possibility. In fact, the interpolation weights are equal, if they refer to geometrically similar configurations. Therefore, the interpolation weights have to be computed only for certain typical configurations but not for every irregular node. Furthermore, they are invariant with respect to the grid scaling (grid size).

Theorem 7.1. *Let $x_0,...,x_n \in \mathbb{R}$ and $\xi_0,...,\xi_n \in \mathbb{R}$ be two sets of sampling nodes and $x, \xi \in \mathbb{R}$ two interpolation points. Moreover $w_0,...,w_n$ and $\omega_0,...,\omega_n$ denote the associated interpolation weights of x and ξ. Suppose there exists $a \in \mathbb{R}\setminus\{0\}$ and $b \in \mathbb{R}$, such that*

$$\xi = ax + b \quad \textit{and} \quad \xi_i = ax_i + b \quad \textit{for}\, i \in \{0,...,n\}.$$

Then x and ξ have the same interpolation weights, i.e. $w_i = \omega_i$ for $i \in \{0,...,n\}$.

The statement of the theorem can also be formulated in terms of the Vandermonde matrix with $\mathbf{r}(x) := (1, x, ..., x^n)^T$ and $\mathbf{w} := (w_0, w_1, ..., w_n)^T$.

$$V(x_0,...,x_n)^T \cdot \mathbf{w} = \mathbf{r}(x) \quad \Rightarrow \quad V(ax_0+b,...,ax_n+b)^T \cdot \mathbf{w} = \mathbf{r}(ax+b)$$

Proof: Let L_i and Λ_i denote the corresponding i'th Lagrange polynomial. Using the assumption of the theorem we obtain

$$\begin{aligned}\omega_i &= \Lambda_i(\xi) = \prod_{\substack{j=0\\j\neq i}}^{n} \frac{\xi-\xi_j}{\xi_i-\xi_j} \\ &= \prod_{\substack{j=0\\j\neq i}}^{n} \frac{(ax+b)-(ax_j+b)}{(ax_i+b)-(ax_j+b)} \\ &= \prod_{\substack{j=0\\j\neq i}}^{n} \frac{x-x_j}{x_i-x_j} \;=\; L_i(x) \;=\; w_i \;,\end{aligned}$$

which yields the assertion. ■

An analog result can be derived for the interpolation of functions depending on several variables. The next proposition is of practical interest for finite-difference methods based on uniform grids.

Proposition 7.1. *Let $x_0 < ... < x_n$ be equidistant, that means $|x_{i-1} - x_i| = h$ for $i \in \{1, ..., n\}$ with $h > 0$ fixed. The interpolation weights of x_k, $k \in \{0, ..., n\}$, with respect to $\{x_0, ..., x_n\} \setminus \{x_k\}$ are given by*

$$w_{k,i} = (-1)^{i+k+1} \frac{\binom{n}{i}}{\binom{n}{k}} \;. \tag{7.1}$$

According to the definition of the interpolation weights we obtain

$$P(f|x_0, ..., \widehat{x_k}, ..., x_n)(x_k) = \sum_{\substack{i=0\\i\neq k}}^{n} w_{k,i} f(x_i) \;. \tag{7.2}$$

Note that $\widehat{\ }$ indicates, that x_k has to be deleted[3] from the list of sampling nodes. One can easily see, that the sum of the $w_{k,i}$'s over i is equal to one, what reflects the fact, that constant functions are exactly interpolated. Expanding $0 = (1-1)^n$ by the binomial formula, solving for $\binom{n}{k}$ and dividing by it yields the desired equation.

Proof: By theorem 7.1 it is sufficient to consider the special case of $x_i = i$ for $i \in \{0, 1, ..., n\}$ such that $h = 1$. As we have n sampling points $\{x_0, ..., x_n\} \setminus \{x_k\}$ the interpolation will be of order n. Thus, if f in (7.2) is a polynomial with $\deg f \leq n-1$, f will agree with its interpolation polynomial, i.e. $f = P(f)$. Taking in particular the first n monomials $1, x, ..., x^{n-1}$, the following n equalities are required to hold:

$$k^\ell \stackrel{!}{=} \sum_{\substack{i=0\\i\neq k}} w_{k,i}\, i^\ell \qquad \text{for } 0 \leq \ell \leq n-1. \tag{7.3}$$

This provides a linear system that could be solved for the weights.

[3] Figuratively, x_k has to 'take its hat' – i.e. to leave – as it is formulated in the German edition of [12], volume 1.

Since the system matrix is the transposed *Vandermonde* matrix $V^T(0, ..., \widehat{k}, ..., n)$ with $V^T_{\ell i} = i^\ell$ the linear system has a unique solution. Hence we are done, as soon as we show, that (7.1) is a solution.
If we insert (7.1) into (7.3) and multiply by $\binom{n}{k}$ we get:

$$\begin{aligned} \binom{n}{k} k^\ell &\stackrel{!}{=} (-1)^{k+1} \sum_{\substack{i=0 \\ i \neq k}}^{n} \binom{n}{i} (-1)^i i^\ell \\ \Leftrightarrow \qquad 0 &\stackrel{!}{=} (-1)^{k+1} \sum_{\substack{i=0 \\ i \neq k}}^{n} \binom{n}{i} (-1)^i i^\ell \; - \; \binom{n}{k} k^\ell \end{aligned}$$

Writing the minus sign in front of the binomial coefficient as $(-1)^{2k+1}$ yields,

$$\Leftrightarrow \qquad 0 \stackrel{!}{=} (-1)^{k+1} \sum_{\substack{i=0 \\ i \neq k}}^{n} \binom{n}{i} (-1)^i i^\ell \; + \; (-1)^{2k+1} \binom{n}{k} k^\ell$$

such that the term containing k can be incorporated into the sum:

$$\Leftrightarrow \qquad 0 \stackrel{!}{=} \sum_{i=0}^{n} \binom{n}{i} (-1)^i i^\ell \; = \; \ell!(-1)^n P^{(n)}_\ell(1)$$

The polynomial $P^{(n)}_\ell$ is defined in (7.4). Since $P^{(n)}_\ell(1) = 0$ for $\ell < n$, the assertion is a direct consequence of lemma 7.1. ■

In order to evaluate certain sums (involving binomial coefficients) that we encounter in the proofs of proposition 7.1 and 7.2, it is advantageous to introduce for $n \in \mathbb{N}_0$ and $\ell \in \mathbb{N}_0$ the following family of polynomials:

$$P^{(n)}_\ell(x) := \tfrac{1}{\ell!} (x\partial)^\ell (x-1)^n \,. \tag{7.4}$$

Notice that the degree of $P^{(n)}_\ell$ is n, as any differentiation is succeeded by a multiplication by x.

Lemma 7.1. *The subsequent equality holds true*

$$P^{(n)}_\ell(1) = \tfrac{1}{\ell!} \sum_{k=0}^{n} (-1)^{n+k} \binom{n}{k} k^\ell \,, \quad \textit{in particular} \quad P^{(n)}_\ell(1) = \begin{cases} 0: & 0 \leq \ell < n \\ 1: & \ell = n \end{cases} .$$

Proof: Expanding $(x-1)^n$ by means of the binomial formula yields

$$\begin{aligned} P^{(n)}_\ell(x) = \tfrac{1}{\ell!} (x\partial)^\ell (x-1)^n \; &= \tfrac{1}{\ell!} (x\partial)^\ell \sum_{k=0}^{n} \binom{n}{k} x^k (-1)^{n-k} \\ = \tfrac{1}{\ell!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} (x\partial)^\ell x^k \; &= \tfrac{1}{\ell!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} k^\ell x^k \,. \end{aligned}$$

The assertion is obtained by setting $x = 1$ and observing that $(-1)^{n-k} = (-1)^{n+k}$.
By induction one can show, that the highest derivative occurring in the differential operator $(x\partial_x)^\ell$ is of order ℓ. More exactly it can be expanded into

$$(x\partial)^\ell = x^\ell \partial^\ell + \tfrac{\ell(\ell-1)}{2} x^{\ell-1} \partial^{\ell-1} + ... + x\partial \,.$$

Furthermore we have

$$\partial^\ell (x-1)^n = n(n-1)\cdot \ldots \cdot (n-\ell+1)(x-1)^{n-\ell} = \ell!\binom{n}{\ell}(x-1)^{n-\ell}$$

As $(x-1)^n$ has an n-fold zero for $x = 1$, we obtain $\partial^\ell (x-1)^n = 0$ for $0 \le \ell < n$ and $P_n^{(n)} = \frac{1}{n!}n!\binom{n}{n}(x-1)^0 = 1$. ■

Let us now turn to the numeric computation of derivatives (numeric differentiation), that is classically done by finite differences (difference-quotients). In this context the stencil notation is popular.

Convention: A *difference stencil* is a finite, ordered collection of real numbers $a_1, \ldots, a_n$ written in the form $[a_1, \ldots, \overset{*}{a_k}, \ldots, a_n]$. The number $n \in \mathbb{N}$ is the *length* of the stencil; the asterisk $*$ indicates its *center*. This becomes important, if the stencil is applied as a (linear) operator.

1) $[a_1, \ldots, \overset{*}{a_k}, \ldots, a_n] : \mathcal{F}(\mathbb{Z}) \to \mathcal{F}(\mathbb{Z})$

$$\Big([a_1, \ldots, \overset{*}{a_k}, \ldots, a_n]f\Big)(j) \equiv [a_1, \ldots, \overset{*}{a_k}, \ldots, a_n]f(j) := \sum_{i=1}^{n} a_i f(j+i-k) \tag{7.5}$$

2) $[a_1, \ldots, \overset{*}{a_k}, \ldots, a_n]_h : \mathcal{F}(\mathbb{R}) \to \mathcal{F}(\mathbb{R})$

$$\Big([a_1, \ldots, \overset{*}{a_k}, \ldots, a_n]_h f\Big)(x) \equiv [a_1, \ldots, \overset{*}{a_k}, \ldots, a_n]_h f(x) := \sum_{i=1}^{n} a_i f\big(x+(i-k)h\big) \tag{7.6}$$

In the latter case an additional parameter $h > 0$, the *step-size*, must be specified. For ease of notation stencils are here defined as operators on $\mathcal{F}(\mathbb{Z})$ or $\mathcal{F}(\mathbb{R})$ respectively in order to be applicable everywhere and thus to avoid incompatibilities with possible boundaries.

Theorem 7.2. *For $n \in \mathbb{N}_0$ and $m \in \mathbb{N}$ there is a unique* right-sided (forward) *stencil $R^{n,m}$ of length $n+m$, such that for all $u \in \mathcal{C}^{n+m}(\mathbb{R})$*

$$h^{-n}R^{n,m}u = \partial^n u + \mathrm{O}(h^m)\ .$$

Similar statements hold of course for stencils, that are left-sided or whose center is located at another position.

Proposition 7.2. *The right-sided stencils of first order accuracy are given by*

$$R^{n,1} = \Big[\overbrace{(-1)^n\tbinom{n}{0}}^{*},\ (-1)^{n+1}\tbinom{n}{1},\ \ldots\ ,(-1)^{2n}\tbinom{n}{n}\Big]\ .$$

Proof: It is to show that $h^{-n}R^{n,1}u(x) \overset{!}{=} \partial^n u(x) + \mathrm{O}(h)$ for any $u \in \mathcal{C}^{n+1}(\mathbb{R})$. According to equation 7.6 we have:

$$h^{-n}R^{n,1}u(x) = h^{-n}\sum_{k=0}^{n}(-1)^{n+k}\tbinom{n}{k}u(x+kh)$$

Substituting for $u(x+kh)$ its truncated Taylor expansion and

$$h^{-n}R^{n,1}u(x) = h^{-n}\sum_{k=0}^{n}(-1)^{n+k}\tbinom{n}{k}\Big\{\sum_{\ell=0}^{n}\tfrac{1}{\ell!}k^\ell h^\ell \partial^\ell u(x) \;+\; \mathrm{O}(h^{n+1})\Big\}$$

interchanging the order of the sums,

$$h^{-n}R^{n,1}u(x) = \sum_{\ell=0}^{n}\tfrac{1}{\ell!}\Big(\sum_{k=0}^{n}(-1)^{n+k}\tbinom{n}{k}k^\ell\Big)\partial^\ell u(x)h^{\ell-n} \;+\; \mathrm{O}(h)$$

yields finally

$$h^{-n}R^{n,1}u(x) = \sum_{\ell=0}^{n}P_\ell^{(n)}(1)\,\partial^\ell u(x)h^{\ell-n} \;+\; \mathrm{O}(h) \;= \partial^n u(x) + \mathrm{O}(h)\ ,$$

where we have employed lemma 7.1. ■

Remark 7.1. The right-sided stencils for higher derivatives can be derived by an heuristic reasoning. As higher differentiation operators are simply the operator product (composition) of first order differentiation operators, the stencils should be obtained similarly by concatenation or composition. For instance:

$$R^{1,1}\circ R^{1,1} = -[\overset{*}{-1},\, 1]\circ[\overset{*}{-1},\, 1] = [\overset{*}{-1},\, 1] + [\overset{*}{0},\, -1,\, 1] = [\overset{*}{1},\, -2,\, 1] = R^{2,1}.$$

$$R^{1,1}\circ R^{1,1}\circ R^{1,1} = [\overset{*}{-1},\, 1]\circ[\overset{*}{1},\, -2,\, 1] = -[\overset{*}{1},\, -2,\, 1] + [\overset{*}{0},\, 1,\, -2,\, 1] = [\overset{*}{-1},\, 3,\, -3,\, 1] = R^{3,1}$$

Notice that attention has to be paid to the centers of the stencils when adding or composing them. If we set $R^{n,1} = [\overset{*}{r}{}_0^{(n)}, r_1^{(n)}, ..., r_n^{(n)}]$, the concatenation $R^{n+1,1} = R^{1,1}\circ R^{n,1}$ results in the following recursive rule

$$r_k^{(n+1)} = r_{k-1}^{(n)} - r_k^{(n)} \quad \text{for } 1 \le k \le n$$

to compute the stencil coefficients. This relation is complemented by setting $r_0^{(n+1)} = (-1)^{n+1}$ and $r_{n+1}^{(n+1)} = 1$. Moreover the recursion is started with $r_0^{(1)} = -1$ and $= r_1^{(1)} = 1$ since $R^{1,1} = [-1\,,1]$. The coefficients of the stencils can be arranged in a triangular pattern similar to Pascal's triangle. Unlike the recurrence relation associated to Pascal's triangle, which is a sum and not a difference, we obtain also binomial coefficients for the $r_k^{(n)}$'s but with alternating signs (cf. proposition 7.2).

Instead of searching for stencils computing arbitrary derivatives with first order accuracy, one can also ask for stencils approximating a given derivative with arbitrary accuracy. In fact, there is a nice formula pertaining to the first derivative.

Conjecture 7.1. *The right-sided stencil of minimal length, approximating the first derivative with the accuracy of order n, is given by*

$$R^{1,n} \;=\; -\sum_{k=1}^{n}(-1)^k\tfrac{1}{k}R^{k,1} \;=\; R^{1,1} - \tfrac{1}{2}R^{2,1} + \tfrac{1}{3}R^{3,1} - \tfrac{1}{4}R^{4,1} + ... + (-1)^{n+1}\tfrac{1}{n}R^{n,1}\,. \tag{7.7}$$

Remark 7.2. Similar expressions can be found for stencils associated to higher derivatives, but the occurring coefficients do not obey such an easy rule. The first terms of the Taylor expansion of $R^{n,1}$ applied to a function u are

$$h^{-n}R_h^{n,1}u = \partial^n u + hP_{n+1}^{(n)}(1)\partial^{n+1}u + h^2P_{n+2}^{(n)}(1)\partial^{n+2}u + ...\ .$$

In order to eliminate the leading error, which is essentially given by the $(n+1)$'th derivative, we can use the stencil $R^{n+1,1}$. Hence

$$R^{n,2} = R^{n,1} - P_{n+1}^{(n)}(1)\cdot R^{n+1,1}\ .$$

Note that the right hand side represents a stencil of the length $n+2$, what agrees with the length of the minimal stencil for the n'th derivative with second order accuracy. Taylor expansion reveals once more the leading error.

$$h^{-n} R_h^{n,2} u = \partial^n u + h^2 \left(P_{n+2}^{(n)}(1) - P_{n+1}^{(n)}(1) \cdot P_{n+2}^{(n+1)}(1) \right) \partial^{n+2} u + \mathrm{O}(h^3)$$

Analogously, this can be used for a second correction leading to $R^{n,3}$. Moreover it becomes clear, how a (rather complicated) recursion relation for the coefficients could be derived.

Remark 7.3. In the case of stencils with higher accuracy the approach of concatenation does not work, what is documented by the following example:
Computing $R^{1,2}$ at first by means of (7.7)

$$R^{1,2} = R^{1,1} - \tfrac{1}{2} R^{2,1} = [\overset{*}{-1},\ 1] - \tfrac{1}{2}[\overset{*}{1},\ -2,\ 1] = \tfrac{1}{2}[\overset{*}{-3},\ 4,\ -1],$$

self-concatenation yields

$$\begin{aligned} R^{1,2} \circ R^{1,2} &= \tfrac{1}{2}[\overset{*}{-3},\ 4,\ -1] \circ \tfrac{1}{2}[\overset{*}{-3},\ 4,\ -1] \\ &= -\tfrac{3}{4}[\overset{*}{-3},\ 4,\ -1] \ +\ \tfrac{4}{4}[\overset{*}{0},\ -3,\ 4,\ -1] \ -\ \tfrac{1}{4}[\overset{*}{0},\ 0,\ -3,\ 4,\ -1] \\ &= \tfrac{1}{4}[\overset{*}{9},\ -24,\ 22,\ -8,\ 1]\ . \end{aligned}$$

This stencil is already by one entry too long to agree with $R^{2,2} = [2,\ -5,\ 4,\ -1]$, that comes out by solving the defining linear system. Hence we have corroborated that $R^{1,2} \circ R^{1,2} \neq R^{2,2}$. Nevertheless $R^{1,2} \circ R^{1,2}$ can be used to approximate the second derivative with second order accuracy.

Remark 7.4. It is possible to derive $R^{2,2}$ from $R^{1,2} \circ R^{1,2}$ by combining the concatenated stencil with an extrapolation stencil, that is used to reduce the length. From proposition 7.1 we can read off the following interpolation weights, that we write now in stencil form

$$[-1,\ 4,\ -6,\ \overset{*}{4}]_h u(x+3h) = u(x+4h) + \mathrm{O}(h^4).$$

The residual $\mathrm{O}(h^4)$ can be checked by Taylor expansion. Observing that this equation is equivalent to $[\overset{*}{-1},\ 4,\ -6,\ 4,\ -1]_h u(x) = \mathrm{O}(h^4)$, we obtain

$$h^{-2} R_h^{1,2} \circ R_h^{1,2} u + \tfrac{1}{4} h^{-2} [\overset{*}{-1},\ 4,\ -6,\ 4,\ -1]_h u = \partial^2 u + \mathrm{O}(h^2).$$

Obviously, summing yields $\tfrac{1}{4}[\overset{*}{9},\ -24,\ 22,\ -8,\ 1] + \tfrac{1}{4}[\overset{*}{-1},\ 4,\ -6,\ 4,\ -1] = [\overset{*}{2},\ -5,\ 4,\ -1,\ 0]$ which is equivalent to $R^{2,2}$ as the 0 at right can be discarded.

We have seen in remark 7.2 that the knowledge of the numbers $P_{n+\nu}^{(n)}(1)$ is important for producing stencils of higher accuracy by the presented correction procedure based on Taylor expansion. Therefore it is worthwile to look for properties of these numbers, that might simplify their computation. The following conjecture summarizes an experimental observation, that we have not yet proved.

Conjecture 7.2. *For $n \in \mathbb{N}_0, \nu \in \mathbb{N}_0$ the sequence $n \mapsto P_n^{(n+\nu)}(1)$ is an* arithmetic sequence[4] *of degree ν with the step-size $2^{-\nu}$.*

[4] A sequence $(a_n)_{n\in\mathbb{N}_0}$ is called an arithmetic sequence, if the difference sequence Da with $(Da)_n := a_{n+1} - a_n =: d_n = d$ is constant. d is the step-size. A sequence $(b_n)_{n\in\mathbb{N}_0}$ is referred to as an arithmetic sequence of degree $\delta \in \mathbb{N}$, if $D^\delta b$ is an arithmetic sequence, where D^δ means δ-times repeated application of the difference operator.

7.2 Model problem: 1D Poisson equation with Dirichlet BC

As model problem being subject to later numerical investigations, we choose the *Poisson equation* with *Dirichlet boundary conditions* on the interval $\mathcal{I} := [0, L]$. The classical variant of this boundary value problem (BVP) is:

For a given data function $f \in \mathcal{C}\big((0, L)\big)$ and boundary values $u_0, u_L \in \mathbb{R}$ find a function $u \in \mathcal{C}\big([0, L]\big) \cap \mathcal{C}^2\big((0, L)\big)$, that satisfies the BVP

$$u(0) = u_0,\ \ u(L) = u_L \quad \text{and} \quad \partial^2 u(x) = f(x) \quad \text{for } x \in (0, L),\ L > 0. \tag{7.8}$$

For reasons that become clear further below, it is advantageous to consider uniformly continuous sources $f \in \mathcal{C}\big([0, L]\big)$ implying $u \in \mathcal{C}^2([0, L])$. Introducing the operator

$$D : \mathcal{C}^2\big([0, L]\big) \to \mathbb{R} \times \mathcal{C}\big([0, L]\big) \times \mathbb{R} \quad \text{with} \quad u \mapsto Du := \big(u(0), u'', u(L)\big), \tag{7.9}$$

equation (7.8) can be reformulated into

$$Du = (u_0, f, u_L). \tag{7.10}$$

The operator D, consisting of the second derivative (Laplacian) together with the projection onto the boundary values, admits an *inverse operator* $G := D^{-1}$, that can be computed analytically. G turns out to be an integral operator, whose kernel function is referred to as the *Green's function* $\mathscr{G} : [0, L] \times [0, L] \to \mathbb{R}$:

$$\mathscr{G}(x, \xi) := \frac{x\xi}{L} - \min\{x, \xi\} \ = \begin{cases} \frac{1}{L}\xi(x - L) & \text{for } 0 \le \xi \le x \le L \\ \frac{1}{L}x(\xi - L) & \text{for } 0 \le x \le \xi \le L \end{cases} \tag{7.11}$$

Important properties of the Green's function are its symmetry $\mathscr{G}(x, \xi) = \mathscr{G}(\xi, x)$, its piecewise linearity and its non-positivity[5]. It will be proved in theorem 7.4, that $G : \mathbb{R} \times \mathcal{C}\big([0, L]\big) \times \mathbb{R} \to \mathcal{C}^2\big([0, L]\big)$ is given by

$$G : (u_0, f, u_L) \mapsto u \ \text{ with } \ u(x) = u_0 + \tfrac{x}{L}(u_L - u_0) + \int_0^L \mathscr{G}(x, \xi)\, f(\xi)\, d\xi. \tag{7.12}$$

If one of the arguments of $\mathscr{G}$ is kept fixed, the resulting function exhibits a kink in its extremum ($x = \xi$), where its derivative makes a leap by 1. This jump condition is the crucial characteristic[6] of G, exploited to prove (7.12).

Formula (7.12) guarantees the existence of a solution $u \in \mathcal{C}^2([0, L])$ for a right hand side $f \in \mathcal{C}^0([0, L])$ being continuous on the *compact* interval[7]. Moreover, the differential equation enforces, that $f \in \mathcal{C}^n([0, L])$ entails $u \in \mathcal{C}^{n+2}([0, L])$ for any $n \in \mathbb{N}_0$.

[5] Instead of (7.8) many authors consider the equation $-\partial^2 u = f$ (notice the minus sign), since $-\partial^2$ is a positive operator implying that the Green's function is non-negative.

[6] Formally, the derivative of a jump function (Heaviside function) yields a δ-function. So $\partial_x^2 \mathscr{G}(x, \xi) = \delta(x - \xi)$.

[7] In contrast, if f is only continuous on the *open* interval, then f need not be integrable over $(0, L)$ and thence the formula might not be applicable.

Therefore the inverse operator G acts in a smoothing way.

In the framework of finite differences, the solution u is normally required to be at least in $\mathcal{C}^4([0,L])$. These stronger assumptions are necessary to be able to deploy Taylor expansion as a simple but mighty analytic tool[8] (confer section 7.6).

Theorem 7.3. *(uniqueness) The Poisson equation with Dirichlet boundary conditions (7.8) has at most one solution in* $\mathcal{C}([0,L]) \cap \mathcal{C}^2\big((0,L)\big)$.

Proof: Suppose there exist two different solutions $u \neq v$. It is to show, that they are actually equal. Consider the difference $w := u - v$ satisfying the homogeneous Poisson equation. From $w'' \equiv 0$ follows $w' \equiv const$ and hence $w(x) = c_1 x + c_0$ with some constants c_0, c_1. The homogeneous boundary conditions $w(0) = w(L) = 0$ require however $c_1 = c_0 = 0$. Thus $w \equiv 0$ which is equivalent to $u \equiv v$. ■

For general Sturm-Liouville differential operators of second order the construction of the Green's function and the inverse operator is done by means of a *fundamental system.* In the present case, we can however gain (7.12) by a repeated integration of the differential equation.

Theorem 7.4. *Let $f \in \mathcal{C}([0,L])$. Then the Poisson equation with Dirichlet boundary conditions (7.8) admits a solution $u \in \mathcal{C}^2([0,L])$ which is given by (7.12).*

Proof: A first representation of the solution is obtained by integrating the Poisson equation $u'' = f$ twice resulting in

$$u(x) = \int_0^x \int_0^\zeta f(\xi)\, d\xi d\zeta \;+\; C_1 x \;+\; C_0 \,.$$

The integration constants C_1, C_0 have to be adapted to the boundary conditions. This yields:

$$u(x) = u_0 \;+\; \int_0^x \int_0^\zeta f(\xi)\, d\xi d\zeta \;-\; \frac{x}{L}\left[\int_0^L \int_0^\zeta f(\xi)\, d\xi d\zeta \;+\; u_0 - u_L\right] \tag{7.13}$$

According to the fundamental theorem of calculus, u is in $\mathcal{C}^2([0,L])$ and solves (7.8). It remains to be shown that (7.13) can be transformed into (7.12). For this we have to interchange the order of integration using Fubini's theorem.

$$u(x) = u_0 + \int_0^x \int_\xi^x f(\xi)\, d\zeta d\xi - \frac{x}{L}\int_0^L \int_\xi^L f(\xi)\, d\zeta d\xi - \tfrac{x}{L} u_0 + \tfrac{x}{L} u_L$$

Performing the integration with respect to ζ gives

$$u(x) = u_0 + \frac{x}{L}(u_L - u_0) \;+\; \int_0^x (x-\xi) f(\xi)\, d\xi \;-\; \frac{x}{L}\int_0^L (L-\xi) f(\xi)\, d\xi$$

The two integrals abbreviated by I_1, I_2 have to be considered more carefully:

$$I_1 + I_2 = \; x\int_0^x f(\xi)\, d\xi \;-\; \int_0^x \xi f(\xi)\, d\xi \;-\; x\int_0^L f(\xi)\, d\xi \;+\; \frac{x}{L}\int_0^L \xi f(\xi)\, d\xi$$

By splitting the third integral from 0 to x and from x to L, the first integral is canceled:

$$= -\int_0^x \xi f(\xi)\, d\xi \;-\; x\int_x^L f(\xi)\, d\xi \;+\; \frac{x}{L}\int_0^L \xi f(\xi)\, d\xi \,,$$

[8] The significance of Taylor expansion is based on its capability to link the *realm of differences* with the *realm of derivatives.* As long as we are primarily interested in the difference schemes to approximate differential equations this is an indispensable tool.

Decomposing likewise the last integral and drawing together the integrals over the same domain produces finally the Green's function as kernel function under the integral.

$$= \frac{x-L}{L}\int_0^x \xi f(\xi)\,d\xi \;+\; \frac{x}{L}\int_x^L (\xi - L)f(\xi)\,d\xi \;=\; \int_0^L \mathscr{G}(x,\xi)f(\xi)\,d\xi$$

This finishes the proof of the theorem. ■

Actually, the applicability of formula (7.12) is not restricted to source functions in $\mathcal{C}([0,L])$ but holds for a broader class of functions. This is supported by the following examples:

$$\begin{array}{lllll} f(x) = x^{-1} & \wedge\; u_0 = 0 & \wedge\; u_1 = 0 & \Rightarrow & u(x) = x\log(x) \\ f(x) = x^{-3/2} & \wedge\; u_0 = 0 & \wedge\; u_1 = 0 & \Rightarrow & u(x) = -2x^{1/2} + 2x \end{array}$$

As (7.12) can obviously be applied to source functions that become singular at the boundary, the condition requiring f to be continuous on the *compact* interval $[0,L]$ should be relaxed.

For this, let us define a space of functions being locally integrable[9] and being additionally integrable over $[0,L]$ if multiplied by a $\mathcal{C}^1([0,L])$ (weight)-function, that is supposed to vanish at the boundary points 0 and L:

$$\mathcal{L}^1_w\big((0,L)\big) := \Big\{ g \in \mathcal{L}^1_{loc}\big((0,L)\big) :\; x \mapsto xg(x) \in \mathcal{L}^1_{loc}\big([0,L)\big) \;\wedge\; x \mapsto (L-x)g(x) \in \mathcal{L}^1_{loc}\big((0,L]\big) \Big\}$$

The boundary constraints on the admissible weight functions serve to compensate possible singularities of the source functions. It is obvious that (7.12) is well defined for functions of this space. In contrast it appears questionable, whether (7.12) is then also a solution of the Poisson equation even if the source is continuous in $(0,L)$.

This problem is avoided by considering the following subspace of $\mathcal{L}^1_w\big((0,L)\big)$. We write $f \in \mathcal{L}^1_{dom}\big((0,L)\big)$, if there exist $\alpha > 0$ and constants $C_0, C_L > 0$ such that $f \in \mathcal{L}^1_{loc}\big((0,L)\big)$ satisfies

$$|f(x)| < C_0 x^{-2+\alpha} \quad \text{for } 0 < x \le \delta \quad \text{and} \quad |f(x)| < C_L(L-x)^{-2+\alpha} \quad \text{for } L-\delta \le x < L$$

Lemma 7.2. *Assume* $f \in \mathcal{L}^1_{dom}\big((0,L)\big)$ *and* $z \in (0,L)$. *Then:*

$$\begin{array}{llll} \textit{i)} & \displaystyle\lim_{\epsilon\downarrow 0}\int_0^\epsilon |\xi f(\xi)|\,d\xi = 0 & \textit{and} & \displaystyle\lim_{\epsilon\downarrow 0}\int_{L-\epsilon}^L |(L-\xi)f(\xi)|\,d\xi = 0 \\ \textit{ii)} & \displaystyle\lim_{\epsilon\downarrow 0}\,\epsilon\int_\epsilon^z |f(\xi)|\,d\xi = 0 & \textit{and} & \displaystyle\lim_{\epsilon\downarrow 0}\,\epsilon\int_z^{L-\epsilon} |f(\xi)|\,d\xi = 0 \end{array}$$

Proof: The proof is straightforward by employing the majorants. ■

Theorem 7.5. *The Poisson equation with Dirichlet boundary conditions has, for any* $f \in \mathcal{C}\big((0,L)\big) \cap \mathcal{L}^1_{dom}\big((0,L)\big)$, *a solution* $u \in \mathcal{C}^2\big((0,L)\big) \cap \mathcal{C}\big([0,L]\big)$ *that is represented by (7.12).*

Proof: In order to shorten the writing we consider the case of homogeneous boundary conditions $u(0) = u(L) = 0$ without restriction of generality. Let us define u for any $x \in (0,L)$:

$$u(x) = \frac{x-L}{L}\int_0^x \xi f(\xi)\,d\xi \;+\; \frac{x}{L}\int_x^L (\xi - L)f(\xi)\,d\xi\;.$$

From the assumption follows $f \in \mathcal{L}^1_w\big((0,L)\big)$; thence the integrals are well defined. By the fundamental theorem of calculus, an integral represents a differentiable function with respect to its upper and lower integration bound if the integrand is continuous[10]. Applying the product rule

[9] integrable over every *compact* subset

[10] If the integrand should be singular at the boundary, the integral interval may be decomposed, in order to verify differentiability at arbitrary but fixed $x \in (0,L)$. For instance: as x lies in the open interval we can always find an intermediate position $0 < z < x$, such that $\int_0^x = \int_0^z + \int_z^x$. The first integral is then a constant while the second one can be differentiated with respect to x.

yields:

$$\begin{aligned}
u'(x) &= \frac{1}{L}\int_0^x \xi f(\xi)\,d\xi \;+\; \frac{x-L}{L}xf(x) \;+\; \frac{1}{L}\int_0^x \xi f(\xi)\,d\xi \;-\; \frac{x}{L}(x-L)f(x) \\
&= \frac{1}{L}\int_0^x \xi f(\xi)\,d\xi \;+\; \frac{1}{L}\int_0^x \xi f(\xi)\,d\xi \\
u''(x) &= \frac{1}{L}xf(x) \;-\; \frac{1}{L}(x-L)f(\xi) \;=\; f(x)
\end{aligned}$$

As $f \in \mathcal{C}\big((0,L)\big)$ we see that $u \in \mathcal{C}^2\big((0,L)\big)$. Moreover the Poisson equation is obviously satisfied. The continuity of u holds by construction in the open interval $(0,L)$. To verify the continuity at the two boundary points we consider:

$$\lim_{\epsilon\downarrow 0} |u(\epsilon)| = \lim_{\epsilon\downarrow 0}\Big|\frac{\epsilon-L}{L}\int_0^{\epsilon} \xi f(\xi)\,d\xi \;+\; \frac{\epsilon}{L}\int_{\epsilon}^{L} (\xi-L)f(\xi)\,d\xi\Big|$$

Applying the triangle inequality and decomposing the integration domain of the second integral into $(\epsilon,z)\cup(z,L)$, yields:

$$\lim_{\epsilon\downarrow 0} |u(\epsilon)| \le \lim_{\epsilon\downarrow 0}\Big\{\frac{L-\epsilon}{L}\int_0^{\epsilon} |\xi f(\xi)|\,d\xi \;+\; \frac{\epsilon}{L}\int_{\epsilon}^{z} |(\xi-L)f(\xi)|\,d\xi \;+\; \frac{\epsilon}{L}\int_{z}^{L} |(\xi-L)f(\xi)|\,d\xi\Big\}$$

The first addend vanishes in the limit since the integral tends to zero by part i) of lemma 7.2. The same is true for the third addend due to the factor ϵ. Observe that the integral remains bounded because $(\xi-L)f(\xi) \in \mathcal{L}^1_{loc}((0,L])$. The second addend becomes:

$$\frac{\epsilon}{L}\int_{\epsilon}^{z} |(\xi-L)f(\xi)|\,d\xi \;=\; \frac{\epsilon}{L}\int_{\epsilon}^{z} |\xi f(\xi)|\,d\xi \;+\; \epsilon\int_{\epsilon}^{z} |f(\xi)|\,d\xi$$

The factor ϵ makes the first addend vanish, as the integral is bounded due to $\xi f(\xi) \in \mathcal{L}^1_{loc}([0,L))$. Furthermore the second addend goes to zero by part ii) of lemma 7.2. Hence we get $\lim_{\epsilon\downarrow 0} u(\epsilon) = u(0) = 0$. An analog computation can be performed for the right boundary at L. This shows $u \in \mathcal{C}([0,L])$ as asserted. ■

7.3 Discretization of Dirichlet boundary conditions

In order to discretize the BVP (7.8) by finite differences, the interval $\mathcal{I} = [0,L]$ is covered by an equidistant grid $\mathcal{I}_h := \{x_0,...,x_N\} \subset \mathcal{I}$ containing $N+1$ nodes $x_i := ih$ with $i \in \{0,...,N\}$ and $h := \frac{L}{N}$. The numerical solution is computed as a grid function $v : \mathcal{I}_h \to \mathbb{R}$. Instead of $v(x_i)$ we write also v_i. Note, that v as any other discrete quantity or operator on $\mathcal{F}(\mathcal{I}_h)$ depends already by its domain of definition on the number of grid nodes or equivalently on the grid spacing h. Nevertheless we abstain from writing v_h, in order to avoid a too cumbersome and overloaded notation.

A function[11] $g : \mathcal{I} \to \mathbb{R}$ defined on the spatial domain $\mathcal{I}$ can be "compared" with a grid function in the easiest way by restricting[12] it onto the grid, i.e. $\hat{g} : \mathcal{I}_h \to \mathbb{R}$ with $\hat{g}_i \equiv \hat{g}(x_i) := g(x_i)$.

[11] In particular the solution u of the BVP (7.8)

[12] The projection from $\mathcal{F}(\mathcal{I})$ to $\mathcal{F}(\mathcal{I}_h)$ is called *sampling*.

The second derivative is usually discretized by the three-point central difference quotient. This leads to the *standard discretization* of BVP (7.8)

$$v_0 = u_0, \quad v_N = u_L, \qquad \frac{v_{i-1} - 2v_i + v_i}{h^2} = f(ih) \quad \text{for } i \in \{1, ..., N-1\}. \tag{7.14}$$

The matrix, pertaining to the above linear system, discretizes the operator D defined in (7.10) and is therefore abbreviated by $\hat{D}$.
The Dirichlet boundary conditions fit ideally with the three-point stencil. The values at the boundary nodes x_0, x_N are already prescribed by the boundary conditions, whereas the interior nodes $x_1, ..., x_{N-1}$ have only to "communicate" with their direct neighbors.

The situation gets more sophisticated if the stencil becomes broader. For this, let us consider the symmetric five-point stencil

$$[a,\ b,\ \overset{*}{c},\ b,\ a] = \tfrac{1}{12}[-1,\ 16,\ -\overset{*}{30},\ 16,\ -1],$$

that offers another, more accurate discretization of the second derivative. The remainder of this section we will be concerned with the construction of appropriate boundary conditions. The following sections are then devoted to their theoretical and experimental analysis.

Lemma 7.3. *Let $I \subset \mathbb{R}$ be an interval and $g \in \mathcal{C}^6(I)$. For any $x \in \overset{\circ}{I}$ there exists $h_{max} = h_{max}(x) > 0$ and a constant $C = C(x) > 0$, such that the following estimate holds true for any h with $|h| \leq h_{max}$:*

$$\left| \tfrac{1}{12h^2}[-1,\ 16,\ -\overset{*}{30},\ 16,\ -1]_h\, g(x) \ -\ g''(x) \right| \ <\ C\, h^4\,.$$

C can be chosen independently from $x \in \overset{\circ}{I}$, if I is compact!

The verification can be done by writing out the stencil and replacing $g(x-2h), g(x-h), g(x+h)$ and $g(x+2h)$ by the truncated Taylor expansion up to sixth order (the remainder term is the sixth order). Let us here proceed the other way round by demonstrating exemplarily, how the stencil is constructed. For this we make the ansatz of a symmetric five-point stencil.

$$\tfrac{1}{h^2}[a,\ b,\ \overset{*}{c},\ b,\ a]_h\ g(x) \overset{!}{=} g''(x) + \mathrm{O}(h^4)$$

This ansatz is justified by the following reasons:

i) Odd derivatives will be canceled, because they occur with opposite sign in the truncated Taylor series of $g(x-2h), g(x+2h)$ and $g(x-h), g(x+h)$.

ii) To get a fourth order accurate approximation of the second derivative, the zeroth and fourth derivative must be canceled, while the coefficients in front of the second derivative should sum up to one.

Therefore we obtain three equations for three unknowns.

$$\begin{array}{lcccccc} g(x) & : & 2a + 2b + c & = & 0 & & \\ \partial^2 g(x) : & & 2 \cdot \frac{2^2}{2!}a + 2 \cdot \frac{1}{2!}b & = & 2a + \frac{1}{2}b & = & 1 \\ \partial^4 g(x) : & & 2 \cdot \frac{4^2}{4!}a + 2 \cdot \frac{1}{4!}b & = & \frac{4}{3}a + \frac{1}{12}b & = & 0 \end{array}$$

The solution of this system is just given by: $a = -\frac{1}{12}, b = \frac{16}{12}, c = -\frac{30}{12}$.
Choosing $h_{max} < 2\,\mathrm{dist}(\partial I, x)$, $g(x \pm 2h)$ and $g(x \pm h)$ are defined for every h with $|h| \leq h_{max}$. The truncated Taylor expansion yields

$$\tfrac{1}{h^2}\left[a, b, \overset{*}{c}, b, a\right]_h\ g(x) = g''(x) + h^4\left(\tfrac{2^6 a}{6!}\partial^6 g(\theta_2) + \tfrac{b}{6!}\partial^6 g(\theta_1) + \tfrac{b}{6!}\partial^6 g(\eta_1) + \tfrac{2^6 a}{6!}\partial^6 g(\eta_2)\right),$$

where $\theta_2 \in (x-2h,x), \theta_1 \in (x-h,x), \eta_1 \in (x,x+h)$ and $\eta_2 \in (x,x+2h)$ are appropriate intermediate positions. Since the continuous function $|\partial^6 g|$ attains a maximum on the compact subinterval $[x-h_{max}, x+h_{max}] \subset\subset I$, the remainder term can be estimated as asserted with

$$C(x) := 2\frac{2^6+1}{6!} \max\{a,b\} \max_{|\xi-x|<h_{max}} |\partial^6 g(\xi)| .$$

If I is compact, $\partial^6 g$ has a maximum on the entire interval. Hence the constant C can be defined equally for all $x \in \mathring{I}$.

In contrast to the three-point stencil, the five-point stencil leads to a conflict with the Dirichlet boundary conditions. As every node needs two neighbors on both sides, the stencil can not be applied in x_1 and x_{N-1}. There are several possibilities to overcome this difficulty:

1) Apply the three-point stencil in x_1 and x_{N-1}.

2) Extrapolate to get v_{-1}, v_{N+1}; apply the five-point stencil in x_1, x_{N-1}.

3) Use asymmetric or one-sided finite difference quotients in x_1, x_{N-1}.

The first idea leads directly to the scheme:

$$\begin{aligned}
v_0 &= u_0 \\
\tfrac{1}{h^2}[1,\ \overset{*}{-2},\ 1]v_1 &= f(h) \\
\tfrac{1}{12h^2}[-1,\ 16,\ \overset{*}{-30},\ 16,\ -1]v_i &= f(ih) \qquad \text{for } i \in \{2,...,N-2\} \\
\tfrac{1}{h^2}[1,\ \overset{*}{-2},\ 1]v_{N-1} &= f\big((N-1)h\big) \\
v_N &= u_L
\end{aligned} \tag{7.15}$$

For the second idea we have to specify how the extrapolation shall be performed. Let g be a function defined in a vicinity of $x \in \mathbb{R}$, containing the equidistant points $x-h, x, x+h, ..., x+4h$ for $h > 0$ small enough. If $x-h$ plays the role of the extrapolation point, while $x, x+h, ...$ are taken as the sampling nodes, proposition 7.1 yields the following extrapolation stencils:

$$\begin{array}{lcllc}
g(x-h) &=& [\overset{*}{1}]_h\ g(x) & +\,\mathrm{O}(h) & \text{nearest-neighbor} \\
g(x-h) &=& [\overset{*}{2},-1]_h\ g(x) & +\,\mathrm{O}(h^2) & \text{linear} \\
g(x-h) &=& [\overset{*}{3},-3,1]_h\ g(x) & +\,\mathrm{O}(h^3) & \text{quadratic} \\
g(x-h) &=& [\overset{*}{4},-6,4,-1]_h\ g(x) & +\,\mathrm{O}(h^4) & \text{cubic} \\
g(x-h) &=& [\overset{*}{5},-10,10,-5,1]_h\ g(x) & +\,\mathrm{O}(h^5) & \text{quartic}
\end{array}$$

Assuming g to be sufficiently smooth, it can be shown by straight-forward Taylor expansion, that the difference between the actual value $g(x-h)$ and the result of the extrapolation is of the indicated order.
Because of theorem 7.1 these extrapolation formulas are also valid, if h is replaced by

$-h$. Then the right-sided stencils become left-sided with the coefficients in converse order. For instance:

$$g(x+h) = [\,1,\ -5,\ 10,\ -10,\ \overset{*}{5}\,]_h\ g(x)\ +\ \mathrm{O}(h^5)$$

Since the extrapolation can obviously be applied in x_{N+1} analogously to x_{-1}, we only need to consider the left boundary in detail.
Let us assume, that x_{-1} would belong to the grid. Then the stencil in x_1 would be

$$\tfrac{1}{h^2}[a,\ b,\ \overset{*}{c},\ b,\ a]v_1 = \tfrac{1}{h^2}\Big(av_{-1} + bv_0 + cv_1 + bv_2 + av_3\Big) = f(h)\ . \tag{7.16}$$

As v_{-1} is only fictitious, we must replace it by an extrapolating expression accessing the values at the right neighbors of x_{-1}. According to the preparative excursion we can consider the following options:

$v_{-1} \to v_0$	nearest-neighbor
$v_{-1} \to 2v_0 - v_1$	linear
$v_{-1} \to 3v_0 - 3v_1 + v_2$	quadratic
$v_{-1} \to 4v_0 - 6v_1 + 4v_2 - v_3$	cubic
$v_{-1} \to 5v_0 - 10v_1 + 10v_2 - 5v_3 + v_4$	quartic

Hence, substituting v_{-1}, equation (7.16) becomes by utilizing the

0) nearest-neighbor (0^{th} order/constant polynomial) extrapolation:

$$\begin{aligned} \tfrac{1}{h^2}\Big((b+a)v_0 + cv_1 + bv_2 + av_3\Big) &= f(h) \\ \Leftrightarrow \qquad \tfrac{1}{12h^2}[15,\ -\overset{*}{30},\ 16,\ -1]\ v_1 &= f(h) \end{aligned}$$

I) linear (1^{st} order polynomial) extrapolation:

$$\begin{aligned} \tfrac{1}{h^2}\Big((b+2a)v_0 + (c-a)v_1 + bv_2 + av_3\Big) &= f(h) \\ \Leftrightarrow \qquad \tfrac{1}{12h^2}[14,\ -\overset{*}{29},\ 16,\ -1]\ v_1 &= f(h) \end{aligned}$$

II) quadratic (2^{nd} order polynomial) extrapolation:

$$\begin{aligned} \tfrac{1}{h^2}\Big((b+3a)v_0 + (c-3a)v_1 + (b+a)v_2 + av_3\Big) &= f(h) \\ \Leftrightarrow \qquad \tfrac{1}{12h^2}[13,\ -\overset{*}{27},\ 15,\ -1]\ v_1 &= f(h) \end{aligned}$$

III) cubic (3^{rd} order polynomial) extrapolation:

$$\begin{aligned} \tfrac{1}{h^2}\Big((b+4a)v_0 + (c-6a)v_1 + (b+4a)v_2 + (a-a)v_3\Big) &= f(h) \\ \Leftrightarrow \qquad \tfrac{1}{12h^2}[12,\ -\overset{*}{24},\ 12,\ 0]\ v_1 \ = \ \tfrac{1}{h^2}[1,\ -\overset{*}{2},\ 1]\ v_1 &= f(h) \end{aligned}$$

Remarkably, the cubic extrapolation leads to the same equation in x_1 that occurs in the mixed stencil scheme (7.15), where the five-point stencil in x_1, x_{N-1} has been supplanted by the three-point stencil (first idea).

IV) quartic (4^{th} order polynomial) extrapolation:

$$\begin{aligned} \tfrac{1}{h^2}\Big((b+5a)v_0 + (c-10a)v_1 + (b+10a)v_2 - 4av_3 + av_4\Big) &= f(h) \\ \Leftrightarrow \qquad \tfrac{1}{12h^2}[11,\ -\overset{*}{20},\ 6,\ -4,\ 1]\, v_1 &= f(h) \end{aligned}$$

Summarizing, we obtain the following class of *extrapolation schemes*

$$\begin{aligned} v_0 &= u_0 \\ \tfrac{1}{h^2}[\varrho_0^\mu,\ \overset{*}{\varrho_1^\mu},\ \varrho_2^\mu,\ \varrho_3^\mu,\ \varrho_4^\mu]v_1 &= f(h) \\ \tfrac{1}{h^2}[a,\ b,\ \overset{*}{c},\ b,\ a]v_i &= f(ih) \qquad \text{for } i \in \{2, ..., N-2\} \\ \tfrac{1}{h^2}[\varrho_4^\mu,\ \varrho_3^\mu,\ \varrho_2^\mu,\ \overset{*}{\varrho_1^\mu},\ \varrho_0^\mu]v_{N-1} &= f(h) \\ v_N &= u_L \end{aligned} \tag{7.17}$$

with

$$[\varrho_0^\mu,\ \overset{*}{\varrho_1^\mu},\ \varrho_2^\mu,\ \varrho_3^\mu,\ \varrho_4^\mu] := \begin{cases} \mu = e_0: & [b+a,\ \overset{*}{c},\ b,\ a,\ 0] \\ \mu = e_1: & [b+2a,\ \overset{*}{c-a},\ b,\ a,\ 0] \\ \mu = e_2: & [a+3a,\ \overset{*}{c-3a},\ b+a,\ a,\ 0] \\ \mu = e_3: & [a+4a,\ \overset{*}{c-6a},\ b+4a,\ 0,\ 0] \\ \mu = e_4: & [a+5a,\ \overset{*}{c-10a},\ b+10a,\ -4a,\ a] \end{cases} . \tag{7.18}$$

Like $\hat{D}$ the associated system matrices $\hat{D}_\mu$ represent discretizations of the operator D.

$$\hat{D}_\mu := \frac{1}{h^2}\begin{pmatrix} h^2 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \varrho_0^\mu & \varrho_1^\mu & \varrho_2^\mu & \varrho_3^\mu & \varrho_4^\mu & 0 & 0 & \dots & 0 \\ a & b & c & b & a & 0 & 0 & \dots & 0 \\ 0 & a & b & c & b & a & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a & b & c & b & a & 0 \\ 0 & \dots & 0 & 0 & a & b & c & b & a \\ 0 & \dots & 0 & 0 & \varrho_4^\mu & \varrho_3^\mu & \varrho_2^\mu & \varrho_1^\mu & \varrho_0^\mu \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & h^2 \end{pmatrix} \tag{7.19}$$

For the sake of completeness let us mention the outcome of the third idea. The ansatz of an *asymmetric* difference quotient like $\frac{1}{h^2}[\alpha,\ \overset{*}{\beta},\ \gamma,\ \delta,\ (\epsilon)]$ leads to

$$\begin{aligned} \tfrac{1}{h^2}[1,\ -\overset{*}{2},\ 1,\ 0]_h\, g(x) &= g''(x) + O(h^2)\ , \\ \tfrac{1}{h^2}[\,\tfrac{11}{12},\ -\overset{*}{\tfrac{5}{3}},\ \tfrac{1}{2},\ \tfrac{1}{3},\ -\tfrac{1}{12}\,]_h\, g(x) &= g''(x) + O(h^3) \end{aligned}$$

Comparing these stencils with those of the cubic and quartic extrapolation scheme in x_1, we recognize, that this approach reproduces only what we have already found.

In contrast, the ansatz of *one-sided* difference quotients $\frac{1}{h^2}[\overset{*}{\alpha},\ \beta,\ \gamma,\ (\delta)]$ generates alternative schemes, that will not be studied in the sequel, however. From remark 7.1 and 7.4 we get the subsequent stencils for the second derivative

$$\begin{aligned} \tfrac{1}{h^2}[\overset{*}{1},\ -2,\ 1]_h\, g(x) &= g''(x) + O(h)\ , \\ \tfrac{1}{h^2}[\overset{*}{2},\ -5,\ 4,\ -1]_h\, g(x) &= g''(x) + O(h^2)\ . \end{aligned}$$

The one-sided three-point stencil is just the classical central three-point stencil for the second derivative, shifted to the right and losing therefore one order of accuracy.

7.4 Stability of extrapolation schemes

In the last section we have constructed different discretizations of the operator D introduced in equation (7.10). It is generally known, that differential operators are not bounded in the supremum norm. This property is inherited to the discretizations. Surely, the discrete operators $\hat{D}, \hat{D}_\mu$ act on finite dimensional spaces and are therefore bounded. However, their norm grows over all bounds, if the grid spacing h tends to 0 (i.e. with increasing dimension of the system). This can be easily seen as follows: The discrete operators $\hat{D}, \hat{D}_\mu$ are represented by band matrices whose band width does not change with the grid spacing h. As the coefficients depend on the grid spacing only by the factor $\frac{1}{h^2}$ (compare with (7.19)), the induced $\|\cdot\|_\infty$ matrix norms behave like $\mathrm{O}(h^{-2})$ and "explode" in the limit $h \downarrow 0$.
In opposition to D, its inverse operator G is bounded as being an integral operator. Therefore the question arises, whether the norm of the inverse operators of $\hat{D}, \hat{D}_\mu$ remains bounded with respect to h. Clearly, this property should be expected, as $\hat{D}^{-1}, \hat{D}_\mu^{-1}$ are supposed to approximate G.

Definition 7.1. *A scheme discretizing the BVP (7.8) is called* stable*, if the inverse of its system matrix can be estimated independently of the grid spacing with respect to the $\|\cdot\|_\infty$ operator matrix norm.*
In particular, the standard scheme and the extrapolation schemes are stable, if

$$\|\hat{D}^{-1}\|_\infty = \mathrm{O}(1) \qquad \textit{and} \qquad \|\hat{D}_\mu^{-1}\|_\infty = \mathrm{O}(1) \qquad \textit{for}\ h \downarrow 0.$$

As G is the inverse of D, it is reasonable to assume, that a discretization (sampling) of G is somehow close to the inverse of $\hat{D}, \hat{D}_\mu$. Let us therefore project the Green's function onto the grid:

$$\hat{\mathscr{G}}_{ij} := \mathscr{G}(ih, jh) = \mathscr{G}(x_i, \xi_j) = ijh^2L^{-1} - h\min\{i,j\} \ \text{ for } \ i,j \in \{0,1,...,N\} \quad (7.20)$$

The obvious symmetry property $\mathscr{G}(x,\xi) = \mathscr{G}(\xi,x)$ entails $\hat{\mathscr{G}}_{ij} = \hat{\mathscr{G}}_{ji}$. There is another symmetry property[13], that should be at our disposition for later purposes. We check it for the discrete case.

[13]The "continuous" analogon to lemma 7.4 is $\mathscr{G}(x, L-\xi) = \mathscr{G}(L-x,\xi)$ and $\mathscr{G}(x,\xi) = \mathscr{G}(L-x, L-\xi)$ respectively.

Lemma 7.4. *Let $\hat{\mathscr{G}}$ denote the $(N+1)\times(N+1)$ matrix representing the restriction of the Green's function on a uniform grid with $N+1$ nodes in each direction. Then*

$$\text{i) } \hat{\mathscr{G}}_{k(N-l)} = \hat{\mathscr{G}}_{(N-k)l}, \qquad \text{ii) } \hat{\mathscr{G}}_{kl} = \hat{\mathscr{G}}_{(N-k)(N-l)} \qquad \text{for } k,l \in \{0,...,N\}.$$

Proof: By (7.20) the assertion i) is equivalent to

$$k(N-l)hL^{-1} - \min\{k, N-l\} = (N-k)lhL^{-1} - \min\{N-k, l\}\ , \tag{7.21}$$

where both sides have been divided by h. In order to evaluate the minimum, two cases have to be distinguished:

$$\text{case 1:} \quad k \le N-l \quad \Leftrightarrow \quad l \le N-k$$
$$\text{case 2:} \quad k > N-l \quad \Leftrightarrow \quad l > N-k$$

Using $h = \frac{L}{N}$ or $hL^{-1} = \frac{1}{N}$ respectively, we get from (7.21) in the case of

$$1) \qquad \tfrac{k(N-l)}{N} - k = \tfrac{(N-k)l}{N} - l \quad \Leftrightarrow \quad k - \tfrac{kl}{N} - k = l - \tfrac{kl}{N} - l$$
$$2) \quad \tfrac{k(N-l)}{N} - (N-l) = \tfrac{(N-k)l}{N} - (N-k) \ \Leftrightarrow\ k - \tfrac{kl}{N} - N + l = l - \tfrac{kl}{N} - N + k.$$

So we end up with the same terms on each side for both cases.
As for ii) we have to show

$$h^{-1}\mathscr{G}_{kl} = \tfrac{kl}{N} - \min\{k,l\} \stackrel{!}{=} \tfrac{(N-k)(N-l)}{N} - \min\{N-k, N-l\} = h^{-1}\mathscr{G}_{(N-k)(N-l)}\ .$$

Since the expressions are symmetric with respect to k, l we can suppose $k \le l \ \Leftrightarrow\ N-k \ge N-l$ without loss of generality. Thus

$$\tfrac{kl}{N} - k \stackrel{!}{=} \tfrac{N^2 - Nl - Nk + kl}{N} - (N-l)$$

This equation is verified by elementary cancellations. ■

Note that the two properties require each other. Setting for instance $\kappa = N-k$ we have for $\kappa, k, l \in \{0,...,N\}$: $\quad \hat{\mathscr{G}}_{k(N-l)} = \hat{\mathscr{G}}_{(N-k)l} \ \stackrel{\kappa=N-k}{\Leftrightarrow}\ \hat{\mathscr{G}}_{(N-\kappa)(N-l)} = \hat{\mathscr{G}}_{\kappa l}\ .$

In order to discretize G, we approximate the integral by the left-point quadrature (rectangular rule), such that (7.12) becomes[14]

$$u(ih) = \Big(1 - \frac{x_i}{L}\Big)u_0 + \frac{x_i}{L}u_L \ + \ h\sum_{j=0}^{N-1} \hat{\mathscr{G}}_{ij} f(jh) \ + \ \mathrm{O}(h)\ .$$

Observing that the sum starts actually with $j=1$ ($\mathscr{G}_{i0} = 0$ as the Green's function vanishes at the boundary of $[0,L]\times[0,L]$) this equation takes the matrix vector form

$$\begin{pmatrix} \hat{u}_0 \\ \hat{u}_1 \\ \vdots \\ \hat{u}_{N-1} \\ \hat{u}_N \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ \frac{N-1}{N} & h\hat{\mathscr{G}}_{11} & \dots & h\hat{\mathscr{G}}_{1(N-1)} & \frac{1}{N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{N} & h\hat{\mathscr{G}}_{(N-1)1} & \dots & h\hat{\mathscr{G}}_{(N-1)(N-1)} & \frac{N-1}{N} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}}_{=:\hat{G}} \begin{pmatrix} u_0 \\ \hat{f}_1 \\ \vdots \\ \hat{f}_{N-1} \\ u_L \end{pmatrix} + \mathrm{O}(h)$$

defining the *sampled inverse operator* $\hat{G}$. Note that $\hat{G}_{kl}$ is thus given by

$$\hat{G}_{kl} = \tfrac{N-k}{N}\delta_{l0} + h\hat{\mathscr{G}}_{kl} + \tfrac{k}{N}\delta_{lN}\ . \tag{7.22}$$

[14] Actually, the right hand side approximates $u(ih)$ with second order accuracy, i.e. $\mathrm{O}(h)$ could be replaced by $\mathrm{O}(h^2)$. By theorem 7.6 we see, that the right hand side corresponds just to the numeric solution provided by the standard scheme, which is second order accurate.

Remark 7.5. $\hat{G}$ inherits the symmetry property treated in lemma 7.4 from $\hat{\mathscr{G}}$, since it is also satisfied by its left- and rightmost columns, i.e. by the matrix $\left(\frac{N-k}{N}\delta_{l0} + \frac{k}{N}\delta_{lN}\right)_{k,l\in\{0,\ldots,N\}}$.
It can be read off from equation (7.19) and from a matrix representation of $\hat{D}$ respectively, that the $\hat{D}_\mu$'s and $\hat{D}$ possess the same symmetry property (i.e. $(\hat{D}_\mu)_{mk} = (\hat{D}_\mu)_{(N-m)(N-k)}$). We will benefit from this circumstance in the proofs of the lemmata 7.6 and 7.7.

Lemma 7.5. *The norm of the sampled inverse operator can be estimated independently of the grid spacing h or the number of grid nodes respectively. More exactly:*

$\|\hat{G}\|_\infty \leq 1 + 2L^2$ *with L denoting the length of the discretized interval $\mathcal{I} = [0, L]$.*

Proof: In order to compute the $\|\cdot\|_\infty$-operator matrix norm, we have to compute for each row the sum of the moduli of its elements. The norm is then given as the maximum of these sums.
In each row of $\hat{G}$ the sum of the leftmost and rightmost element is one. Therefore it remains to estimate the maximal contribution of the sampled Green's function $\hat{\mathscr{G}}$. Thus:

$$
\begin{aligned}
\|\hat{G}\|_\infty &= 1 + \max_{i=1}^{N-1}\left\{h\sum_{j=1}^{N-1}|\mathscr{G}_{ij}|\right\} = 1 + \max_{i=1}^{N-1}\left\{h\sum_{j=1}^{N-1}\left|ijh^2L^{-1} - h\min\{i,j\}\right|\right\} \\
&\leq 1 + \max_{i=1}^{N-1}\left\{\tfrac{h^3}{L}\, i\sum_{j=1}^{N-1} j + h^2\sum_{j=1}^{N-1}\min\{i,j\}\right\} \\
&\leq 1 + \tfrac{L^3}{N^3}\tfrac{1}{L}\, N\,\tfrac{N(N+1)}{2} + \tfrac{L^2}{N^2}N^2 \\
&\leq 1 + \tfrac{1}{2}L^2\tfrac{N+1}{N} + L^2 \leq 1 + 2L^2
\end{aligned}
$$

■

We have introduced $\hat{G}$, because it seems promising to take it as a guess for approximating $\hat{D}^{-1}$ and $\hat{D}_\mu^{-1}$. Indeed, we have hit the nail on the top as regards the standard scheme (7.14).

Theorem 7.6. ***(Stability of the Standard Scheme)*** *The inverse of the matrix associated with the scheme (7.14) is the sampled inverse operator, i.e.*

$$\hat{D}^{-1} = \hat{G} \qquad \Leftrightarrow \qquad \hat{D}\cdot\hat{G} = I \;\; \textit{(identity)}\,.$$

In combination with lemma 7.5, this implies immediately the stability of the standard method.

Proof: We have to show that $\hat{D}_{kl}G_{lm} = \delta_{km}$, where the product on the left hand side is to be summed over $l \in \{0, ..., N\}$ (convention of summation). As it seems to us more cumbersome to write down a general formula for the elements of $\hat{D}$, we show this in three steps.

1[st] case: $k = 0$

$$\hat{D}_{0l}\hat{G}_{lm} = \delta_{0l}\cdot\left(\tfrac{N-l}{N}\delta_{m0} + h\hat{\mathscr{G}}_{ml} + \tfrac{l}{N}\delta_{mN}\right) = \tfrac{N-0}{N}\delta_{m0} + h\hat{\mathscr{G}}_{m0} + \tfrac{0}{N}\delta_{mN} = \delta_{m0}$$

2^{nd} case: $1 \leq k \leq N-1$

$$
\begin{aligned}
\hat{D}_{kl}\hat{G}_{lm} &= h^{-2}(\delta_{k-1l} - 2\delta_{kl} + \delta_{k+1l}) \cdot (\tfrac{N-l}{N}\delta_{m0} + h\hat{\mathscr{G}}_{ml} + \tfrac{l}{N}\delta_{mN}) \\
&= \tfrac{N-(k-1)\,-\,2(N-k)\,+\,N-(k+1)}{N}\delta_{m0} \;+\; h^{-1}(\delta_{k-1l} - 2\delta_{kl} + \delta_{k+1l})\hat{\mathscr{G}}_{ml} \;+\; \tfrac{(k-1)\,-\,2k\,+\,(k+1)}{N}\delta_{m0} \\
&= (\delta_{k-1l} - 2\delta_{kl} + \delta_{k+1l}) \cdot (\tfrac{ml}{N} - \min\{l,m\}) \\
&= m\tfrac{(k-1)-2k+(k+1)}{N} \;-\; \min\{k-1,m\} + 2\min\{k,m\} - \min\{k+1,m\} \\
&= \big[-(k-1) + 2k - (k+1)\big] + [-2+2]\delta_{(k-1)m} + \delta_{km} = \delta_{km}
\end{aligned}
$$

3^{rd} case: $k = N$

$$\hat{D}_{Nl}\hat{G}_{lm} = \delta_{Nl} \cdot (\tfrac{N-l}{N}\delta_{m0} + h\hat{\mathscr{G}}_{ml} + \tfrac{l}{N}\delta_{mN}) = \tfrac{N-N}{N}\delta_{m0} + h\hat{\mathscr{G}}_{mN} + \tfrac{N}{N}\delta_{mN} = \delta_{mN}$$

For the 1^{st} and 3^{rd} case we have used, that $\mathscr{G}_{ij}$ is zero whenever $i \in \{0, N\}$ or $j \in \{0, N\}$. Note in the 2^{nd} case the representation of the Kronecker-δ by means of the three minima. ■

In the case of the extrapolation schemes based on the five-point stencil, the sampled inverse operator $\hat{G}$ renders also good services, that constitute the first step towards the proof of stability.

Lemma 7.6. *Let $\hat{D}_\mu$ as given in (7.19) with $\mu \in \{e_1, e_2, e_3, e_4\}$. Multiplying $\hat{D}_\mu$ from the right by $\hat{G}$ results in* strictly diagonal-dominant *band matrices J_μ, whose band width is reduced by two in comparison with $\hat{D}_\mu$. More exactly:*

$$
\textit{I)}\quad \hat{D}_{e_1} \cdot \hat{G} \;=:\; J_{e_1} \;=\; \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & \frac{7}{6} & -\frac{1}{12} & 0 & 0 & \dots & 0 \\
0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 & \dots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \dots & 0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 \\
0 & \dots & 0 & 0 & -\frac{1}{12} & \frac{7}{6} & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

$$
\textit{II)}\quad \hat{D}_{e_2} \cdot \hat{G} \;=:\; J_{e_2} \;=\; \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & \frac{13}{12} & -\frac{1}{12} & 0 & 0 & \dots & 0 \\
0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 & \dots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \dots & 0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 \\
0 & \dots & 0 & 0 & -\frac{1}{12} & \frac{13}{12} & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

$$
\textit{III)}\quad \hat{D}_{e_3} \cdot \hat{G} \;=:\; J_{e_3} \;=\; \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & \dots & 0 \\
0 & 1 & 0 & 0 & 0 & \dots & 0 \\
0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 & \dots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \dots & 0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 \\
0 & \dots & 0 & \dots & 0 & 1 & 0 \\
0 & \dots & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

$$IV)\quad \hat{D}_{e_4}\cdot\hat{G} \;=:\; J_{e_4} \;=\; \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \frac{11}{12} & \frac{1}{6} & -\frac{1}{12} & 0 & \dots & 0 \\ 0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 \\ 0 & \dots & 0 & -\frac{1}{12} & \frac{1}{6} & \frac{11}{12} & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

Proof: The verification proceeds similarly to the proof of theorem 7.6.
Ad I) For ease of notation let us set throughout the subsequent calculations: $\bar{D} := D_{e_1}$. The assertion $\bar{D}\hat{G} = \bar{J}$ is equivalent to the following componentwise equations for $k, m \in \{0, ..., N\}$,

$$\begin{array}{lrcll} \text{i)} & \bar{D}_{0l}\,\hat{G}_{lm} & \overset{!}{=} & \delta_{0m} & \\ \text{ii)} & \bar{D}_{1l}\,\hat{G}_{lm} & \overset{!}{=} & \frac{7}{6}\delta_{1m} - \frac{1}{12}\delta_{2m} & \\ \text{iii)} & \bar{D}_{kl}\,\hat{G}_{lm} & \overset{!}{=} & -\frac{1}{12}\delta_{(k-1)m} + \frac{7}{6}\delta_{km} - \frac{1}{12}\delta_{(k+1)m} & \text{for } 2 \le k \le N-2 \\ \text{iv)} & \bar{D}_{(N-1)l}\,\hat{G}_{lm} & \overset{!}{=} & -\frac{1}{12}\delta_{(N-2)m} + \frac{7}{6}\delta_{(N-1)m} & \\ \text{v)} & \bar{D}_{Nl}\,\hat{G}_{lm} & \overset{!}{=} & \delta_{Nm} & \end{array}$$

where we adopt the convention of summation, i.e. summing over the index l from 0 to N.
Remember: $\hat{\mathscr{G}}_{lm} = 0$ whenever $l = 0$ or $l = N$, $\quad 2a + 2b + c = 0$ and $h^{-1}\mathscr{G}_{lm} = \frac{lm}{N} - \min\{l, m\}$.

Check of i): $\bar{D}_{0l}\,\hat{G}_{lm} = \delta_{0l}\left[\frac{N-l}{N}\delta_{0m} + h\hat{\mathscr{G}}_{lm} + \frac{l}{N}\delta_{Nm}\right] = \frac{N-0}{N}\delta_{0m} + h\hat{\mathscr{G}}_{0m} + \frac{0}{N}\delta_{Nm} = \delta_{0m}$

Check of ii): Because of

$$\begin{aligned} h^2\bar{D}_{1l}\left[\tfrac{N-l}{N}\delta_{0m} + \tfrac{l}{N}\delta_{Nm}\right] &= \left[(b+2a)\delta_{0l} + (c-a)\delta_{1l} + b\delta_{2l} + a\delta_{3l}\right]\left[\delta_{0m} - \tfrac{l}{N}\delta_{0m} + \tfrac{l}{N}\delta_{Nm}\right] \\ &= (2a+2b+c)\delta_{0l}\delta_{0m} - (2a+2b+c)\tfrac{1}{N}\delta_{0m} + (2a+2b+c)\tfrac{1}{N}\delta_{Nm} \\ &= 0 \end{aligned}$$

we have only to consider the part of $\hat{G}_{lm}$ which comes from the Green's function $\hat{\mathscr{G}}$:

$$\begin{aligned} \bar{D}_{1l}\hat{G}_{lm} &= h^{-2}\left[(b+2a)\delta_{0l} + (c-a)\delta_{1l} + b\delta_{2l} + a\delta_{3l}\right]\cdot h\hat{\mathscr{G}}_{lm} \\ &= \left[(b+2a)\delta_{0l} + (c-a)\delta_{1l} + b\delta_{2l} + a\delta_{3l}\right]\cdot\left[\tfrac{lm}{N} - \min\{l,m\}\right] \\ &= (c-a)\tfrac{m}{N} + b\tfrac{2m}{N} + a\tfrac{3m}{N} - (c-a)\min\{1,m\} - b\min\{2,m\} - a\min\{3,m\} \\ &= (2a+2b+c)\tfrac{m}{N} - (c-a)\min\{1,m\} - b\min\{2,m\} - a\min\{3,m\} \\ &= -(c+b)\delta_{1m} - (c+2b+a)\delta_{2m} \;=\; (b+2a)\delta_{1m} + a\delta_{2m} \end{aligned}$$

Inserting $a = -\frac{1}{12}$ and $b = \frac{16}{12}$ yields the assertion.

Check of iii): Let us split $\hat{G}_{lm}$ as above:

$$\begin{aligned} h^2\bar{D}_{kl}\left[\tfrac{N-l}{N}\delta_{0m} + \tfrac{l}{N}\delta_{Nm}\right] &= \left[a\delta_{(k-2)l} + b\delta_{(k-1)l} + c\delta_{kl} + b\delta_{(k+1)l} + a\delta_{(k+2)l}\right]\left[\tfrac{N-l}{N}\delta_{0m} + \tfrac{l}{N}\delta_{Nm}\right] \\ &= (2a+2b+c)\tfrac{N-k}{N}\delta_{0m} + (2a+2b+c)\tfrac{k}{N}\delta_{Nm} \\ &= 0 \end{aligned}$$

Hence:

$$\begin{aligned} \bar{D}_{kl}\hat{G}_{lm} &= h^{-2}\left[a\delta_{(k-2)l} + b\delta_{(k-1)l} + c\delta_{kl} + b\delta_{(k+1)l} + a\delta_{(k+2)l}\right]\cdot h\hat{\mathscr{G}}_{lm} \\ &= \left[a\delta_{(k-2)l} + b\delta_{(k-1)l} + c\delta_{kl} + b\delta_{(k+1)l} + a\delta_{(k+2)l}\right]\left[\tfrac{lm}{N} - \min\{l,m\}\right] \\ &= (2a+2b+c)\tfrac{km}{N} - a\min\{k-2,m\} - b\min\{k-1,m\} \\ &\qquad - c\min\{k,m\} - b\min\{k+1,m\} - a\min\{k+2,m\} \\ &= a\delta_{(k-1)m} + (2a+b)\delta_{km} + a\delta_{(k+1)m} \end{aligned}$$

Setting $a = -\frac{1}{12}$ and $2a + b = \frac{7}{6}$ we obtain iii).
We annotate that an evaluation of the min's results in the following case differentiation equivalently expressed by means of the Kronecker-δ's.

$$\left.\begin{array}{c} a\min\{k-2,m\}+b\min\{k-1,m\} \\ +\ c\min\{k,m\}\ + \\ b\min\{k+1,m\}+a\min\{k+2,m\} \end{array}\right\} = \left\{\begin{array}{lll} m \le k-2: & (a+b+c+b+a)m & = \ 0 \\ m = k-1: & a(k-2)+(b+c+b+a)(k-1) & = -a \\ m = k \quad : & a(k-2)+b(k-1)+(c+b+a)k & = -2a-b \\ m = k+1: & a(k-2)+b(k-1)+ck(b+a)(k+1) & = -a \\ m \ge k+2: & (a+b+c+b+a)(k+2) & = \ 0 \end{array}\right.$$

The symmetry properties of $\bar{D}$ and $\hat{G}$ offer an alternative to a direct check of **assertion iv)** and **v)**. Summing the product in iv) in reverse order ($l \to N-l$) yields:

$$\bar{D}_{(N-1)l}\hat{G}_{lm} = \bar{D}_{(N-1)(N-l)}\hat{G}_{(N-l)m}$$

Exploiting the symmetry property (confer remark 7.5)

$$= \bar{D}_{1(N-(N-l))}\hat{G}_{l(N-m)} \;=\; \bar{D}_{1l}\hat{G}_{l(N-m)}$$

using assertion ii) with $m \to N-m$ and observing that generally $\delta_{i(N-j)} = \delta_{(N-i)j}$

$$= \tfrac{7}{6}\delta_{1(N-m)} - \tfrac{1}{12}\delta_{2(N-m)} = \tfrac{7}{6}\delta_{(N-1)m} - \tfrac{1}{12}\delta_{(N-2)m}$$

delivers finally iv). Analogously v) is checked.

As the $\hat{D}_\mu$'s differ from one another only in the row of index 1 and $N-1$ it remains to verify assertion ii) and iv) for each D_μ separately. However the symmetry argument works equally for all $\hat{D}_\mu$'s, such that we can confine ourselves to inspecting only ii). Since the computations are accomplished in a similar way, we present only the step:

Ad II)

$$\begin{aligned}(\hat{D}_{e_2})_{1l}\hat{G}_{lm} &= -(c-3a)\min\{1,m\} - (b+a)\min\{2,m\} - a\min\{3,m\} \\ &= -(c-3a) - 2(b+a) - 3a + \big[(c-3a) - 2(b+a) - 3a\big]\delta_{0m} + \big[(b+a)+2a\big]\delta_{1m} + a\delta_{2m} \\ &= (b+3a)\delta_{1m} + a\delta_{2m} = \tfrac{13}{12}\delta_{1m} - \tfrac{1}{12}\delta_{2m}\end{aligned}$$

Ad III)

$$\begin{aligned}(\hat{D}_{e_3})_{1l}\hat{G}_{lm} &= -(c-6a)\min\{1,m\} - (b+4a)\min\{2,m\} \\ &= \big[-(c-6a) - 2(b+4a)\big] + \big[(c-6a)+2(b+4a)\big]\delta_{0m} + (b+4a)\delta_{1m} \\ &= \delta_{1m}\end{aligned}$$

Ad IV)

$$\begin{aligned}(\hat{D}_{e_4})_{1l}\hat{G}_{lm} &= (10a-c)\min\{1,m\} - (b+10a)\min\{2,m\} + 4a\min\{3,m\} - a\min\{4,m\} \\ &= \big[(10a-c) - 2(b+10a) + 12a - 4a\big] + \big[(10a-c) - 2(b+10a) + 12a - 4a\big]\delta_{0m} \\ &\quad + \big[(b+10a) - 8a + 3a\big]\delta_{1m} + \big[-4a+2a\big]\delta_{2m} + a\delta_{3m} \\ &= \tfrac{11}{12}\delta_{1m} + \tfrac{1}{6}\delta_{2m} - \tfrac{1}{12}\delta_{3m}\end{aligned}$$

■

Note that it is important to multiply with $\hat{G}$ *from the right*. The multiplication from the left leads to more complicated matrices. It is remarkable, that the matrix $\hat{D}_{e_0}$ of the nearest-neighbor extrapolation scheme exposes a different bahavior. Although $\hat{D}_{e_0} \cdot \hat{G}$ is in the lines of index $2 \le i \le N-2$ also equal to the tridiagonalmatrix $\mathrm{tridiag}\big(-\frac{1}{12}, \frac{7}{6}, -\frac{1}{12}\big)$ the two lines of index $i = 1$ and $i = N-1$ are full. It turns out that the product-matrix is neither diagonal-dominant nor even bounded. Therefore

$\hat{D}_{e_0}$ will undergo an individual treatment in lemma 7.7.

Proposition 7.3. *Let* $A = (a_{ij})_{ij} \in \mathbb{C}^{d\times d}$ *be a* strictly diagonal-dominant *matrix, i.e.*

$$\alpha := \min_i \left\{ |a_{ii}| - \sum_{j\neq i} |a_{ij}| \right\} > 0\,. \qquad \Rightarrow \qquad \begin{cases} i) & \forall x \in \mathbb{C}^d\colon\ \|Ax\|_\infty \geq \alpha \|x\|_\infty \\ ii) & A \text{ is invertible} \\ iii) & \|A^{-1}\|_\infty \leq \frac{1}{\alpha} \end{cases}$$

Proof:

ad i) Without loss of generality let $\|x\|_\infty = 1$ (otherwise consider $x/\|x\|$). Then $|x_j| \leq 1$ for all $j \in \{1, ..., d\}$ and there is at least one index i with $|x_i| = 1$.

$$\begin{aligned} \|Ax\|_\infty \;&\geq\; |(Ax)_i| \;=\; \big|\sum_j a_{ij} x_j\big| \;\geq\; |a_{ii}||x_i| - \big|\sum_{j\neq i} a_{ij}x_j\big| \\ &\geq\; |a_{ii}| - \sum_{j\neq i} |a_{ij}||x_j| \;\geq\; |a_{ii}| - \sum_{j\neq i}|a_{ij}| \;\geq\; \alpha \end{aligned}$$

ad ii) Due to i) the kernel (nullspace) of A must be trivial.

ad iii) Setting in ii) $x = A^{-1}z$ for some $z \in \mathbb{C}^d$, we obtain:

$$\|z\|_\infty = \|AA^{-1}z\|_\infty \geq \alpha\|A^{-1}z\|_\infty \qquad \Rightarrow \qquad \|A^{-1}z\|_\infty \leq \tfrac{1}{\alpha}\|z\|_\infty .$$

As the inequality holds true, while z ranges over the whole $\mathbb{C}^d$, the assertion follows.

■

Theorem 7.7. ***(Stability of the higher order extrapolation methods)*** *The matrices* $\hat{D}_\mu$ *with* $\mu \in \{e_1, e_2, e_3, e_4\}$ *defined in (7.19) are invertible. Moreover the following estimates hold* independently *of the grid spacing* $h = \frac{L}{N}$ *or the number of nodes respectively:*

$$\begin{array}{lllll} I) & \|\hat{D}_{e_1}^{-1}\|_\infty & \leq & 1 + 2L^2 & \textit{(5-point stencil, linear extrapol.)} \\ II) & \|\hat{D}_{e_2}^{-1}\|_\infty & \leq & 1 + 2L^2 & \textit{(5-point stencil, qudrat. extrapol.)} \\ III) & \|\hat{D}_{e_3}^{-1}\|_\infty & \leq & 1 + 2L^2 & \textit{(5-point stencil, cubic extrapol.)} \\ IV) & \|\hat{D}_{e_4}^{-1}\|_\infty & \leq & \frac{3}{2}(1 + 2L^2) & \textit{(5-point stencil, quartic extrapol.)} \end{array}$$

Proof: Since the matrices J_μ in lemma 7.6 are diagonal-dominant, they are invertible according to proposition 7.3 ii). Hence, it can be inferred from the lemma 7.6, that the $\hat{D}_\mu$'s are invertible too, with

$$\hat{D}_\mu^{-1} = \hat{G} \cdot J_\mu^{-1}\,.$$

Now, $\|\cdot\|_\infty$ is an operator matrix norm, since it is induced by the maximum norm for vectors. So it is *submultiplicative* yielding with lemma 7.5 the inequality

$$\|\hat{D}_\mu^{-1}\|_\infty \;\leq\; \|\hat{G}\|_\infty\, \|\hat{J}_\mu^{-1}\|_\infty \;\;\leq\; (1 + 2L^2)\|\hat{J}_\mu^{-1}\|_\infty .$$

In order to estimate $\|\hat{J}_\mu^{-1}\|_\infty$, proposition 7.3 iii) can be used once more, such that the claimed estimates ensue directly. ■

Let us now tackle the stability of the nearest-neighbor extrapolation scheme.

Lemma 7.7. *Let $\hat{D}_{e_0}$ be given as in (7.19) with $\mu = e_0$. The multiplication of $\hat{D}_{e_0}$ from the left with $\hat{G}$ yields the following matrix:*

$$\hat{G} \cdot \hat{D}_{e_0} =: K_{e_0} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ * & * & -\frac{1}{12} & 0 & 0 & 0 & 0 & \dots & 0 & * & * \\ * & * & \frac{7}{6} & -\frac{1}{12} & 0 & 0 & 0 & \dots & 0 & * & * \\ * & * & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 & 0 & \dots & 0 & * & * \\ * & * & 0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & 0 & \dots & 0 & * & * \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ * & * & 0 & \dots & 0 & -\frac{1}{12} & -\frac{7}{6} & -\frac{1}{12} & 0 & * & * \\ * & * & 0 & \dots & 0 & 0 & -\frac{1}{12} & \frac{7}{6} & -\frac{1}{12} & * & * \\ * & * & 0 & \dots & 0 & 0 & 0 & -\frac{1}{12} & \frac{7}{6} & * & * \\ * & * & 0 & \dots & 0 & 0 & 0 & 0 & -\frac{1}{12} & * & * \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The asterisks indicate non-zero elements given in the two leftmost columns by

$$(K_{e_0})_{10} = -\tfrac{1}{6} + \tfrac{1}{12N}, \qquad (K_{e_0})_{k0} = -\tfrac{1}{12}\tfrac{N-k}{N} \quad \text{for } k \geq 2,$$
$$(K_{e_0})_{11} = \tfrac{15}{12} - \tfrac{1}{12N}, \quad (K_{e_0})_{21} = -\tfrac{1}{6N}, \quad (K_{e_0})_{k1} = \tfrac{1}{12}\tfrac{N-k}{N} \quad \text{for } k \geq 3.$$

Furthermore the elements in the two rightmost columns are determined by the symmetry property $(K_{e_0})_{km} = (K_{e_0})_{(N-k)(N-m)}$ holding for $k, m \in \{0, ..., N\}$.

Proof: In order to facilitate the writing of indices let us introduce the shortcuts $\bar{D} := \hat{D}_{e_0}$ and $\bar{K} := K_{e_0}$. The verification is divided into several steps. We start by computing individually the three leftmost columns of $\bar{K}$ (of indices $m = 0, 1, 2$). Since the subsequent columns of $\bar{D}$ (up to the $(N-3)$'th one) contain the full five-point stencil, we can treat together the corresponding columns of $\bar{K}$ in a single *coup*. The proof is finished by checking the symmetry property.

Remember that $h^{-1}\mathscr{G}_{ij} = \frac{ij}{N} - \min\{i, j\}$ (cf. (7.22)) and $2a + 2b + c = 0$. Furthermore recall from (7.19) and (7.18) the representation of the columns of $\bar{D}$. As in the previous proofs we use the summation convention with respect to the index l (to be summed from 0 to N).

Column $m = 0$:

$$\begin{aligned} \bar{K}_{k0} &= \hat{G}_{kl}\bar{D}_{l0} = \left(\tfrac{N-k}{N}\delta_{l0} + h\hat{\mathscr{G}}_{kl} + \tfrac{k}{N}\delta_{lN}\right) \cdot \left(\delta_{l0} + h^{-2}(b+a)\delta_{l1} + h^{-2}a\delta_{l2}\right) \\ &= \tfrac{N-k}{N} + \left((b+a)h^{-1}\hat{\mathscr{G}}_{k1} + ah^{-1}\hat{\mathscr{G}}_{k2}\right) \\ &= \tfrac{N-k}{N} + (b+a)\tfrac{k}{N} - (b+a)\min\{k, 1\} + a\tfrac{2k}{N} - a\min\{k, 2\} \\ &= 1 + (b+3a-1)\tfrac{k}{N} - (b+3a) + (b+3a)\delta_{k0} + a\delta_{k1} \\ &= -\tfrac{1}{12}\tfrac{N-k}{N} + \tfrac{13}{12}\delta_{k0} - \tfrac{1}{12}\delta_{k1} \end{aligned}$$

Column $m = 1$:

$$\begin{aligned} \bar{K}_{k1} &= \hat{G}_{kl}\bar{D}_{l1} = \left(\tfrac{N-k}{N}\delta_{l0} + h\hat{\mathscr{G}}_{kl} + \tfrac{k}{N}\delta_{lN}\right)h^{-2}\left(c\delta_{l1} + b\delta_{l1} + a\delta_{l2}\right) \\ &= ch^{-1}\hat{\mathscr{G}}_{k1} + bh^{-1}\hat{\mathscr{G}}_{k2} + ah^{-1}\hat{\mathscr{G}}_{k3} \\ &= (c+2b+3a)\tfrac{k}{N} - c\min\{1, k\} - b\min\{2, k\} - a\min\{3, k\} \\ &= -\tfrac{1}{12}\tfrac{k}{N} - (c+2b+3a) + (c+2b+3a)\delta_{k0} + (b+2a)\delta_{k1} + a\delta_{k2} \\ &= \tfrac{1}{12}\tfrac{N-k}{N} - \tfrac{1}{12}\delta_{k0} + \tfrac{7}{6}\delta_{k1} - \tfrac{1}{12}\delta_{k2} \end{aligned}$$

Column $m = 2$**:**

$$\begin{aligned}
\bar{K}_{k2} &= \hat{G}_{kl}\bar{D}_{l2} = \left(\tfrac{N-k}{N}\delta_{0l} + h\hat{\mathscr{G}}_{kl} + \tfrac{k}{N}\delta_{Nl}\right)\cdot h^{-2}\left(b\delta_{l1} + c\delta_{l2} + b\delta_{l3} + a\delta_{l4}\right)\\
&= h^{-1}\hat{\mathscr{G}}_{kl}\left(b\delta_{l1} + c\delta_{l2} + b\delta_{l3} + a\delta_{l4}\right)\\
&= (b + 2c + 3b + 4a)\tfrac{k}{N} - b\min\{1,k\} - c\min\{2,k\} - b\min\{3,k\} - a\min\{4,k\}\\
&= -(b + 2c + 3b + 4a) + (b + 2c + 3b + 4a)\delta_{k0} + (c + 2b + 3a)\delta_{k1} + (b + 2a)\delta_{k2} + a\delta_{k3}\\
&= -\tfrac{1}{12}\delta_{k1} + \tfrac{7}{6}\delta_{k2} - \tfrac{1}{12}\delta_{k3}
\end{aligned}$$

Columns $3 \le m \le N-3$**:**

$$\begin{aligned}
\bar{K}_{km} &= \hat{G}_{kl}\bar{D}_{lm} = \left(\tfrac{N-k}{N}\delta_{0l} + h\hat{\mathscr{G}}_{kl} + \tfrac{k}{N}\delta_{Nl}\right)\cdot h^{-2}\left(a\delta_{l(m-2)} + b\delta_{l(m-1)} + c\delta_{lm} + b\delta_{l(m+1)} + a\delta_{l(m+2)}\right)\\
&= h^{-1}\hat{\mathscr{G}}_{kl}\left(a\delta_{l(m-2)} + b\delta_{l(m-1)} + c\delta_{lm} + b\delta_{l(m+1)} + a\delta_{l(m+2)}\right)\\
&= \left[a(m-2) + b(m-1) + cm + b(m+1) + a(m+2)\right]\tfrac{k}{N}\\
&\quad - a\min\{k,m-2\} - b\min\{k,m-1\} - c\min\{k,m\} - b\min\{k,m+1\} - a\min\{k,m+2\}\\
&= m(2a + 2b + c) - \left[a(m-2) + b(m-1) + cm + b(m+1) + a(m+2)\right]\\
&\quad + [b + 2c + 3b + 4a]\delta_{k(m-2)} + [c + 2b + 3a]\delta_{k(m-1)} + [b + 2a]\delta_{km} + \delta_{k(m+1)}\\
&= a\delta_{k(m-1)} + (b + 2a)\delta_{km} + a\tfrac{1}{12}\delta_{k(m+1)} = -\tfrac{1}{12}\delta_{k(m-1)} + \tfrac{7}{6}\delta_{km} - \tfrac{1}{12}\delta_{k(m+1)}
\end{aligned}$$

Symmetry property: Using the symmetry properties of $\hat{G}$ and $\bar{K}$ mentioned in remark 7.5 and "rebaptizing" the summation index from $N-l$ to l, we obtain:

$$\bar{K}_{(N-k)(N-m)} := \hat{G}_{(N-k)l}\bar{D}_{l(N-m)} = \hat{G}_{k(N-l)}\bar{D}_{(N-l)m} = \hat{G}_{kl}\bar{D}_{lm} =: \bar{K}_{km}\ . \qquad \blacksquare$$

Lemma 7.8. *The matrix* $K_{e_0} = \hat{G}\cdot\hat{D}_{e_0}$ *is* strictly diagonal-dominant *and satisfies*

$$\|K_{e_0}^{-1}\|_\infty < \tfrac{6}{5} \qquad \textit{and} \qquad \|K_{e_0} - I\|_\infty = \|\hat{G}\cdot(\hat{D}_{e_0} - \hat{D})\|_\infty \;=\; \tfrac{1}{2}.$$

Proof: From lemma 7.7 we can read off the elementwise representation of the matrix K_{e_0} with $m \in \{0, ..., N\}$ and $k \in \{1, ..., N-1\}$:

$$\begin{aligned}
(K_{e_0})_{0m} &= \delta_{0m}\\
(K_{e_0})_{km} &= -\tfrac{1}{12}\tfrac{N-k}{N}\delta_{0m} + \tfrac{1}{12}\tfrac{N-k}{N}\delta_{1m} - \tfrac{1}{12}\delta_{(k-1)m} + \tfrac{7}{6}\delta_{km} - \tfrac{1}{12}\delta_{(k+1)m} + \tfrac{1}{12}\tfrac{k}{N}\delta_{0m} - \tfrac{1}{12}\tfrac{k}{N}\delta_{1m}\\
(K_{e_0})_{Nm} &= \delta_{Nm}
\end{aligned}$$

We observe that all matrix elements are *non-negative.* Since $\frac{7}{6} > \frac{4}{12}$ the matrix is strictly diagonal-dominant and proposition 7.3 iii) ensures $\|K_{e_0}^{-1}\| < (\frac{7}{6} - \frac{4}{12})^{-1} = \frac{6}{5}$. Furthermore we get
$\|K_{e_0} - I\|_\infty = \max_{k=1}^{N-1}\left[\frac{1}{12}\frac{N-k}{N} + \frac{1}{12}\frac{N-k}{N} + \frac{1}{12} + (\frac{7}{6} - 1) + \frac{1}{12} + \frac{1}{12}\frac{k}{N} + \frac{1}{12}\frac{k}{N}\right] = \frac{4}{12} + \frac{1}{6} = \frac{1}{2}$. $\blacksquare$

Theorem 7.8. ***(Stability of the nearest-neighbor extrapolation method)*** *The system matrix $\hat{D}_{e_0}$ of the nearest-neighbor extrapolation scheme is invertible. Moreover we have independently of the number of grid nodes*

$$\hat{D}_{e_0}^{-1} = K_{e_0}^{-1}\hat{G} \qquad \textit{and} \qquad \|\hat{D}_{e_0}^{-1}\|_\infty < \tfrac{6}{5}(1 + 2L^2).$$

Proof: Analogously to theorem 7.7 the assertion follows directly from lemma 7.5 and lemma 7.8. $\blacksquare$

Remark 7.6. There is another possibility to prove theorem 7.8, however, with a slightly worse (larger) constant in the estimate. The idea is based on a Neumann sequence. We present it briefly, since it might be helpful to have it in mind, when the argument using diagonal-dominance fails (see for example comments on corollary 7.2).
Obviously we have

$$\hat{D}_{e_0} = \hat{D} + \hat{D}_{e_0} - \hat{D} = \hat{D}\cdot\left(I + \hat{D}^{-1}(\hat{D}_{e_0} - \hat{D})\right) = \hat{D}\cdot\left(I + \hat{G}(\hat{D}_{e_0} - \hat{D})\right) = \hat{D}\cdot(I + P)$$

where we introduce the abbreviation $P := \hat{G}(\hat{D}_{e_0} - \hat{D})$. Lemma 7.8 yields $\|P\|_\infty = \frac{1}{2}$. Therefore the Neumann series of P converges and is equal to the inverse of $I + P$:

$$(I+P)^{-1} = \sum_{k\geq 0}(-1)^k P^k \quad \Rightarrow \quad \|(I+P)^{-1}\|_\infty \;\leq\; \sum_{k\geq 0}\|P\|_\infty^k \;=\; \tfrac{1}{1-\|P\|_\infty} \;=\; \tfrac{1}{1-\frac{1}{2}} \;=\; 2$$

From $\hat{D}_{e_0} = \hat{D}\cdot(I+P)$ we can conclude $\hat{D}_{e_0}^{-1} = (I+P)^{-1}\cdot\hat{D}^{-1} = (I+P)^{-1}\cdot\hat{G}$
The submultiplicativity of the norm $\|\cdot\|_\infty$ and lemma 7.5 lead eventually to the estimate:

$$\|D_{e_0}^{-1}\|_\infty \;\leq\; \|(I+P)^{-1}\|_\infty\,\|\hat{G}\|_\infty \;\leq\; 2(1+2L^2) = 2+4L^2\ .$$

7.5 Damping property of discrete inverse operators

The Lipschitz-continuity of the Green's operator G guarantees that the solution of the Poisson problem is bounded, whenever the input data – the source and the boundary values – are bounded too. This property is transmitted to the numeric schemes by the stability estimates, which provide an upper bound for the norm of the discrete inverse operators under grid refinement. Such stability estimates are expected to hold for any reasonable discretization of a well-posed differential equation problem.
In order to prove the convergence of the five-point stencil extrapolation schemes we need to consider also (point) sources that become singular under grid refinement. These sources are of course not bounded in the $\|\cdot\|_\infty$-norm; that is why the stability argument fails to infer the boundedness of the corresponding solutions. A more thorough knowledge of the discrete inverse operators is therefore necessary. Typically, estimates of their matrix elements are required.

The problem sketched so far for the discrete case has a continuous analogon, that we want to present at first to settle the idea. Let us consider the Poisson problem with homogeneous boundary conditions and a "point source" represented by *Dirac's δ-function*[15]. How does the solution behave, when the position of the point source is shifted towards the left boundary point 0, while its strength is increased reciprocally to its distance ϵ from 0?

$$u(0) = u(L) = 0, \qquad \partial^2 u(x) = \tfrac{1}{\epsilon}\delta(x-\epsilon) \quad x \in (0,L)$$

This problem is interesting because of two antagonistic conditions: on the one hand the boundary condition tries to keep the (continuous) solution close to zero in the vicinity of the boundary, on the other hand the singular source tends to infinity.
The solution is just $1/\epsilon$ times the Green's function, that is constructed by the *fundamental solution*[16] plus a regular part adjusting the boundary conditions. A

[15]Correctly, one should speak of the *δ-distribution*. It can be regularized by a sequence of functions, that vanish outside a shrinking vicinity of a given center point, such that their integral over $\mathbb{R}$ remains constant one. A simple possibility to realize such a sequence parametrized by $h \downarrow 0$ is given by rectangular functions of height h^{-1}, whose support is an interval of length h around the center. This sequence is well suited to discretize the δ-function on a grid of mesh-width h.

[16]The fundamental solution $\mathscr{F} = \mathscr{F}(x,\xi)$ pertaining to a differential operator Δ satisfies $\Delta_x\mathscr{F}(x,\xi) = \delta(x-\xi)$. The Green's function $\mathscr{G}$ refers to the differential operator and a prescribed choice of homogeneous boundary conditions. We have $\mathscr{G}(x,\xi) = \mathscr{F}(x,\xi) + \mathscr{U}(x,\xi)$ with $\Delta_x\mathscr{U}(x,\xi) = 0$, such that $\mathscr{G}$ satisfies the boundary conditions with respect to x for any ξ.

formal computation with the δ-function yields:

$$u(x) \;=\; \tfrac{1}{\epsilon}\int_0^L \mathscr{G}(x,\xi)\delta(\xi-\epsilon)\,d\xi \;=\; \tfrac{1}{\epsilon}\mathscr{G}(x,\epsilon) \;=\; \tfrac{1}{\epsilon}\tfrac{x\epsilon}{L} - \tfrac{1}{\epsilon}\min\{x,\epsilon\} \tag{7.23}$$

$$\Rightarrow |u(x)| \;=\; \begin{cases} \left|\tfrac{x}{L}-\tfrac{x}{\epsilon}\right| \;=\; \tfrac{x}{\epsilon}-\tfrac{x}{L} \;<\; 1 & \text{for } 0\le x\le\epsilon \\ \left|\tfrac{x}{L}-1\right| \;=\; 1-\tfrac{x}{L} \;<\; 1 & \text{for } \epsilon\le x\le L \end{cases}$$

Indeed the solution remains bounded. The crucial point is, that the singular factor $1/\epsilon$ tends as fast to infinity as the Greens function vanishes in approaching the boundary. Thus, the question, which inverse powers of ϵ still generate bounded solutions, is therefore related to the decay rate of the Green's function near the boundary.

Let us return to the discrete case. The discretized version of equation (7.23) could be written in the form

$$\hat{D}_\mu v = \tfrac{1}{h^2}\delta_1 \quad \text{with } \delta_i \in \mathcal{F}(\mathcal{I}_h), \quad \delta_i(x_j) := \delta_{ij} \text{ for } 0\le i,j\le N.$$

Here the grid spacing $h = \frac{L}{N}$ has adopted the role of the parameter ϵ. Furthermore the discrete counterpart of $\delta(x-h)$ is $\frac{1}{h}\delta_1$. After the consideration of the continuous case one is strongly inclined to suppose that the associated solutions remain uniformly bounded for $h\downarrow 0$. Before we come to the proof of this result our observation prompts a clear definition, characterizing the right hand side.

Definition 7.2. *For $h\downarrow 0$ let $g\in\mathcal{F}(\mathcal{I}_h)$ be a (sequence of) grid function(s) satisfying the following conditions:*

i) $\|g\|_\infty = \mathrm{O}(\frac{1}{h^2})$

ii) $\mathrm{dist}(\mathrm{supp}(g),\partial\mathcal{I}) = \mathrm{O}(h), \quad \partial\mathcal{I}=\{0,L\}$

iii) $\#\mathrm{supp}(g) = \mathrm{O}(1)$

The discrete operator $A:\mathcal{F}(\mathcal{I}_h)\to\mathcal{F}(\mathcal{I}_h)$ is said to have the damping property *if $\|Ag\|_\infty = \mathrm{O}(1)$ under grid refinement $h\downarrow 0$.*

The term "damping property" tries to express figuratively, that the singularity of the source is "damped" by the application of A, as ironing flattens a crease. If we want to waive condition ii) and permit arbitrarily located singular sources, the singularity must be mildened. This means that the norm of g may only be of the order $\mathrm{O}(\frac{1}{h})$, what corresponds to the δ-function without any singular factor.

Theorem 7.9. ***(Damping property of $\hat{G}$)*** *The sampled Green's operator, which is the inverse operator to the standard scheme, possesses the damping property.*

Proof: A grid function g fulfilling the conditions of the previous definition is written as a finite linear combination of $\frac{1}{h^2}\delta^m$, where the index m varies in $\{0,1,...,M\}\cup\{N-M,...,N\}$ and M is a fixed integer while $h\downarrow 0$ and $N\to\infty$. By linearity it is sufficient to consider the case $g=\frac{1}{h^2}\delta^m$. We have to show that $\|\hat{G}g\|_\infty$ is bounded independently of h.

$$\begin{aligned}
\|\tfrac{1}{h^2}\hat{G}\delta^m\|_\infty &= \max_{k=1}^{N} \big|\tfrac{1}{h^2}\sum_{l=0}^{N} \hat{G}_{kl}\delta_{ml}\big| \\
&= \max_{k=1}^{N} \big|\tfrac{1}{h^2}\sum_{l=0}^{N} \big(\tfrac{N-k}{N}\delta_{l0} + h\hat{\mathscr{G}}_{kl} + \tfrac{k}{N}\delta_{lN}\big)\delta_{ml}\big| \\
&= \max_{k=1}^{N} \big|\tfrac{1}{h}\sum_{l=0}^{N} \hat{\mathscr{G}}_{kl}\delta_{ml}\big| \\
&= \max_{k=1}^{N} \big|\tfrac{1}{h}\sum_{l=0}^{N} \big(klh^2L^{-1} - h\min\{k,l\}\big)\delta_{ml}\big| \\
&= \max_{k=1}^{N} \big|\tfrac{k}{N}m - \min\{k,m\}\big| \\
&= \max_{k=1}^{N} \big\{\min\{k,m\} - \tfrac{k}{N}m\big\} = m - \tfrac{m}{N}m \\
&\leq m
\end{aligned}$$

Note that $0 \leq k, m \leq N$, therefore $\frac{k}{N}, \frac{m}{N} < 1$. By reasons of symmetry we expect this to hold also for the right boundary at L.

$$\begin{aligned}
\|\tfrac{1}{h^2}\hat{G}\delta^{N-m}\|_\infty &= \max_{k=1}^{N} \big\{\min\{k, N-m\} - \tfrac{k}{N}(N-m)\big\} \\
&= \begin{cases} k \leq N-m: & k - k(1-\tfrac{m}{N}) \;=\; \tfrac{k}{N}m \\ k > N-m: & (1-\tfrac{k}{N})(N-m) \;=\; \tfrac{N-k}{N}(N-m) \end{cases} \\
&\leq \begin{cases} k \leq N-m: & m \\ k > N-m: & \tfrac{m}{N}(N-m) = m - \tfrac{m}{N}m \leq m \end{cases} \\
&\leq m
\end{aligned}$$

■

Remark 7.7. This result can be used to prove the damping property for a whole class of operators. Suppose that $A \in \mathcal{F}(\mathcal{I}_h) \to \mathcal{F}(\mathcal{I}_h)$ factorizes in the form $A = B \cdot \hat{G}$ with $B \in \mathcal{F}(\mathcal{I}_h) \to \mathcal{F}(\mathcal{I}_h)$ uniformly bounded under grid refinement and $\hat{G}$ as the *right* factor. Then:

$$\|h^{-2}A\delta_m\|_\infty \;\leq\; \|B\|_\infty \|h^{-2}\hat{G}\delta_m\|_\infty \;=\; \mathrm{O}(1)\cdot\mathrm{O}(1) \;=\; \mathrm{O}(1)\,.$$

Corollary 7.1. *The inverse operator $\hat{D}_{e_0}^{-1}$ of the nearest-neighbor extrapolation scheme has the damping property.*

Proof: In theorem 7.8 we have established the decomposition $\hat{D}_{e_0}^{-1} = K_{e_0}^{-1}\hat{G}$. Therefore the assertion follows directly from remark 7.7. ■

Corollary 7.2. *The inverse operators associated with the* linear, quadratic, cubic *and* quartic extrapolation schemes *defined in (7.17) and (7.19) dispose of the damping property.*

Instead of hurling ourselves immediately into tedious computations to prove this result, we content ourselves with some comments on it.

1) The factors in $\hat{D}_{e_1} = J_{e_1}\hat{D}$ (see lemma 7.6) or equivalently in $\hat{D}_{e_1}^{-1} = \hat{G}J_{e_1}^{-1}$ appear just in the wrong order to apply remark 7.7. So we consider $K_{e_1} := \hat{G}\hat{D}_{e_1}$ being also a tridiagonal matrix equal to J_{e_1} with the exception of the non-vanishing elements $(K_{e_1})_{10} = (K_{e_1})_{(N-1)N} = -\frac{1}{12}$. Using K_{e_1} to show stability, the damping property comes for free.

2) The situation is similar for the quadratic extrapolation scheme. However, $K_{e_2} := \hat{G}\hat{D}_{e_2}$ is no more a band-matrix, but its three left- and rightmost columns are full similar to K_{e_1} in lemma 7.7. Nevertheless, the matrix is still strictly diagonal-dominant, such that it can be handled analogously.

3) The situation becomes more complicated for $K_{e_3} := \hat{G}\hat{D}_{e_1}$ and $K_{e_4} := \hat{G}\hat{D}_{e_1}$, since the number of full columns on each side increases to four and five respectively. Moreover the matrices lose their strict diagonal-dominance[17] what requires more effort to invert them. In fact, the *Neumann series* (similar to remark 7.6) offers a possibility, but there remains an additional cliff to circumnavigate. The full work is done for the quartic extrapolation scheme in proof of lemma 7.9.

Lemma 7.9. *The system matrix $\hat{D}_{e_4}$ of the the quartic extrapolation scheme allows the factorization $\hat{D}_{e_4} = \hat{D}K_{e_4}$, where the matrix K_{e_4} is invertible and $\|K_{e_4}^{-1}\|_\infty < 12$ independently of the grid spacing.*

Proof: It is to show that the product matrix $\hat{G}\hat{D}_{e_4} =: K_{(e_4)}$ is invertible satisfying moreover the asserted estimate. In order to compute the product we introduce two $(N+1)\times(N+1)$ matrices:

$$\tilde{D} := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 1 & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -2 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 & -2 & 1 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$M := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0\dots0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4}\frac{N-1}{N} & \frac{3}{4}\frac{1-N}{N} & \frac{5}{6}\frac{N-1}{N} & \frac{5}{12}\frac{1-N}{N} & \frac{1}{12}\frac{N-1}{N} & 0\dots0 & \frac{1}{12}\frac{1}{N} & \frac{5}{12}\frac{-1}{N} & \frac{5}{6}\frac{1}{N} & \frac{3}{4}\frac{-1}{N} & \frac{1}{4}\frac{1}{N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\dots\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{4}\frac{1}{N} & \frac{3}{4}\frac{-1}{N} & \frac{5}{6}\frac{1}{N} & \frac{5}{12}\frac{-1}{N} & \frac{1}{12}\frac{1}{N} & 0\dots0 & \frac{1}{12}\frac{N-1}{N} & \frac{5}{12}\frac{1-N}{N} & \frac{5}{6}\frac{N-1}{N} & \frac{3}{4}\frac{1-N}{N} & \frac{1}{4}\frac{N-1}{N} \\ 0 & 0 & 0 & 0 & 0 & 0\dots0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The vertical dots in the non-zero columns indicate that the matrix elements decrease or increase linearly with their row indices, e.g. $M_{k0} = \frac{1}{4}\frac{N-k}{N}$ and $M_{kN-4} = \frac{1}{12}\frac{k}{N}$ for $1 \le k \le N-1$. At the end of a longish computation we find

$$K_{e_4} = \hat{G}\hat{D}_{e_4} = M + (I - \tfrac{1}{12}\tilde{D}) = I + (M - \tfrac{1}{12}\tilde{D}).$$

We are interested in the inverse of this matrix. As it is not diagonal-dominant, we try to invert it by computing the *Neumann series* for $M - \frac{1}{12}\tilde{D}$, which is known to be convergent iff the spectral radius is less than one. A sufficient criterion consists in checking a matrix norm. Unfortunately we have

$$\|M - \tfrac{1}{12}\tilde{D}\|_\infty = \tfrac{8}{3} > 1$$

[17] At least, $\hat{G}\hat{D}_{e_3}$ is diagonal-dominant but *not* strictly diagonal dominant.

Since it is still sufficient if the norm of some power of $M - \frac{1}{12}\tilde{D}$ is less than one, we estimate $\|(M - \frac{1}{12}\tilde{D})^2\|_\infty$. For this we expand the square and compute the addends. After a good deal of arithmetics we find:

$$M^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \dots 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{16}\frac{1-N}{N} & \frac{3}{16}\frac{N-1}{N} & \frac{5}{24}\frac{1-N}{N} & \frac{5}{48}\frac{N-1}{N} & \frac{1}{48}\frac{1-N}{N} & 0 \dots 0 & \frac{1}{48}\frac{-1}{N} & \frac{5}{48}\frac{1}{N} & \frac{5}{24}\frac{-1}{N} & \frac{3}{16}\frac{1}{N} & \frac{1}{16}\frac{-1}{N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \dots \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{16}\frac{-1}{N} & \frac{3}{16}\frac{1}{N} & \frac{5}{24}\frac{-1}{N} & \frac{5}{48}\frac{1}{N} & \frac{1}{48}\frac{-1}{N} & 0 \dots 0 & \frac{1}{48}\frac{1-N}{N} & \frac{5}{48}\frac{N-1}{N} & \frac{5}{24}\frac{1-N}{N} & \frac{3}{16}\frac{N-1}{N} & \frac{1}{16}\frac{1-N}{N} \\ 0 & 0 & 0 & 0 & 0 & 0 \dots 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$M\tilde{D} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \dots 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{4}\frac{1-N}{N} & \frac{7}{3}\frac{N-1}{N} & \frac{17}{6}\frac{1-N}{N} & \frac{7}{4}\frac{N-1}{N} & \frac{7}{12}\frac{1-N}{N} & \frac{1}{12}\frac{N-1}{N} & 0 \dots 0 & \frac{1}{12}\frac{1}{N} & \frac{7}{12}\frac{-1}{N} & \frac{7}{4}\frac{1}{N} & \frac{17}{6}\frac{-1}{N} & \frac{7}{3}\frac{1}{N} & \frac{3}{4}\frac{-1}{N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \dots \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{3}{4}\frac{-1}{N} & \frac{7}{3}\frac{1}{N} & \frac{17}{6}\frac{-1}{N} & \frac{7}{4}\frac{1}{N} & \frac{7}{12}\frac{-1}{N} & \frac{1}{12}\frac{1}{N} & 0 \dots 0 & \frac{1}{12}\frac{N-1}{N} & \frac{7}{12}\frac{1-N}{N} & \frac{7}{4}\frac{N-1}{N} & \frac{17}{6}\frac{1-N}{N} & \frac{7}{3}\frac{N-1}{N} & \frac{3}{4}\frac{1-N}{N} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \dots 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$\tilde{D}M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ -\frac{1}{4} & \frac{3}{4} & -\frac{5}{6} & \frac{5}{12} & -\frac{1}{12} & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ & \dots & 0 & -\frac{1}{12} & \frac{5}{12} & -\frac{5}{6} & \frac{3}{4} & -\frac{1}{4} \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Observe that $\tilde{D}M$ is almost zero. This is due to the fact that $\tilde{D}$ is essentially a finite difference matrix. Therefore $\tilde{D}M$ gives, so to say, the second derivative of the column vectors of M. Since these depend linearly on the row index, the product matrix elements are zero except at those positions, where the matrix elements of M jump in the top- and bottommost line from zero to some other value. In contrast M^2 and $\tilde{D}M$ exhibit nearly the same structure. This facilitates estimating the norm of their sum, which is much smaller than the sum of their norms:

$$\begin{aligned} \|M^2 - \tfrac{1}{12}M\tilde{D}\|_\infty = \max_{k=1}^{N-1} \Big(&\big| -\tfrac{1}{16} + \tfrac{1}{12}\tfrac{3}{4}\big| + \big|\tfrac{3}{16} - \tfrac{1}{12}\tfrac{7}{3}\big| + \big| -\tfrac{5}{24} + \tfrac{1}{12}\tfrac{17}{6}\big| + \\ &+ \big|\tfrac{5}{48} - \tfrac{1}{12}\tfrac{7}{4}\big| + \big| -\tfrac{1}{48} + \tfrac{1}{12}\tfrac{7}{12}\big| + \big|0 - \tfrac{1}{12}\tfrac{1}{12}\big|\Big) \cdot \Big(\tfrac{N-k}{N} + \tfrac{k}{N}\Big) = \tfrac{1}{9}. \end{aligned}$$

Employing the triangle inequality we get:

$$\begin{aligned} \|(M^2 - \tfrac{1}{12}\tilde{D})^2\|_\infty &= \|M^2 - \tfrac{1}{12}M\tilde{D} - \tfrac{1}{12}\tilde{D}M + \tfrac{1}{144}\tilde{D}^2\|_\infty \\ &\leq \|M^2 - \tfrac{1}{12}M\tilde{D}\|_\infty + \tfrac{1}{12}\|\tilde{D}M\|_\infty + \tfrac{1}{144}\|\tilde{D}\|_\infty^2 \end{aligned}$$

The estimate is completed by reading off the norms of $\tilde{D}M$ and $\tilde{D}$

$$\|(M^2 - \tfrac{1}{12}\tilde{D})^2\|_\infty \leq \tfrac{1}{9} + \tfrac{1}{12}\tfrac{7}{3} + \tfrac{4^2}{144} = \tfrac{15}{36} < 1$$

which brings the desired result. In order to show that $K_{e_4}^{-1}$ exists, we have to verify the convergence of the Neumann series. Additionally, we will obtain an upper bound for $\|K_{e_4}^{-1}\|_\infty$.

$$K_{e_4}^{-1} \overset{!}{=} 1 + \sum_{n=1}^{\infty}(\hat{G}\hat{D}_{e_4} - I)^n = 1 + \sum_{n=1}^{\infty}(M - \tfrac{1}{12}\tilde{D})^n$$

Taking the norm on both sides and applying the triangle inequality to get the norm inside the sum yields:

$$\begin{aligned}
\|K_{e_4}^{-1}\|_\infty &\le 1+\sum_{n=1}^{\infty}\left\|(M-\tfrac{1}{12}\tilde{D})^n\right\|_\infty \\
&\le 1+\|M-\tfrac{1}{12}\tilde{D}\|_\infty+\sum_{m=1}^{\infty}\left\|(M-\tfrac{1}{12}\tilde{D})^{2m}\right\|_\infty+\sum_{m=1}^{\infty}\left\|(M-\tfrac{1}{12}\tilde{D})^{2m+1}\right\|_\infty \\
&\le 1+\tfrac{8}{3}+\left(1+\|M-\tfrac{1}{12}\tilde{D}\|_\infty\right)\cdot\sum_{m=1}^{\infty}\left\|(M-\tfrac{1}{12}\tilde{D})^{2m}\right\|_\infty \\
&\le \tfrac{11}{3}+\tfrac{11}{3}\sum_{m=1}^{\infty}\|(M-\tfrac{1}{12}\tilde{D})^2\|_\infty^m \;\le\; \tfrac{11}{3}+\tfrac{11}{3}\sum_{m=1}^{\infty}\left(\tfrac{15}{36}\right)^m
\end{aligned}$$

The last sum is just a geometric series. Hence:

$$\|K_{e_4}^{-1}\|_\infty \le \tfrac{11}{3}+\tfrac{11}{3}\left(1-\tfrac{15}{36}\right)^{-1} \;=\; \tfrac{11}{3}+\tfrac{11}{3}\tfrac{36}{21}<12,$$

Note that we have used the submultiplicativity of $\|\cdot\|_\infty$ twice. The Neumann series is therefore absolutely convergent and defines a Cauchy sequence which is convergent. ∎

Observe that the lemma implies not only the damping property for the inverse matrix of the quartic extrapolation scheme but guarantees also its stability. Nevertheless, in section 7.4, we have given a separate proof for the stability, which requires less computation and permits a common treatment of the four higher order extrapolation schemes.

7.6 Asymptotic expansions and convergence

Now, we have all ingredients to derive an asymptotic expansion of the numeric solution v and to prove the convergence of the five-point stencil extrapolation schemes as a byproduct. The presentation will be given in detail for the nearest neighbor extrapolation scheme. It is possible to adapt the proof for the other extrapolation schemes without any difficulty.

Let us start with the definition of two useful grid functions:

$$\chi:\mathcal{I}_h\to\mathbb{R},\qquad \chi(x_i)\equiv\chi_j:=\begin{cases}0 & \text{for } j\in\{0,1,N-1,N\}\\ 1 & \text{for } 2\le j\le N-2\end{cases}\tag{7.24}$$

$$\delta_i:\mathcal{I}_h\to\mathbb{R},\qquad \delta_j(x_i)\equiv\delta_{ij}:=\begin{cases}0 & \text{for } i\ne j\\ 1 & \text{for } i=j\end{cases}\tag{7.25}$$

The nodes, where χ does not vanish, are called *bulk nodes.* Furthermore we will make use of the following terminology.

Convention: May $G_h\equiv G:\mathcal{I}_h\to\mathbb{R}$ denote a grid function (more precisely a family of grid functions) for any admissible[18] grid spacing h. G is a *sampled* grid function, if there exists (for the whole family) a continuous (smooth) function $g:\mathcal{I}\to\mathbb{R}$ independent of the mesh-width h, such that $G=\hat{g}$, where $\hat{g}$ designates the restriction (sampling) of g onto the grid. Otherwise G is called a *pure* grid

[18] The length of $\mathcal{I}$ should be an natural multiple of the grid spacing. Furthermore the grid spacing is generally supposed to be less than one.

function. In the first case G is also said to be *regular*, while it is *irregular* in the latter one.
Furthermore, let G be uniformly bounded with respect to a norm $\|\cdot\|$ on $\mathcal{F}(\mathcal{I}_h)$. If G admits for some $n \in \mathbb{N}_0$ a representation of the form

$$G = A^{(0)} + hA^{(1)} + ... + h^n A^{(n)}$$

we call the right hand side a *(truncated asymptotic) expansion* of the order n. The coefficient functions $A^{(\nu)} \equiv A_h^{(\nu)}$ for $0 \leq \nu \leq n$ are referred to as *(asymptotic) order functions*. They are supposed to be uniformly bounded[19] with respect to the same norm, although they may explicitly depend on h. In particular they might be regular or irregular. This convention is naturally extendible to expansions with fractional or negative powers of h. For the latter case we have to waive the assumption that G is uniformly bounded.

Remark 7.8. In order to include also the case of irregular asymptotic order function, we have to admit, that the order functions depend not only by their domain of definition on the grid spacing but also numerically. Therefore the asymptotic expansion is not uniquely defined, for instance we may write

$$G = A^{(0)}+h\underbrace{\left(A^{(1)} + hA^{(2)}\right)}_{=:\tilde{A}^{(1)}}+h^3A^{(3)} \quad \text{instead of} \quad G = A^{(0)}+hA^{(1)}+h^2A^{(2)}+h^3A^{(3)} \; .$$

The left expansion pretends that the second order vanishes, while the right expansion takes the existence of a second order into account, that can be split off from the first one. Both forms may be correct with the difference that the right one is more precise. Generally, asymptotic expansions make only sense, if the asymptotic orders vary with the grid spacing as little as possible. The notion of a *complete* asymptotic expansion (up to a certain accuracy) can only be vaguely defined. Roughly speaking it means, that one tries to split off as many orders as possible (up to a given accuracy). So, orders should neither be hidden in the preceding ones nor artificially generated. It should be pointed out, that regular order functions can be uniquely defined.

Recall that the nearest neighbor extrapolation scheme discretizes the boundary value problem (7.8) in the following way:

$$\hat{D}_{e_0} v = \varsigma \;\Leftrightarrow\; \begin{cases} v_0 & = \; u_0 \\ \frac{1}{h^2}[a+b,\; \overset{*}{c}, b,\; a]v_1 & = \; f(h) \\ \frac{1}{h^2}[a,\; b,\; \overset{*}{c},\; b,\; a]v_i & = \; f(ih) \quad i \in \{2, ..., N-2\} \\ \frac{1}{h^2}[a,\; b,\; \overset{*}{c},\; b+a]v_{N-1} & = \; f(Nh-h) \\ v_N & = \; u_L \end{cases} \tag{7.26}$$

The *discrete source function* $\varsigma : \mathcal{I}_h \to \mathbb{R}$ contains the boundary values as well as the source f of the original problem, i.e. $\varsigma = (u_0, \hat{f}_1, ..., \hat{f}_{N-1}, u_L)^\top$.

[19] If there is a common bound for all functions of the family.

Definition 7.3. *Let $g : \mathcal{I}_h \to \mathbb{R}$ be a grid function and $\varsigma : \mathcal{I}_h \to \mathbb{R}$ a given source function. The grid function $\hat{D}_{e_0} g - \varsigma$ is referred to as the* residual *of g with respect to $\hat{D}_{e_0}$ and ς.*

Although given in a special context, this definition applies analogously to all similar cases.

The following theorem is central for the whole section. It postulates an asymptotic expansion with irregular order functions for the numeric solution v. The result will be refined later on by extending the expansion to further orders.

Theorem 7.10. *Let $f \in \mathcal{C}^4(\mathcal{I})$ and $u \in \mathcal{C}^6(\mathcal{I})$ be the solution of the BVP (7.8). Then the solution $v \in \mathcal{F}(\mathcal{I}_h)$ of the scheme (7.26) admits the expansion*

$$v = \hat{u} + hs^{(1)} + h^2 s^{(2)} + h^3 s^{(3)} + \mathrm{O}(h^4), \tag{7.27}$$

where $s^{(1)}, s^{(2)}, s^{(3)} \in \mathcal{F}(\mathcal{I}_h)$ are grid functions, that are bounded with respect to the $\|\cdot\|_\infty$-norm independently *of the grid spacing h and thus in particular for $h \downarrow 0$.*

Remark 7.9. In practice, the numeric solution v is used as an approximation for the the solution u of the differential equation. Here we adopt the converse perspective[20], what is frequently done in the error analysis.
Moreover it is a matter of the point of view, whether (7.27) is considered to expand the numeric solution v or the *error* $v - \hat{u}$ by subtracting $\hat{u}$.

Remark 7.10. The proof is based on the following idea: by Taylor expansion, the residual of $\hat{u}$ naturally comes along in the form of an asymptotic expansion. As the residual of v is zero, it is plausible to assume, that any reasonable approximation of v should have a small residual, i.e. a residual close to the residual of v (continuity argument). Therefore we attempt to modify the grid function $\hat{u}$ in such a way, that the leading terms of its residual do not occur any more. Of course, we try this by an ansatz that reflects the structure of the residual of $\hat{u}$, i.e. an asymptotic expansion. We will see in the subsequent proof, that this procedure leads finally to the claimed expansion.

Proof: Taylor expansion of u and f up to the sixth and fourth derivative respectively yields a residual of the form

$$\hat{D}_{e_0}\hat{u} - \varsigma = h^{-1} r^{(-1)} + \dots + h^3 r^{(3)} + h^4 r^{(4)}. \tag{7.28}$$

Its first five order functions $r^{(\nu)} : \mathcal{I}_h \to \mathbb{R}$ $(-1 \le \nu \le 3)$ are found to be

$$\begin{aligned}
r^{(-1)} &= -\tfrac{1}{12}\big[\partial u(0)\delta_1 - \partial u(L)\delta_{N-1}\big],\\
r^{(0)} &= \tfrac{1}{24}\big[\partial^2 u(0)\delta_1 + \partial^2 u(L)\delta_{N-1}\big],\\
r^{(1)} &= -\tfrac{1}{72}\big[\partial^3 u(0)\delta_1 - \partial^3 u(L)\delta_{N-1}\big],\\
r^{(2)} &= \tfrac{1}{288}\big[\partial^4 u(0)\delta_1 + \partial^4 u(L)\delta_{N-1}\big],\\
r^{(3)} &= -\tfrac{1}{1440}\big[\partial^5 u(0)\delta_1 - \partial^5 u(L)\delta_{N-1}\big],
\end{aligned} \tag{7.29}$$

[20] Remember, that asymptotic similarity is an equivalence relation.

where the stencil at x_1 and x_{N-1} has been expanded around 0 and L respectively (see remark 7.12). Both the boundary conditions being exactly satisfied and the stencil in the bulk nodes – if expanded around its center – do not contribute in these orders. All remainder terms of the Taylor expansion are contained in $r^{(4)}$, which is defined pointwise. At the boundary nodes x_0, x_N and their neighbors x_1, x_{N-1} we obtain

$$\begin{aligned} r^{(4)}(x_0) &= r^{(4)}(x_N) = 0 \\ r^{(4)}(x_i) &= \tfrac{c}{6!}\partial^6 u\big(\eta_1(x_i)\big) + \tfrac{2^6 b}{6!}\partial^6 u\big(\eta_2(x_i)\big) + \tfrac{3^6 a}{6!}\partial^6 u\big(\eta_3(x_i)\big) - \tfrac{1}{4!}\partial^4 f\big(\eta_f(x_i)\big)\ . \end{aligned}$$

with $i \in \{1, N-1\}$ and η_1, η_2, η_3 and η_f being suitable intermediate points[21] depending on x_i.
At the bulk nodes the subsequent remainder is produced $(2 \leq i \leq N-2)$

$$r^{(4)}(x_i) = \tfrac{2^6 a}{6!}\partial^6 u\big(\theta_{-2}(x_i)\big) + \tfrac{b}{6!}\partial^6 u\big(\theta_{-1}(x_i)\big) + \tfrac{b}{6!}\partial^6 u\big(\theta_1(x_i)\big) + \tfrac{2^6 a}{6!}\partial^6 u\big(\theta_2(x_i)\big)\ ,$$

where $\theta_{-2}, \theta_{-1}, \theta_1$ and θ_2 denote appropriate intermediate points[22], whose exact position depends on x_i.
Actually the precise knowledge of the coefficients occurring in the order functions of the residual is not of importance for the moment. We underline, however, that $r^{(-1)}, ..., r^{(3)}$ satisfy condition i) and ii) of definition 7.2. Since their support is given by the interior nodes x_1, x_{N-1} directly adjacent to the boundary nodes x_0, x_N, we get

$$\operatorname{dist}\big(\operatorname{supp}(r^{(\nu)}), \partial\mathcal{I}\big) = h \quad \text{and} \quad \#\operatorname{supp}(r^{(\nu)}) = 2 \quad \text{for} \quad -1 \leq \nu \leq 3. \tag{7.30}$$

Furthermore the six residual order functions are uniformly bounded with respect to $\|\cdot\|_\infty$. This follows clearly from the assumption, that the derivatives of u up to the sixth order and the derivatives of f up to the fourth order are continuous over the *compact* interval $\mathcal{I} = [0, L]$.

In order to generate a residual of magnitude $\mathrm{O}(h^4)$, we consider the ansatz[23]

$$\hat{u} + hs^{(1)} + ... + h^5 s^{(5)}. \tag{7.31}$$

The $s^{(\nu)}$'s represent unknown grid functions, that should turn out to be uniformly

[21] The existence of the intermediate positions is guaranteed by the mean value theorem and the regularity assumptions with respect to u and f. In particular, we have:

$$\begin{array}{cccc} \eta_1(x_1) \in (0,h) & \eta_2(x_1) \in (0,2h) & \eta_3(x_1) \in (0,3h) & \eta_f(x_1) \in (0,h) \\ \eta_1(x_{N-1}) \in (L-h,L) & \eta_2(x_{N-1}) \in (L-2h,L) & \eta_3(x_{N-1}) \in (L-3h,L) & \eta_f(x_{N-1}) \in (L-h,L) \end{array}$$

[22] Analogically to the η's we have:

$$\theta_{-1} \in (x_i - h, x_i) \quad \theta_1 \in (x_i, x_i + h) \quad \theta_{-2} \in (x_i - 2h, x_i) \quad \theta_2 \in (x_i, x_i + 2h)$$

[23] The ansatz is motivated by the observation that the method is convergent. Therefore the ansatz has no singular term and starts with the analytic solution $\hat{u}$ in the zeroth order.

bounded under grid refinement $h \downarrow 0$. Computing the residual of (7.31) yields

$$\hat{D}_{e0}\hat{u} + \sum_{\nu=1}^{5} h^{\nu}\hat{D}_{e0}s^{(\nu)} - \varsigma \;=\; \sum_{\nu=1}^{6} h^{\nu-2}r^{(\nu-2)} + \sum_{\nu=1}^{5} h^{\nu}\hat{D}_{e0}s^{(\nu)} \tag{7.32}$$

$$= \sum_{\nu=1}^{5} h^{\nu}\left[h^{-2}r^{(\nu-2)} + \hat{D}_{e0}s^{(\nu)}\right] \;+\; h^4 r^{(4)} \tag{7.33}$$

Obviously, we can use $\hat{D}_{e0}s^{(\nu)}$ to compensate $h^{-2}r^{(\nu-2)}$ for $\nu \in \{1, ..., 5\}$. So the $s^{(\nu)}$'s are determined by the requirement

$$\hat{D}_{e0}s^{(\nu)} = -h^{-2}r^{(\nu-2)} \quad \Leftrightarrow \quad s^{(\nu)} = -\hat{D}_{e0}^{-1}h^{-2}r^{(\nu-2)}. \tag{7.34}$$

Since $\hat{D}_{e0}$ is invertible (cf. theorem 7.8), the $s^{(\nu)}$'s are well defined. Thus the question arises, whether they are uniformly bounded. This cannot be inferred from the stability of $\hat{D}_{e0}$ since the right hand side of 7.34 (left equation) is of magnitude $\mathrm{O}(h^{-2})$ and therefore gets unbounded if $h \downarrow 0$ (the $r^{(\nu-2)}$'s are however bounded, as we have already remarked). In view of (7.30), $h^{-2}r^{(\nu-2)}$ fulfills all conditions of definition 7.2. According to corollary 7.1 the inverse operator $\hat{D}_{e0}^{-1}$ has the damping property, hence for $h \downarrow 0$

$$\|s^{(\nu)}\|_\infty = \|\hat{D}_{e0}^{-1}h^{-2}r^{(\nu-2)}\|_\infty = \mathrm{O}(1) \qquad \text{with } 1 \le \nu \le 5.$$

Using 7.34, equation 7.33 is equivalent to

$$\varsigma \;=\; \hat{D}_{e0}\big(\hat{u} + hs^{(1)} + ... + h^5 s^{(5)}\big) - h^4 r^{(4)}.$$

Inserting zero in the form of $\hat{D}_{e0}v - \varsigma = 0$ on the left hand side

$$\hat{D}_{e0}v \;=\; \hat{D}_{e0}\big(\hat{u} + hs^{(1)} + ... + h^5 s^{(5)}\big) - h^4 r^{(4)}$$

leads after application of $\hat{D}_{e0}^{-1}$ to the asserted expansion

$$v \;=\; \hat{u} + hs^{(1)} + h^2 s^{(2)} + h^3 s^{(3)} + \underbrace{h^4 s^{(5)} + h^5 s^{(5)} - h^4 \hat{D}_{e0}^{-1} r^{(4)}}_{\mathrm{O}(h^4)}.$$

The underbraced expression is of magnitude $\mathrm{O}(h^4)$, since $s^{(4)}, s^{(5)}$ and $\hat{D}_{e0}^{-1}r^{(4)}$ are uniformly bounded. The latter is true, because $\|\hat{D}_{e0}^{-1}\|_\infty$ is uniformly bounded thanks to the stability provided by theorem 7.8. Therefore the standard estimate yields $\|\hat{D}_{e0}^{-1}r^{(4)}\|_\infty \le \|\hat{D}_{e0}^{-1}\|_\infty \|r^{(4)}\|_\infty < \infty$. ■

Remark 7.11. Notice that $s^{(4)}$ and $s^{(5)}$ are finally absorbed into the $\mathrm{O}(h^4)$-term. This gives rise to the question, whether it would have been necessary to introduce and compute them explicitly. The answer is that they are required by technical reasons, in order to cancel terms of the original residual associated to $\hat{u}$. In the analysis of lattice-Boltzmann schemes we have encountered similar phenomena, where a proof requires the expansion up to higher orders than visible in the final result.

Another issue, closely related to the addressed one, concerns the regularity assumptions. Considering the claimed expansion in theorem 7.10, there is certainly no doubt (by the parallel to Taylor expansion), that every term in the expansion requires one order of regularity concerning f and u. Thus the assumption $f \in \mathcal{C}^2(\mathcal{I})$ and $u \in \mathcal{C}^4(\mathcal{I})$ appears to be sufficient. The cause for stronger regularity assumptions originates from the proof where two subsequent orders internally come into play. Therefore it should be examined whether one could not do better by relaxing the required regularity assumptions which contradict the intuition.

Remark 7.12. The order functions of the residual in 7.29 have been read off from the subsequent Taylor expansion after substituting the numerical values for $a = -\frac{1}{12}, b = \frac{16}{12}$ and $c = -\frac{30}{12}$. Expanding the stencil and f in $x_1 = h$ around 0 yields formally

$$\tfrac{1}{h^2}[a+b,\ \overset{*}{c},\ b,\ a]_h u(h) - f(h) = h^{-1} a\partial u(0) + \sum_{\nu \geq 2} h^{\nu-2}\left[\tfrac{c}{\nu!} + \tfrac{2^\nu b}{\nu!} + \tfrac{3^\nu a}{\nu!} - \tfrac{1}{(\nu-2)!}\right]\partial^\nu u(0). \quad (7.35)$$

Observe, that there is no $\mathrm{O}(h^{-2})$-term at the right hand side since $2a + 2b + c = 0$. By the same reason the coefficient in front of the $\mathrm{O}(h^{-1})$-term simplifies to a. Differentiating formally the Poisson equation $\partial^k f = \partial^{k+2} u$ shows, that f and its derivatives can be expressed by derivatives of u. This is exploited to eliminate f on the right hand side of (7.35). Analogously the stencil in $x_{N-1} = L - h$ can be expanded around L. Since h is then replaced by $-h$ the terms of odd order with respect to h occur with a minus sign (compare the signs in front of $\partial^\nu u(L)$ in (7.29)).

Remark 7.13. Expansion (7.27) contains $\hat{u}$ as a regular order function. In contrast, $s^{(1)}, s^{(2)}$ and $s^{(3)}$ constitute grid functions being not representable as the restriction (sampling) of continuous or even smooth functions onto the grid. This evokes the discrete nature[24] of the scheme.

We have seen in the proof of theorem 7.10, that the coefficient functions $s^{(1)}, s^{(2)}$ and $s^{(3)}$ arise from the residual of $\hat{u}$ generated at the neighbor nodes x_1 and x_{N-1} of the boundary points. Plotting these grid functions over $\mathcal{I}_h$ shows, that they are characterized by large difference quotients[25] near the boundary. So both their origin and their appearance motivate the following convention.
Convention: We call any (part of an) order function of the asymptotic expansion for v a *boundary layer* (german: Rand<u>s</u>chicht) if it is due to the residual in x_1 and x_{N-1}.

Theorem 7.10 implies immediately the convergence. More exactly we obtain:

Corollary 7.3. *Under the hypothesis of theorem 7.10 the scheme (7.26) is first order convergent towards the solution of the BVP (7.8). Hence there is a constant* $C > 0$, *such that*

$$\|v - \hat{u}\|_\infty \ < \ Ch \qquad \textit{for all admissible grid spacings } h < 1.$$

[24]A good numerical scheme is distinguished by the property, that its specifically discrete nature becomes evident as little as possible.

[25]In the next section we consider some examples, that indicate that at least on of the difference quotients $(s_1^{(\nu)} - s_0^{(\nu)})/h$, $(s_N^{(\nu)} - s_{N-1}^{(\nu)})/h$ gets singular for $h \downarrow 0$.

Proof: In the proof of theorem 7.10 we have shown, that the boundary layers $s^{(1)}, s^{(2)}$, and $s^{(3)}$ are uniformly bounded under grid refinement by appropriate, positive constants C_1, C_2 and C_3, i.e. $\|s^{(\nu)}\|_\infty < C_\nu$ for $h \downarrow 0$. Furthermore, equation (7.27) is, by definition of the O-notation, equivalent to

$$\|v - \hat{u} - hs^{(1)} - h^2 s^{(2)} - h^3 s^{(3)}\|_\infty < C_4 h^4$$

with a suitable constant $C_4 > 0$. Adding on both sides $\|hs^{(1)} + h^2 s^{(2)} + h^3 s^{(3)}\|_\infty$ and applying on the left hand side the triangle inequality in the "reverse direction", yields:

$$\|v - \hat{u}\|_\infty < \|hs^{(1)} + h^2 s^{(2)} + h^3 s^{(3)}\|_\infty + C_4 h^4$$

Applying the triangle inequality now on the right hand side and estimating the boundary layers gives

$$\|v - \hat{u}\|_\infty < C_1 h + C_2 h^2 + C_3 h^3 + C_4 h^4.$$

As the function $\kappa \mapsto h^\kappa$ is strictly decreasing for $h < 1$ and $\kappa \in \mathbb{N}$ the assertion ensues with $C := C_1 + C_2 + C_3 + C_4$. ■

Remark 7.14. The residual of $\hat{u}$ in equation 7.28 behaves like $\mathrm{O}(h^{-1})$. So it becomes unbounded (with respect to $\|\cdot\|_\infty$) under grid refinement despite the convergence of the scheme. Therefore the convergence proof cannot be achieved by recourse to the stability of the scheme alone but utilizes essentially the damping property to deal with the singular residual.

Let us quickly examine the convergence of the higher order extrapolation schemes. We anticipate already that their convergence order is equal to the leading order of their residual (generated also in x_1, x_{N-1}) *raised* by two as well, thanks to the effect of the damping property. However, fourth order convergence can not be exceeded due the residual generated in the bulk nodes (see theorem 7.12). Therefore the quartic or even higher extrapolation schemes do not outperform the cubic one.

Theorem 7.11. *Under the assumptions of theorem 7.10, the extrapolation schemes constructed in section 7.3 with the system matrices $\hat{D}_{e_k} (k \in \{1, 2, 3, 4\})$ are convergent. More precisely:*

$$\textit{The} \left\{ \begin{array}{c} \textit{linear} \\ \textit{quadratic} \\ \textit{cubic} \\ \textit{quartic} \end{array} \right\} \textit{extrapolation scheme} \left\{ \begin{array}{c} (\hat{D}_{e_1}) \\ (\hat{D}_{e_2}) \\ (\hat{D}_{e_3}) \\ (\hat{D}_{e_4}) \end{array} \right\} \textit{is} \left\{ \begin{array}{c} 2^{\mathrm{nd}} \\ 3^{\mathrm{rd}} \\ 4^{\mathrm{th}} \\ 4^{\mathrm{th}} \end{array} \right\} \textit{-order accurate.}$$

Proof: (Sketch) Just as for the nearest-neighbor extrapolation scheme an asymptotic expansion of the numeric solution can be established. The arguments apply in complete analogy to the proof of theorem 7.10, because the four schemes are stable according to theorem 7.7 and their corresponding inverse operators satisfy the damping property by corollary 7.2.
Actually, the only point being different is the computation of the residual that results in other expansions accounting for the convergence behavior of each scheme.

Let us therefore have a look at the formal expansions of the different stencils in x_1 analogously to (7.35).

$$\begin{aligned}
&\tfrac{1}{h^2}[\alpha,\, \overset{*}{\beta},\, \gamma,\, \delta,\, \epsilon]_h u(h) - f(h) \\
&\quad = \sum_{\nu\geq 0} h^{\nu-2}\tfrac{1}{\nu!}\left[0^\nu\alpha + \beta + 2^\nu\gamma + 3^\nu\delta + 4^\nu\epsilon\right]\partial^\nu u(0) - \sum_{\nu\geq 0} h^\nu \tfrac{1}{\nu!}\partial^\nu f(0) \\
&\quad = \sum_{\nu\geq 0} h^{\nu-2}\tfrac{1}{\nu!}\left[0^\nu\alpha + \beta + 2^\nu\gamma + 3^\nu\delta + 4^\nu\epsilon\right]\partial^\nu u(0) - \sum_{\nu\geq 2} h^{\nu-2} \tfrac{1}{(\nu-2)!}\partial^\nu u(0)
\end{aligned}$$

The table summarizes the result, that is obtained by evaluating the leading coefficients of this expansion for the four stencils:

extrapolation	stencil in x_1	expansion starts with
linear	$[\alpha,\, \overset{*}{\beta},\, \gamma,\, \delta,\, \epsilon] = [b+2a,\, \overset{*}{c-a},\, b,\, a,\, 0]$	O(1)-term
quadratic	$[\alpha,\, \overset{*}{\beta},\, \gamma,\, \delta,\, \epsilon] = [b+3a,\, \overset{*}{c-3a},\, b+a,\, a,\, 0]$	O(h)-term
cubic	$[\alpha,\, \overset{*}{\beta},\, \gamma,\, \delta,\, \epsilon] = [b+4a,\, \overset{*}{c-6a},\, b+4a,\, 0,\, 0]$	O(h^2)-term
quartic	$[\alpha,\, \overset{*}{\beta},\, \gamma,\, \delta,\, \epsilon] = [b+5a,\, \overset{*}{c-10a},\, b+10a,\, -4a,\, a]$	O(h^3)-term

The right column gives the order of the residual in x_1, x_{N-1}. (Remember, that the residual in the bulk nodes resides in the fourth order independently of the extrapolation.) Taking into account that the residual in x_1, x_{N-1} gives rise to boundary layers, whose order is increased by 2 (damping property), we get the following asymptotic expansions

$$\begin{aligned}
v_{e_1} &= \hat{u} + h^2 s_{e_1}^{(2)} + h^3 s_{e_1}^{(3)} + \mathrm{O}(h^4), \\
v_{e_2} &= \hat{u} + h^3 s_{e_2}^{(3)} + \mathrm{O}(h^4), \\
v_{e_3} &= \hat{u} + \mathrm{O}(h^4), \\
v_{e_4} &= \hat{u} + \mathrm{O}(h^4),
\end{aligned}$$

where v_{e_k} denotes the numeric solution obtained by the scheme e_k. For the cubic extrapolation scheme, the leading boundary layer is in the fourth order, while the first boundary layer of the quartic extrapolation scheme appears only in the fifth order and is therefore absorbed into the O(h^4)-term as well.
Observe also that we have indexed the boundary layers, since they depend on the kind of extrapolation. The assertion is now obvious by proceeding as in the proof of corollary 7.3. ■

Remark 7.15. We emphasize once more, that we could prove neither the convergence of the nearest-neighbor nor of the linear extrapolation scheme without the damping property. Moreover, by ignoring the damping property, the *provable* convergence rates for the quadratic, cubic and quartic extrapolation scheme would be 1, 2 and 3 respectively, which disagrees with the *observable* convergence rates.
Only if the extrapolation is executed by a polynomial of degree five, the residual in x_1 and x_{N-1} is of the order four as in the bulk nodes. Stability provided, we do not need the damping property in this case to show that the convergence rate of the scheme is four. However, the damping property would still be necessary, if one wants to expand the error as we have done it.

Let us return to the nearest-neighbor scheme. Under stronger regularity assumptions the expansion (7.27) can be continued to higher orders. Generally, we get with every additional order of regularity an additional order in the expansion.

We have seen that the stencil in x_1 and x_{N-1} gives rise to the boundary layers $s^{(1)}, s^{(2)}$ and $s^{(3)}$, that correspond to the first three leading orders of the error $v - \hat{u}$. In expansion (7.27) the error generated by the five-point stencil in the bulk nodes $x_2, ..., x_{N-2}$ is hidden in the unspecified $\mathrm{O}(h^4)$-term. If we demand on u for more regularity, we can make explicit the leading orders of the error associated to the stencil in the bulk nodes. It turns out, that the order functions of this error are described by regular grid functions. This contrasts to the boundary layers. To understand this better, let us consider a concrete situation[26], specified in the hypothesis of the following theorem. Examples to illustrate the expansion are presented in the next section.

Theorem 7.12. *May $u \in \mathcal{C}^{10}(\mathcal{I})$ be the solution of the BVP (7.8) with the source $f \in \mathcal{C}^8(\mathcal{I})$. If v denotes the solution of the scheme (7.26) then the truncated expansion*

$$v = \hat{u} + hs^{(1)} + h^2 s^{(2)} + h^3 s^{(3)} + h^4\big[\hat{w} + s^{(4)}\big] + h^5 s^{(5)} + h^6\big[\hat{z} + s^{(6)}\big] + h^7 s^{(7)} + \mathrm{O}(h^8)\,.$$

holds true. The regular parts $\hat{w}$, $\hat{z}$ of the fourth and sixth order function respectively are determined by the subsequent boundary value problems

$$\begin{aligned} w''(x) &= -2\,\tfrac{2^6 a + b}{6!}\,\partial^6 u(x) \quad \textit{for } x \in (0,L) \quad \textit{and } w(0) = w(L) = 0 \\ &\Rightarrow w(x) = \tfrac{1}{90}\Big[\partial^4 u(x) - \tfrac{x}{L}\partial^4 u(L) - \tfrac{L-x}{L}\partial^4 u(0)\Big], \end{aligned}$$

$$\begin{aligned} z''(x) &= -2\,\tfrac{2^8 a + b}{8!}\,\partial^8 u(x) \quad \textit{for } x \in (0,L) \quad \textit{and } w(0) = w(L) = 0 \\ &\Rightarrow z(x) = \tfrac{1}{1008}\Big[\partial^6 u(x) - \tfrac{x}{L}\partial^6 u(L) - \tfrac{L-x}{L}\partial^6 u(0)\Big]. \end{aligned}$$

The boundary layers (irregular parts) are given by:

$$\begin{aligned}
\hat{D}_{e_0} s^{(1)} &= \tfrac{1}{12h^2}\,\big[\partial u(0)\delta_1 - \partial u(L)\delta_{N-1}\big] \\
\hat{D}_{e_0} s^{(2)} &= -\tfrac{1}{24h^2}\,\big[\partial^2 u(0)\delta_1 + \partial^2 u(L)\delta_{N-1}\big] \\
\hat{D}_{e_0} s^{(3)} &= \tfrac{1}{72h^2}\,\big[\partial^3 u(0)\delta_1 - \partial^3 u(L)\delta_{N-1}\big] \\
\hat{D}_{e_0} s^{(4)} &= -\tfrac{1}{288h^2}\,\big[\partial^4 u(0)\delta_1 + \partial^4 u(L)\delta_{N-1}\big] \\
\hat{D}_{e_0} s^{(5)} &= \tfrac{1}{1440h^2}\,\big[\partial^5 u(0)\delta_1 - \partial^5 u(L)\delta_{N-1}\big] + \tfrac{1}{12h^2}\,\big[\partial w(0)\delta_1 - \partial w(L)\delta_{N-1}\big] \\
\hat{D}_{e_0} s^{(6)} &= \tfrac{19}{1728h^2}\,\big[\partial^6 u(0)\delta_1 - \partial^6 u(L)\delta_{N-1}\big] + \tfrac{25}{24h^2}\,\big[\partial^2 w(0)\delta_1 + \partial^2 w(L)\delta_{N-1}\big] \\
\hat{D}_{e_0} s^{(7)} &= \tfrac{172}{15475h^2}\,\big[\partial^6 u(0)\delta_1 - \partial^6 u(L)\delta_{N-1}\big] + \tfrac{71}{72h^2}\,\big[\partial^3 w(0)\delta_1 - \partial^3 w(L)\delta_{N-1}\big] \\
&\qquad + \tfrac{1}{12h^2}\,\big[\partial z(0)\delta_1 - \partial z(L)\delta_{N-1}\big]
\end{aligned}$$

Proof: (Sketch) The proof proceeds anologously to the proof of the previous theorem. Recall that the five-point stencil has the following formal Taylor expansion around its center at x:

$$\tfrac{1}{h^2}[a,\, b,\, \overset{*}{c},\, b,\, a]_h u(x) = \partial^2 u(x) + 2\sum_{\nu \geq 3} h^{2\nu-2}\,\tfrac{2^{2\nu} a + b}{(2\nu)!}\,\partial^{2\nu} u(x). \tag{7.36}$$

Therefore it gives rise to error terms residing in the fourth and sixth order function. For instance, using (7.35) and (7.36) the fourth order function of the residual associated to $\hat{u}$ is

$$\begin{aligned} r^{(4)} &= \big[\tfrac{c}{6!} + \tfrac{2^6 b}{6!} + \tfrac{3^6 a}{6!} - \tfrac{1}{4!}\big]\partial^6 u(0)\delta_1 + \big[\tfrac{c}{6!} + \tfrac{2^6 b}{6!} + \tfrac{3^6 a}{6!} - \tfrac{1}{4!}\big]\partial^6 u(L)\delta_{N-1} + 2\,\tfrac{2^6 a + b}{6!}\,\widehat{\partial^6 u}\,\chi \\ &= \tfrac{1}{288}\partial^6 u(0)\delta_1 + \tfrac{1}{288}\partial^6 u(L)\delta_{N-1} + \tfrac{1}{90}\,\widehat{\partial^6 u}\,\chi\,, \end{aligned}$$

where χ is defined in (7.24). The first two addends can be compensated by means of an appropriately chosen boundary layer as in the proof of theorem 7.10. For the third addend, being different from zero in the bulk nodes, we consider the ansatz $\hat{u} + h^6\hat{w}$. Computing the fourth order function of the residual associated with this ansatz yields the additional term $\widehat{\partial^2 w}\,\chi$, that can be used to compensate $\frac{1}{90}\,\widehat{\partial^6 u}\,\chi$, if w is defined as indicated in the theorem. Analogously one deals with $r^{(6)}$, where z comes into play.

[26]We abstain from treating the general case, since it gets notationally quite cumbersome without providing a better grasp.

As the Dirichlet boundary conditions do not depend on h, they are exactly satisfied by $\hat{u}$ in the zeroth order of the expansion. This implies homogeneous boundary conditions for w, z.
The coefficients in front of the derivatives of w, z occurring in the formulas for $s^{(5)}, s^{(6)}, s^{(7)}$ are determined by expanding the stencil in x_1, x_{N-1} around 0 and L as in (7.35). ■

Remark 7.16. The terms of the residual, generated at the bulk nodes, are associated to terms in the expansion of v that exhibit the same order with respect to h. This distinguishes them from the residual terms produced in x_1, x_{N-1} gaining two orders by the damping property.
Notice also, that only orders of even index in the expansion of v will have regular parts like w, z.

Remark 7.17. The sixth and eighth derivative of u occur as source functions for w and z respectively. If we go to further orders in the expansion of v, derivatives of w and z will occur themselves as source functions. The reason, why w does not appear as a source for z results from the fact, that the stencil (at the bulk nodes) approximates the second derivative with fourth order accuracy.

Remark 7.18. w and z are defined independently of the stencil in x_1, x_{N-1}, i.e. independently of the used extrapolation. So they appear equally in the expansion of all extrapolation schemes. This implies in particular, that their accurary is at most fourth order (since w emerges in the fourth order).

Remark 7.19. Observe that the regular parts of the expansion (e.g. u, w, z) are all defined via the same type of boundary value problem. This might not be amazing, in particular as we consider a linear discretization for a linear BVP. Nevertheless, it is a structural property of the expansion that we already have met in the analysis of lattice-Boltzmann methods.

A correction strategy

On the previous pages we have given a thorough analysis of the nearest-neighbor extrapolation scheme. In comparison to the other extrapolation schemes, this scheme solves the discrepancy of the boundary conditions only in a modest (not to say bad) way, although it might be the most intuitive one.
The (consistency)analysis can be used to improve systematically a numeric scheme and to synthesize thus new, more performant schemes. The suitable correction of the original scheme is suggested by a Taylor expansion of the residual.
Let us demonstrate this exemplarily by means of the nearest neighbor extrapolation scheme. In particular we consider the Taylor expansion of this scheme in x_1.

$$\tfrac{1}{h^2}[b+a,\ \overset{*}{c},\ b,\ a]\hat{u}_1 - \varsigma_1 = h^{-1}a\partial u(0) - \tfrac{a}{2}\partial^2 u(0) + h\tfrac{a}{6}\partial^3 u(0) - h^2\tfrac{a}{24}\partial^4 u(0) + \mathrm{O}(h^3) \tag{7.37}$$

The goal consists in modifying the stencil on the left hand side in such a way, that the terms of the highest order on the right hand side are eliminated. This is done by subtracting appropriate stencils, that approximate the occurring derivatives in (7.37). From proposition 7.2 we get the following right-sided stencils, approximating the first four derivatives with first order accuracy.

$$\begin{aligned}
\tfrac{1}{h}\,[\overset{*}{-1},\ 1]_h u &= \partial u + \tfrac{1}{2}h\partial^2 u + \tfrac{1}{6}h^2\partial^3 u + \tfrac{1}{24}h^3\partial^4 u + \mathrm{O}(h^4)\\
\tfrac{1}{h^2}[\overset{*}{1},\ -2,\ 1]_h u &= \partial^2 u + h\partial^3 u + \tfrac{7}{12}h^2\partial^4 u + \mathrm{O}(h^3)\\
\tfrac{1}{h^3}[\overset{*}{-1},\ 3,\ -3,\ 1]_h u &= \partial^3 u + \tfrac{3}{2}h\partial^4 u + \mathrm{O}(h^2)\\
\tfrac{1}{h^4}[\overset{*}{1},\ -4,\ 6,\ -4,\ 1]_h u &= \partial^4 u + \mathrm{O}(h)
\end{aligned} \tag{7.38}$$

By adding $-\frac{a}{h}$ times the standard right-sided stencil[27] for the first derivative, one gets rid of the leading residual term $h^{-1}a\partial_x u(0)$ in (7.37):

[27]Note: $\frac{1}{h}[\overset{*}{-1},\ 1]_h u(0) = \frac{1}{h}[-1,\ \overset{*}{1}]u_1 = \frac{1}{h}(u_1 - u_0) = \frac{1}{h}\big(u(h) - u(0)\big)$. Even though the stencil is centered in $x_1 = h$ the expansion is done around $x_0 = 0$. This is applied analogically to the other stencils in (7.38).

$$\begin{array}{lcl}
\frac{1}{h^2}[b+a,\ \overset{*}{c},\ b,\ a]\hat{u}_1 - \varsigma_1 & = & h^{-1}a\partial u(0) - \frac{a}{2}\partial^2 u(0) + h\frac{a}{6}\partial^3 u(0) - h^2\frac{a}{24}\partial^4 u(0) + \mathrm{O}(h^3) \\
\frac{1}{h}\ [-1,\ \overset{*}{1}]\hat{u}_1 & = & \partial u(0) + h\frac{1}{2}\partial^2 u(0) + h^2\frac{1}{6}\partial^3 u(0) + h^3\frac{1}{24}\partial^4 u(0) + \mathrm{O}(h^4)\ \big| - \frac{a}{h}\cdot \\
\hline
\frac{1}{h^2}[b+2a,\ \overset{*}{c-a},\ b,\ a]\hat{u}_1 - \varsigma_1 & = & -a\partial^2 u(0) - h^2\frac{a}{12}\partial^4 u(0) + \mathrm{O}(h^3)
\end{array}$$

In order to obtain a residual of the order $\mathrm{O}(h)$, we have to cancel the leading term (second derivative of u) in the expansion of the corrected stencil directly above. This is done analogously utilizing now the stencil for the second derivative in (7.38).

$$\begin{array}{lcl}
\frac{1}{h^2}[b+2a,\ \overset{*}{c-a},\ b,\ a]\hat{u}_1 - \varsigma_1 & = & -a\partial^2 u(0) - h^2\frac{a}{12}\partial^4 u(0) + \mathrm{O}(h^3) \\
\frac{1}{h^2}[1,\ \overset{*}{-2},\ 1]\hat{u}_1 & = & \partial^2 u(0) + h\partial^3 u(0) + h^2\frac{7}{12}\partial^4 u(0) + \mathrm{O}(h^3)\quad \big| + a\cdot \\
\hline
\frac{1}{h^2}[b+3a,\ \overset{*}{c-3a},\ b+a,\ a]\hat{u}_1 - \varsigma_1 & = & ha\partial^3 u(0) + h^2\frac{a}{2}\partial^4 u(0) + \mathrm{O}(h^3)
\end{array}$$

The procedure can be continued. Thus, a second order residual is obtained by

$$\begin{array}{lcl}
\frac{1}{h^2}[b+3a,\ \overset{*}{c-3a},\ b+a,\ a]\hat{u}_1 - \varsigma_1 & = & ha\partial^3 u(0) + h^2\frac{a}{2}\partial^4 u(0) + \mathrm{O}(h^3) \\
\frac{1}{h^3}[-1,\ \overset{*}{3},\ -3,\ 1]\hat{u}_1 & = & \partial^3 u + \frac{3}{2}h\partial^4 u + \mathrm{O}(h^2)\quad \big| - ha\cdot \\
\hline
\frac{1}{h^2}[b+4a,\ \overset{*}{c-6a},\ b+4a,\ 0]\hat{u}_1 - \varsigma_1 & = & -h^2 a\partial^4 u(0) + \mathrm{O}(h^3)
\end{array}$$

Once more this can be corrected, to reach even a residual of third order.

$$\begin{array}{lcl}
\frac{1}{h^2}[b+4a,\ \overset{*}{c-6a},\ b+4a,\ 0]\hat{u}_1 - \varsigma_1 & = & -h^2 a\partial^4 u(0) + \mathrm{O}(h^3) \\
\frac{1}{h^4}[1,\ \overset{*}{-4},\ 6,\ -4,\ 1]\hat{u}_1 & = & \partial^4 u + \mathrm{O}(h)\quad \big| + h^2 a\cdot \\
\hline
\frac{1}{h^2}[b+5a,\ \overset{*}{c-10a},\ b+10a,\ -4a,\ a]\hat{u}_1 - \varsigma_1 & = & \mathrm{O}(h^3)
\end{array}$$

It turns out that the stencils of the higher order extrapolation methods are recovered by this procedure (compare with page 315). This unveils another connection between the extrapolation schemes.

Observe, however, that the correction of the nearest neighbor extrapolation scheme does not lead inevitably to the linear extrapolation scheme. The analysis by Taylor expansion does not exactly prescribe the correction but gives only an indication, i.e. a necessary condition, that the correcting stencil must satisfy. Other possibilities are imaginable. Instead of the standard stencil for the first derivative, we could use one of the following two stencils

$$\begin{array}{lcl}
\frac{1}{2h}\ [\overset{*}{-3},\ 4,\ 1]_h u & = & \partial u - \frac{1}{3}h^2\partial^3 u - \frac{1}{4}h^3\partial^4 u\mathrm{O}(h^4) \\
\frac{1}{2h}\ [-1,\ \overset{*}{0},\ 1]_h u & = & \partial u + \frac{1}{6}h^2\partial^3 u + \mathrm{O}(h^4)
\end{array}$$

to eliminate the $h^{-1}a\partial u(0)$ term in (7.37).

Stability and the damping property provided, this shows finally, that a multitude of different five-point stencil schemes can be generated by modifying only the stencils in x_1 and x_{N-1}. The sole condition to ensure at least convergence of first order is, that the coefficients of the stencil must sum up to zero.

7.7 Numeric experiments

Arriving at the end of this chapter we provide some examples which illuminate the asymptotic expansion that has been established in the previous section, especially in theorem 7.12. In particular, we try to convey a feeling for the two types of grid functions occurring in the expansion: *sampled (regular)* grid functions and *pure (irregular)* grid functions (confer convention on page 330). Beyond that, we want to observe, how the expansion approximates numerically the solution v of the discrete scheme. We restrict the numeric experiments to the five-point stencil with nearest neighbor extrapolation. From the point of asymptotics this is the most interesting method, since its expansion is nontrivial in all orders.

We have selected four test problems to be discussed in the sequel. Besides the asymptotic order functions, introduced in the previous section, the quantities in table 7.1 are of interest for our numeric study.

$$
\begin{aligned}
D_0 &:= v - \hat{u} \quad \leftrightarrow \quad hs^{(1)} + h^2 s^{(2)} + h^3 s^{(3)} + h^4(s^{(4)} + \hat{w}) + h^5 s^{(5)} + h^6(s^{(6)} + \hat{z}) + h^7 s^{(7)} =: T_0 \\
D_1 &:= v - \hat{u} - hs^{(1)} \quad \leftrightarrow \quad h^2 s^{(2)} + h^3 s^{(3)} + h^4(s^{(4)} + \hat{w}) + h^5 s^{(5)} + h^6(s^{(6)} + \hat{z}) + h^7 s^{(7)} =: T_1 \\
D_2 &:= v - \hat{u} - hs^{(1)} - h^2 s^{(2)} \quad \leftrightarrow \quad h^3 s^{(3)} - h^4(s^{(4)} + \hat{w}) - h^5 s^{(5)} - h^6(s^{(6)} + \hat{z}) + h^7 s^{(7)} =: T_2 \\
D_3 &:= v - \hat{u} - hs^{(1)} - h^2 s^{(2)} - h^3 s^{(3)} \quad \leftrightarrow \quad h^4(s^{(4)} + \hat{w}) - h^5 s^{(5)} - h^6(s^{(6)} + \hat{z}) + h^7 s^{(7)} =: T_3 \\
D_4 &:= v - \hat{u} - hs^{(1)} - h^2 s^{(2)} - h^3 s^{(3)} - h^4(s^{(4)} + \hat{w}) \quad \leftrightarrow \quad h^5 s^{(5)} + h^6(s^{(6)} + \hat{z}) + h^7 s^{(7)} =: T_4 \\
D_5 &:= v - \hat{u} - hs^{(1)} - h^2 s^{(2)} - h^3 s^{(3)} - h^4(s^{(4)} + \hat{w}) - h^5 s^{(5)} \quad \leftrightarrow \quad h^6(s^{(6)} + \hat{z}) + h^7 s^{(7)} =: T_5 \\
D_6 &:= v - \hat{u} - hs^{(1)} - h^2 s^{(2)} - h^3 s^{(3)} - h^4(s^{(4)} + \hat{w}) - h^5 s^{(5)} - h^6(s^{(6)} + \hat{z}) \quad \leftrightarrow \quad h^7 s^{(7)} =: T_6 \\
D_7 &:= v - \hat{u} - hs^{(1)} - h^2 s^{(2)} - h^3 s^{(3)} - h^4(s^{(4)} + \hat{w}) - h^5 s^{(5)} - h^6(s^{(6)} + \hat{z}) - h^7 s^{(7)}
\end{aligned}
$$

Table 7.1: Definition of the *deviations* D_k and the *cut tails* T_k to be compared with the D_k's. Note: $D_k = T_k + \mathrm{O}(h^8)$ for $0 \le k \le 6$.

The deviation D_k corresponds to the numeric solution v minus the expansion truncated after the k'th asymptotic order. If we identify v with the infinite expansion, the deviation is just its tail. As we have to dispose of the first eight asymptotic order functions to compute the deviations $D_0, ..., D_7$ we can calculate additionally the associated cut tails $T_0, ..., T_6$. An order by order comparison of the deviations and the cut tails is used to *validate numerically* the expansion in theorem 7.12. This is done by plotting the maximum norm of D_k and T_k versus the grid spacing h in a diagram, whose abscissa and ordinate are logarithmically scaled (see figure 7.3, 7.4).

To get an idea of how the order functions look like, we plot them over the spatial domain, which is chosen as the unit interval. While the zeroth asymptotic order $\hat{u}$ corresponds to the solution of the differential equation and is compared to v, the other asymptotic orders are compared with the rescaled deviations. This means that the deviation D_k is divided by h^{k+1} to obtain a quantity of magnitude $\mathrm{O}(1)$. The advantage of the rescaled deviations is, that they are grid independent with respect to their leading order; so they can be compared more easily on grids of different mesh size (confer e.g. figure 7.1, 7.2).

diagram (subfigure)	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
order function	$\hat{u}$	$s^{(1)}$	$s^{(2)}$	$s^{(3)}$	$s^{(4)} + \hat{w}$	$s^{(5)}$	$s^{(6)} + \hat{z}$	$s^{(7)}$
to be compared with	v	$\frac{1}{h}D_0$	$\frac{1}{h^2}D_1$	$\frac{1}{h^3}D_2$	$\frac{1}{h^4}D_3$	$\frac{1}{h^5}D_4$	$\frac{1}{h^6}D_5$	$\frac{1}{h^7}D_6$

Table 7.2: Quantities being compared in plots over the spatial domain (grid). Asymptotic order functions are indicated by solid lines $-$, while the quantities in the third line are represented by circles $\circ$. Note: $\frac{1}{h^{k+1}}D_k = (k+1)$'th order function $+\,\mathrm{O}(h)$

Sinusoidal test problem (Example 1)

The solution of the test problem over the interval $\mathcal{I} = [0,1]$ (i.e. $L = 1$)

$$u(0) = u(1) = 0, \quad u''(x) = -\pi^2 \sin(\pi x) \quad \Rightarrow \quad u(x) = \sin(\pi x)$$

is given by the sine.
In **Diagram (0)** of **figure 7.1** and **7.2** we recognize the numeric solution v (circles) approximating the analytic solution of the test problem (line) on two different grids.

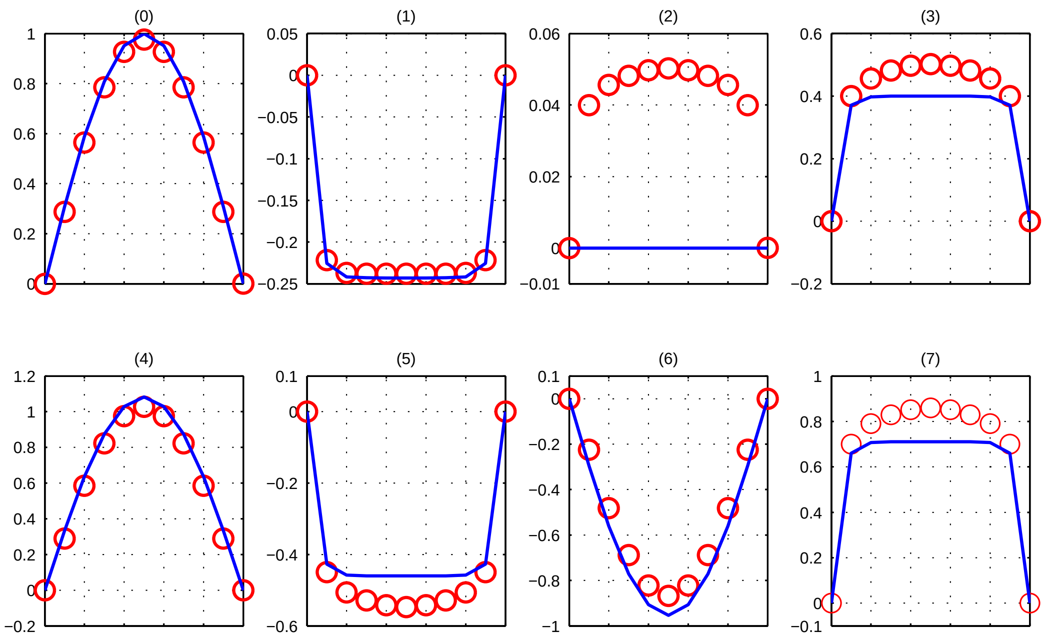


Figure 7.1: Comparison of the asymptotic order functions (lines) and the rescaled deviations (circles) over the spatial domain $[0,1]$ for example 1 (sine). A grid with 11 nodes has been used. The number over each subfigure refers to table 7.2 which specifies the plotted quantities.

Diagram (1) displays the rescaled deviation $h^{-1}D_0$ (circles), which is the error $v - \hat{u}$ divided by the grid spacing. Both of the diagrams look similar, especially the minimum attains almost the same value in both cases; therefore it does not depend significantly on the grid spacing. This can be taken as an experimental evidence

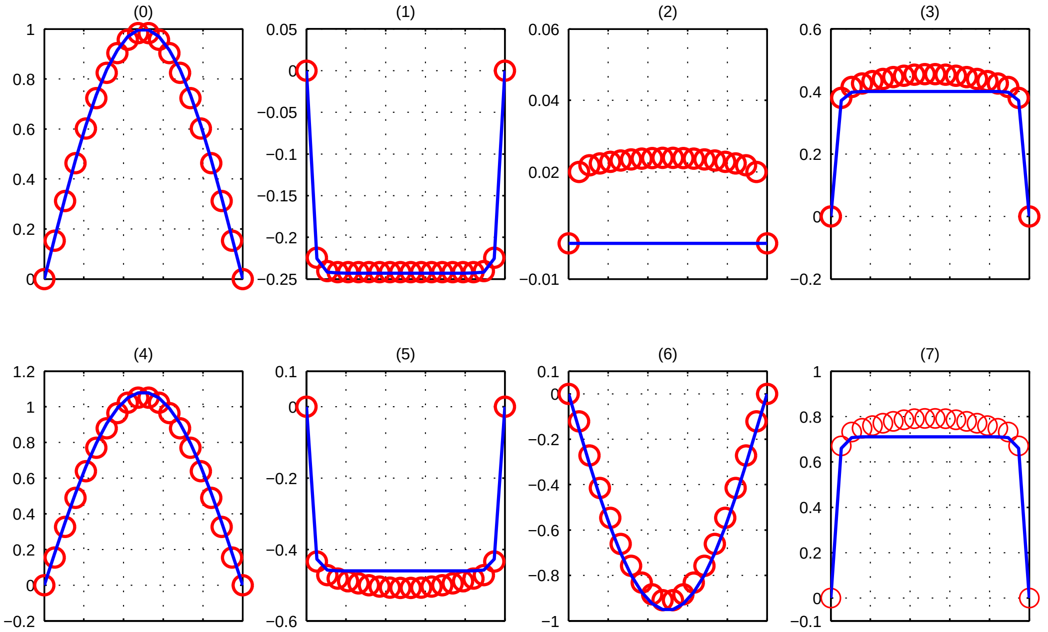


Figure 7.2: Same as in figure 7.1 for a grid with 20 nodes.

for the first order convergence behavior of the scheme, meaning that the error is approximately proportional to the grid spacing.

Another feature is striking: From the left and right boundary we need only one grid step to reach nearly the full depth of the bowel-like curve. Its walls get therefore the steeper the smaller the grid spacing becomes. This behavior can not be described by a smooth function with a gradient being independent of the grid size. In fact, we see here a typical *pure* grid function. Connecting the circles by straight lines yields the graph of a continuous function over the interval $[0, 1]$, whose values depend not only on the spatial position but also essentially on the considered grid (at least near the two boundary points). Such a grid dependent error indicates often a bad behavior of the scheme in solitary nodes as in our case, where the scheme approximates poorly the equation at the direct neighbors of the boundary nodes.

According to theorem 7.12 the main contribution to the total error is given by the first boundary layer $s^{(1)}$. Even if the line[28] marking $s^{(1)}$ seems to match perfectly with the rescaled error $h^{-1}D_0$, a closer look reveals, in **diagram (2)**, that there is still something left, if we subtract $hs^{(1)}$ from the error $v - \hat{u}$ and divide thereafter by h^2. Comparing the resulting rescaled deviation $h^{-2}D_1$ on both grids, we observe that they do not agree with respect to their extremal values. The maximum in figure 7.2 (2) is by a factor of around $\frac{1}{2}$ smaller than in 7.1 (2), what is roughly the ratio of the two grid spacings. Hence there should be no contribution in the second order. In fact, the boundary layer $s^{(2)}$ must vanish, as the second derivative of

[28]The solid line indicating $s^{(1)}$ sholuld not belie the fact, that $s^{(1)}$ is actually a grid function like v, whose values are only defined in the grid nodes. The line has been chosen for the reason of clarity of the diagram!

$\sin(\pi x)$ is zero at the boundary points (compare definition of $s^{(2)}$ in theorem 7.12). Therefore the curve described by the circles in diagram (2) corresponds essentially to the third asymptotic order function multiplied by h.

In **diagram (3)** of figure 7.1 and 7.2 the circles reach approximately the same height. This documents, that the rescaled deviation $h^{-2}R_1$ in (2) becomes an O(1)-quantity after an additional division by h. Observe that $h^{-3}R_1 = h^{-3}R_2$ as $s^{(2)} = 0$.
Unlike diagram (1) the circles in (3) are clearly displaced with respect to the plateau corresponding to $s^{(3)}$. It seems as if the plateau is superimposed by a sinusoidal curve. The displacement fades however with decreasing grid spacing (compare the diagram (3) in both figures!).

Indeed, the **subfigures (4)** corroborate the impression, that the displacement in (3) can be ascribed to the fourth order. In contrast to the previous diagrams the circles representing the rescaled deviation $h^{-4}D_3$ appear to be located on a smooth curve. As the fourth derivative of the analytic solution u vanishes at the boundary points, the boundary layer $s^{(4)}$ is zero. This entails, that the fourth order function comprises only the regular part w.
The fifth and seventh order function in the **diagrams (5)** and **(7)** resemble the third one, while the sixth order function exposes a comportment, that is analogous to the fourth one.
In **figure 7.3** we have plotted the maximum norms of the deviations (circles) and the cut tails versus the grid spacing h in logarithmically scaled coordinate systems. Apparently, in every diagram the circles lie virtually on a *straight* line, that coincides well with the line associated to the corresponding cut tail. Exceptions are only found in diagram (6) and (7), where the values approach the range of machine precision. The straight lines indicate that the deviations and the cut tails depend affine-linearly on some power of the grid spacing, whose exponential is determined by their slopes. Apart from subfigure (1) the lines are almost parallel to the dashed reference lines indicating a slope of $k+1$. This shows that D_k and T_k are in their leading order proportional to h^{k+1}, i.e. $D_k, T_k = \mathrm{O}(h^{k+1})$. The different behavior in diagram (1) is due to the fact, that $s^{(2)} = 0$. Therefore the slope is actually close to 3.

Let us finally throw a glance at the slopes in **table 7.3**. Clearly, the (relative) difference between the deviation D_k and the cut tail T_k is expected to become the less the longer the cut tail is. This explains, why the slopes comply for $k = 0, 1, 2$ with each other up to the fifth and fourth digit respectively. For $k = 1$ and $k = 2$ we obtain equal rates since $s^{(2)} = 0$.
The slope of $\log(\|T_6\|)_\infty$ is 7 within the provided accuracy, because $T_6 = h^7 s^{(7)}$. Note, that this value might differ from 7 choosing other parameters, since the boundary layer $s^{(7)}$ and in particular its extrema might slightly vary with the grid spacing. Therefore strict proportionality of T_6 to h^7 need not always be observed.

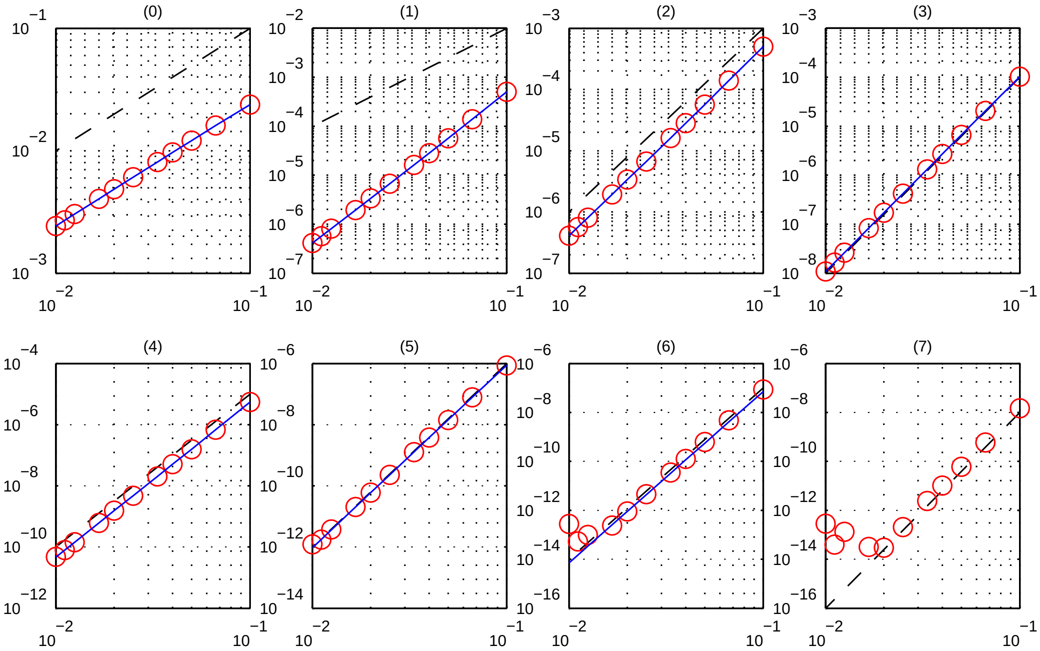


Figure 7.3: Double-logarithmic plot of $\|D_k\|_\infty$ (circles) and $\|T_k\|_\infty$ (lines) for $0 \leq k \leq 7$ versus the grid spacing varying between $\frac{1}{100}$ and $\frac{1}{10}$. A grid sequence with $11, 16, 21, 26, 31, 41, 51, 61, 81, 91$ and 101 nodes has been employed. The associated grid spacings are given by $h = (\#nodes - 1)^{-1}$. The dashed lines, covered in some diagrams, represent reference lines indicating exact proportionality to a power of h, i.e. linear growth in (0), quadratic growth in (1) and so on. Any line running parallel to a reference line expresses an affine-linear dependence on the corresponding power of h.

k	slope of $\log(\|D_k\|_\infty)$	slope of $\log(\|T_k\|_\infty)$
0	0.996972	0.996972
1	3.069949	3.069948
2	3.069949	3.069948
3	3.983730	3.983725
4	5.053250	5.053558
5	5.971200	5.975647
6	7.134063	7.000000
7	8.464278	-

Table 7.3: The slopes have been determined by taking the gradient of the best-fit-straight-line for $(\log h, \log(\|D_k\|_\infty))$ and $(\log h, \log(\|T_k\|_\infty))$ respectively with the grid spacing $h \in \{\frac{1}{10}, \frac{1}{15}, \frac{1}{20}, \frac{1}{25}\}$. This corresponds to the first four grids of the sequence mentioned in the caption of figure 7.3.

Another sinusoidal test problem (Example 2)

Let us proceed to the second example

$$u(0) = u(1) = 0, \quad u''(x) = 4\pi^2 \cos(2\pi x) \quad \Rightarrow \quad u(x) = 1 - \cos(2\pi x)$$

complementing the previous one, as its solution contains the second trigonometric function, the cosine. In contrast to the sine we take the cosine here over a full period.

k	slope of $\log(\|D_k\|_\infty)$	slope of $\log(\|T_k\|_\infty)$
0	2.032256	2.032254
1	2.032256	2.032254
2	3.989398	3.989326
3	3.989398	3.989326
4	5.981233	5.993534
5	5.981233	5.993534
6	7.994524	6.940423
7	7.994524	-

Table 7.4: Slopes of $\log(\|D_k\|_\infty)$ and $\log(\|T_k\|_\infty)$ for the second example (cosine). Note that these quantities depend to a certain extent on which grids are used to determine the best-fit-straight-line. Here we took $h \in \{\frac{1}{25}, \frac{1}{30}, \frac{1}{40}, \frac{1}{50}\}$. This corresponds to the fourth till seventh grid of the sequence that figure 7.4 is based on.

Considering **table 7.4** and **figure 7.4** it seems astonishing at first sight, that v approximates $\hat{u}$ with second order accuracy. The reason for this behavior is found in the fact, that $s^{(1)}$ vanishes for this special case (cf. figure 7.5 (1)), as the derivative of the cosine, namely the sine, vanishes at the boundaries. Similarly, we observe no boundary layers in the third, fifth and seventh order. Instead we get in the even orders a superposition of regular and irregular grid functions, where the latter ones are just the boundary layers.

The value for the slope of $\log(\|T_6\|_\infty)$ in table 7.4 appears confusing, since it is supposed to be close to eight instead to seven. In fact, no relevant meaning inheres in this number, since $s^{(7)}$ and thus $T_6 = h^7 s^{(7)}$ are zero. However $s^{(7)}$ has been determined by the Matlab-script as the solution of a linear system. The small numeric error, $s^{(7)}$ is endued with, gives rise to this imaginary value. Figure 7.4 (7) shows that $\|T_7\|_\infty$ is somewhere in the range between 10^{-20} and 10^{-30} what is very small indeed. Surprisingly, the line representing $\|T_7\|_\infty$ runs very straight and describes no zigzag, like typical roundoff errors, that behave usually rather randomly.

Polynomial test problems (Examples 3 & 4)

As example 3 and 4 we take a polynomial p of degree 6 or 7 respectively solving the following boundary value problem.

$$u(0) = p(0),\ u(1) = p(1), \quad u''(x) = p''(x) \quad \Rightarrow \quad u(x) = p(x)$$

Concretely p has been chosen arbitrarily to be either

$$\begin{aligned} p_6(x) &:= (x+1.00)(x-0.05)(x-0.35)(x-0.89)(x-9.67)(x-15.01) \quad \text{or} \\ p_7(x) &:= -6.67x^7 + 5.89x^6 + 2.48x^5 - 3.46x^4 + 2.69x^3 + 9.71x^2 - 3.45x - 7.32\ . \end{aligned}$$

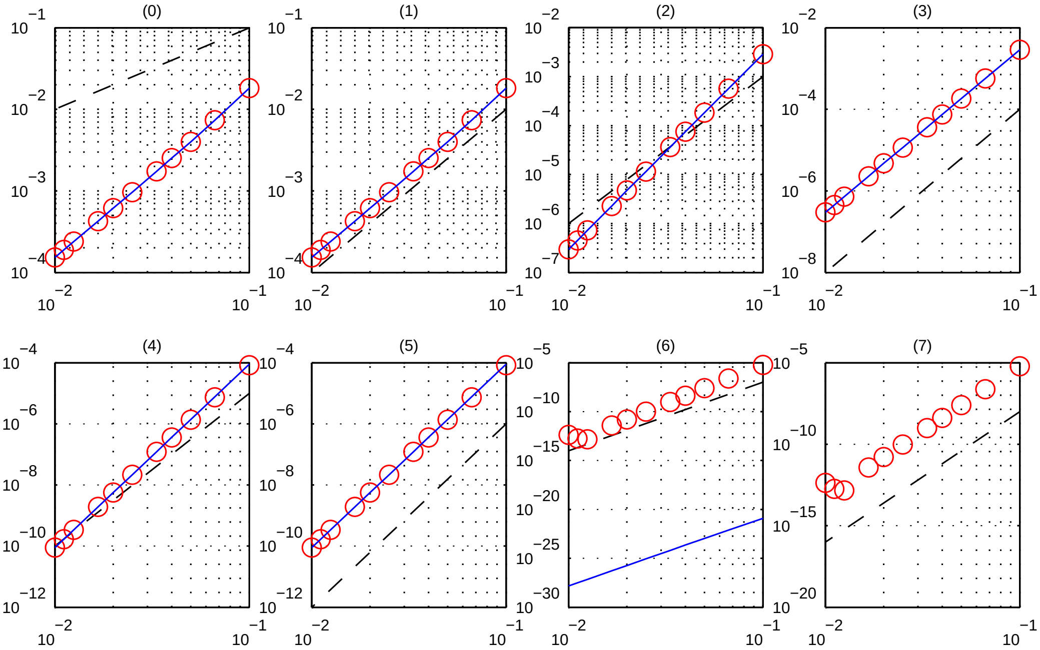


Figure 7.4: Double-logarithmic plot of $\|R_k\|_\infty$ (circles) and $\|T_k\|_\infty$ (lines) for the second example (cosine) using the same grid sequence as for the first example.

Since a polynomial has only finitely many derivatives being different from zero, the numeric solution v has a finite expansion, whenever the source is polynomial. We pay particular attention to the residuals R_6 and R_7, that should vanish in the case of p_6 and p_7 respectively. This provides a good test, whether the coefficients occurring in the expansion have been determined correctly.

Conclusion

We have seen how the asymptotic order functions can be visualized and how they build up the numeric solution v. Furthermore we have tried to encourage the awareness for the two basic kinds of grid functions, that represent different error sources of the scheme: irregular, pure grid functions (boundary layers) generated by the truncation error in singular nodes adjacent to the boundary; and regular, sampled grid functions produced by the truncation error at the bulk nodes.
The interpretation of diagrams should be done with care, because computers are usually restricted with respect to their computational precision. This can provoke artificial, misleading effects.

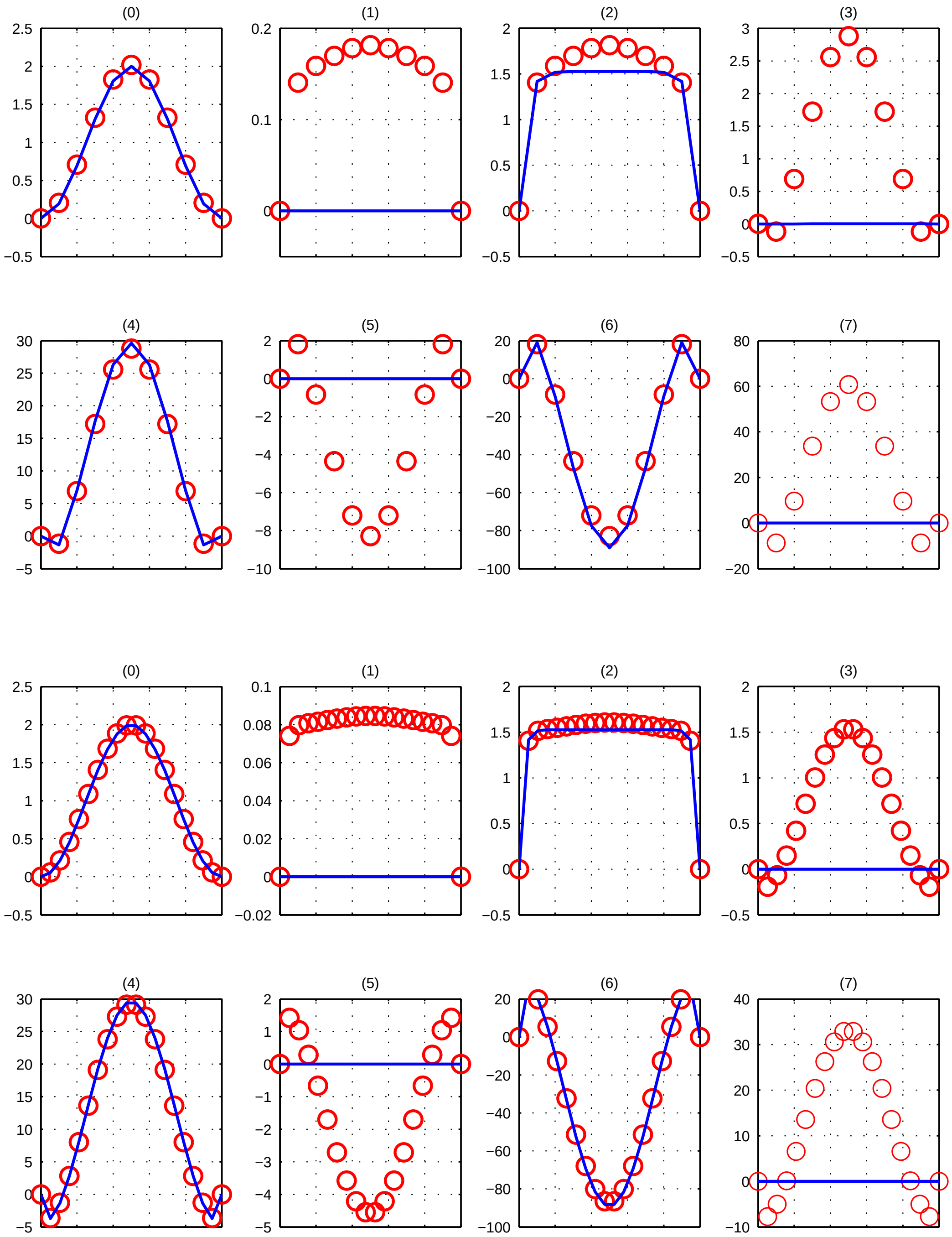


Figure 7.5: Asymptotic order function for example 2 on a grid with 11 (above) and 20 (below) nodes (cf. table 7.2). In contrast to example 1 (figure 7.1,7.2), the odd order functions vanish, while boundary layers and regular parts occur together in the even orders. Note also that the amplitude of the order functions increases much stronger than in example 1. As we have taken the cosine over its full period, differentiating yields an additional factor 2.

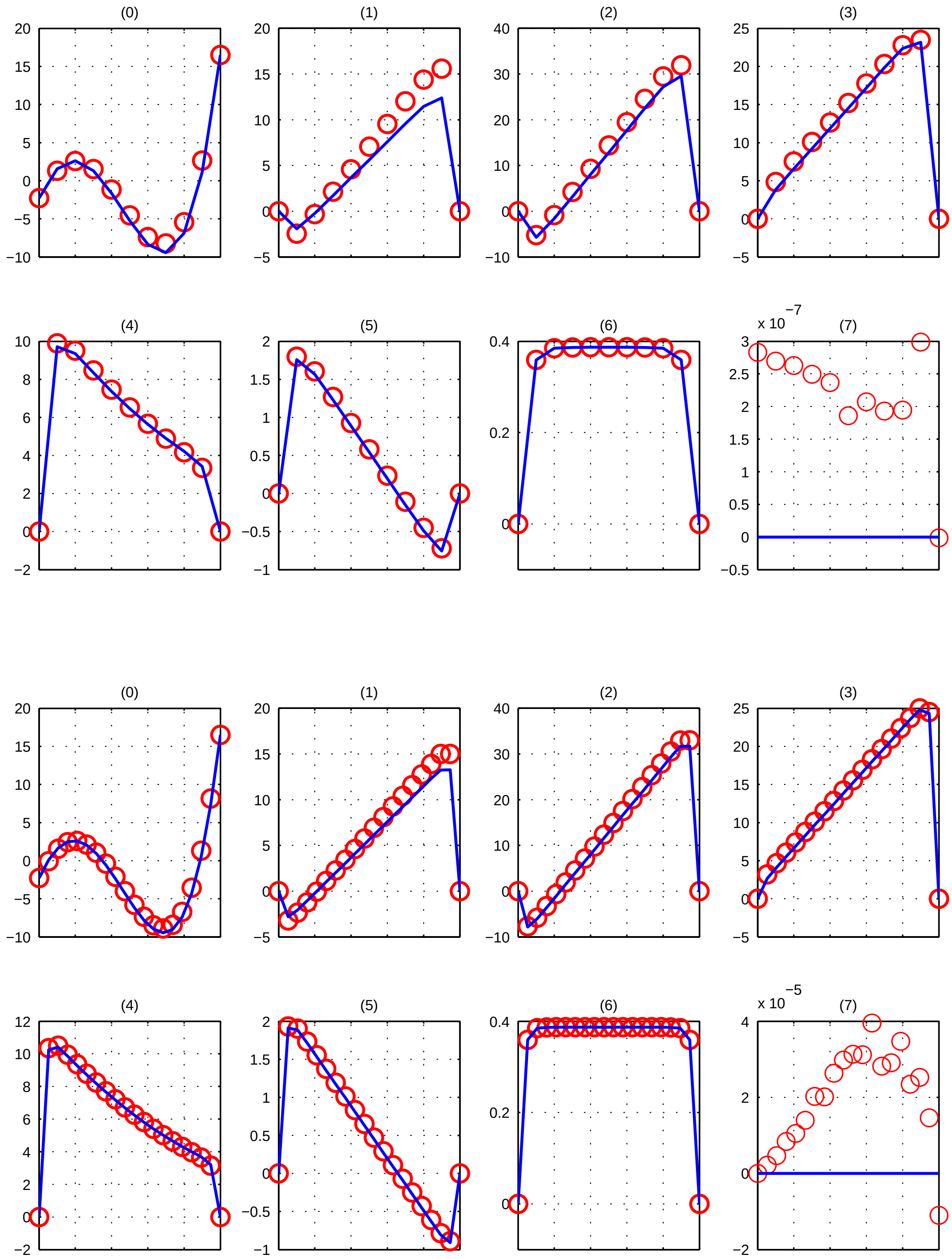


Figure 7.6: Asymptotic order functions for example 3 on a grid with 11 (above) and 20 (below) nodes (cf. table 7.2). As the expansion terminates behind the sixth asymptotic order we have $\frac{1}{h^6}D_5 = s^{(6)}$. Therefore the circles in (6) must coincide with the line. Consequently, in subfigure (7) the circles should be located on the zero line. The observed deviations of the order 10^{-7} and 10^{-5} are probably due to inaccuracies of floating point operations. In particular the linear equations defining the numeric solution v and the boundary layers are solved with a fixed accuracy. This error is amplified when divided by a power of h, what is done to compute the rescaled deviation. The amplification gets the stronger the smaller the grid spacing h becomes.

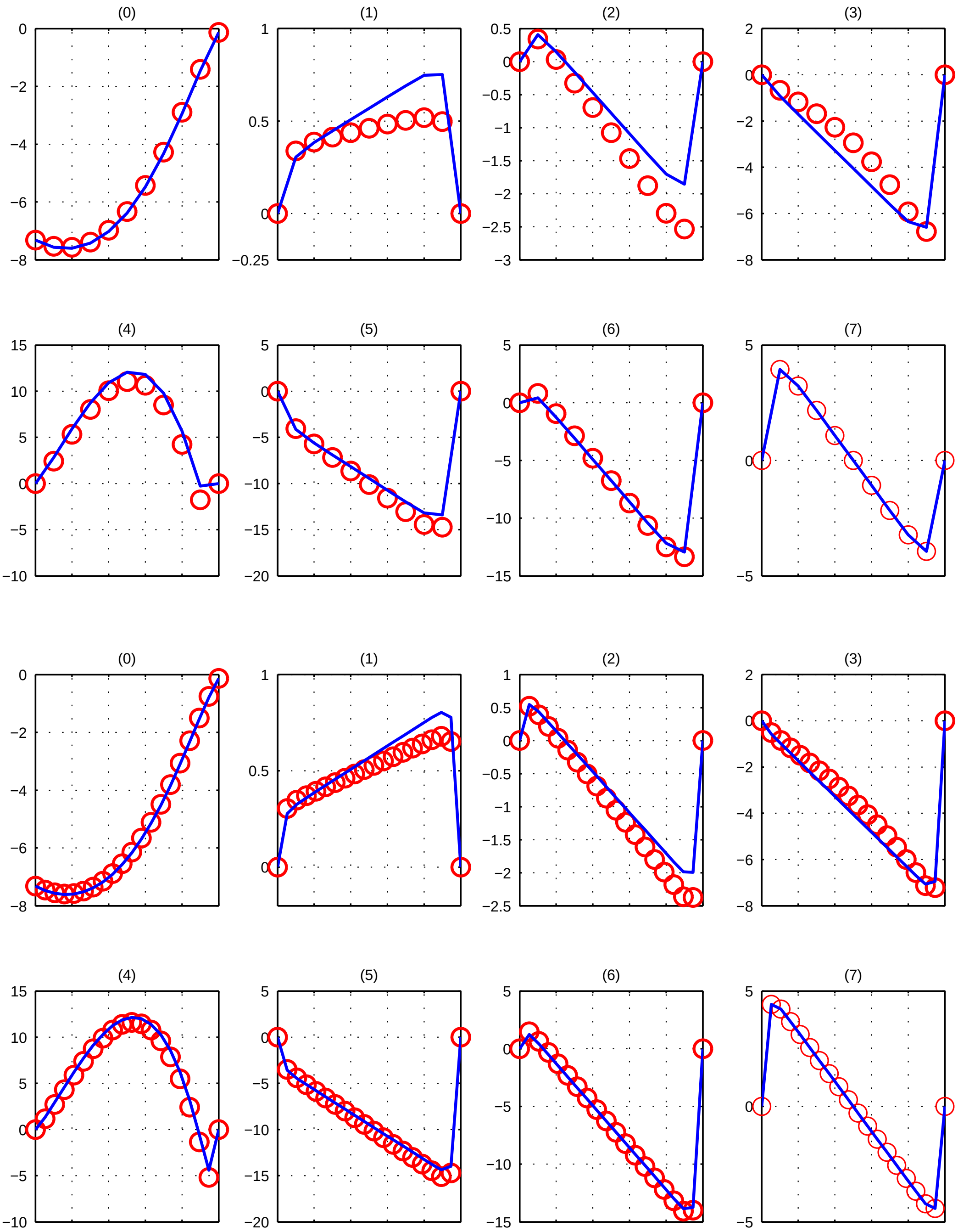


Figure 7.7: Asymptotic orders for example 4 on a grid with 11 (above) and 20 (below) nodes (cf. table 7.2). In subfigure (7) the circles lie quasi exactly on the line. This is checked by computing the deviation D_7. For the first grid we obtain $\|D_7\|_\infty = 2.475875 \cdot 10^{-12}$, whereas $\|D_7\|_\infty = 4.761316e \cdot 10^{-14}$ in the case of the finer, second grid.

Appendix A

Appendix

A.1 Eigenvalues of the advection-diffusion operator

The results of the following eigenvalue computations are used for the purpose of comparison in section 6.2.3.

Non-periodic boundary conditions

Let us consider the *advection-diffusion-reaction* equation in one space dimension

$$\partial_t v \; + \; a\partial_x v \; - \; \nu\partial_x^2 v \; + \; cv \; = \; 0\,, \tag{A.1}$$

with the constants $\nu > 0, a \in \mathbb{R}, c \geq 0$. The unknown $v : \mathbb{R}_0^+ \times [0,1] \to \mathbb{R}$ is a function of the time $t \in \mathbb{R}_0^+$ and of the space $x \in [0,1]$. One of the following boundary conditions (BC) should be imposed:

$$\begin{array}{lr} \text{D) homogeneous Dirichlet BC} \quad : & v(t, x \in \{0,1\}) \; = \; 0 \\ \text{N) homogeneous Neumann BC} \quad : & \partial_x v(t, x \in \{0,1\}) \; = \; 0 \\ \text{R) homogeneous Robin-type BC:} & \big[a - \nu\partial_x\big] v(t, x \in \{0,1\}) \; = \; 0 \end{array}$$

The standard time separation ansatz $v(t,x) = \mathrm{e}^{-\lambda t} u(x)$ leads to the eigenvalue problem of the advection-diffusion-reaction operator in the following function spaces:

$$\begin{aligned} \mathcal{C}_D \; &:= \; \Big\{ w \in \mathcal{C}^2\big([0,1], \mathbb{R}\big) : \qquad\qquad w(x \in \{0,1\}) \; = \; 0 \Big\} \\ \mathcal{C}_N \; &:= \; \Big\{ w \in \mathcal{C}^2\big([0,1], \mathbb{R}\big) : \qquad\quad \partial_x w(x \in \{0,1\}) \; = \; 0 \Big\} \\ \mathcal{C}_R \; &:= \; \Big\{ w \in \mathcal{C}^2\big([0,1], \mathbb{R}\big) : \; \big[a - \nu\partial_x\big]\, w(x \in \{0,1\}) \; = \; 0 \Big\} \end{aligned}$$

So we ask for all $\lambda \in \mathbb{C}$, such that there exists a non-trivial function $u \not\equiv 0$, which belongs to one of the above function spaces and satisfies the subsequent equation.

$$\big(-\nu\,\partial_x^2 \; + \; a\,\partial_x \; + \; c\big)\; u \; = \; \lambda\, u \quad \text{with} \quad \left\{ \begin{array}{ll} u \in \mathcal{C}_D & (D) \\ u \in \mathcal{C}_N & (N) \\ u \in \mathcal{C}_R & (R) \end{array} \right. \tag{A.2}$$

The following theorem provides the answer.

Theorem A.1. *Each of the three eigenvalue problems (D),(N),(R) admits the following eigenvalues being all simple:*

$$\lambda_n = \frac{a^2}{4\nu} \; + \; \pi^2 n^2 \nu \; + \; c, \qquad n \in \mathbb{N}. \tag{A.3}$$

The corresponding eigenfunctions are given by

D) $u_D^n(x) \;=\; \exp(\frac{a}{2\nu}x)\; \sin(\pi n\, x)$

N) $u_N^n(x) \;=\; \exp(\frac{a}{2\nu}x)\; \big[\pi n\, \cos(\pi n\, x) \;-\; \frac{a}{2\nu}\, \sin(\pi n\, x)\big]$

R) $u_R^n(x) \;=\; \exp(\frac{a}{2\nu}x)\; \big[\pi n\, \cos(\pi n\, x) \;+\; \frac{a}{2\nu}\, \sin(\pi n\, x)\big]$

Moreover, $\lambda_0 = c$ *is an additional simple eigenvalue in the case of (N) belonging to the eigenfunction* $u_N^0(x) = 1$ *and in the case of (R) belonging to the eigenfunction* $u_R^0(x) = \mathrm{e}^{\frac{a}{\nu}x}$*. There are no further eigenvalues.*

Proof: The characteristic polynomial of the differential equation (A.2) is $-\nu\zeta^2 + a\zeta + c = \lambda$. Its roots are

$$\zeta_{1,2} \;=\; \tfrac{a}{2\nu} \;\mp\; \sqrt{\tfrac{a^2}{4\nu^2} - \tfrac{\lambda}{\nu} + \tfrac{c}{\nu}} \;=\; \beta \mp \omega \;\; . \tag{A.4}$$

where we have introduced the abbreviations

$$\beta := \tfrac{a}{2\nu} \qquad \omega := \sqrt{\tfrac{a^2}{4\nu^2} - \tfrac{\lambda}{\nu} + \tfrac{c}{\nu}}\,. \tag{A.5}$$

1st case: Let us suppose, that the roots are not equal. According to the theory of ordinary linear differential equations with constant coefficients, a *fundamental system* is given by $f_1(x) := \mathrm{e}^{\zeta_1 x} = \mathrm{e}^{(\beta-\omega)x}$, and $f_2(x) := \mathrm{e}^{\zeta_2 x} = \mathrm{e}^{(\beta+\omega)x}$. Any solution u of the differential equation (A.2) can be written as a linear combination of f_1, f_2.

$$u(x) \;=\; \kappa_1\, f_1(x) \;+\; \kappa_2\, f_2(x) \;=\; \kappa_1\, \mathrm{e}^{(\beta-\omega)x} \;+\; \kappa_2\, \mathrm{e}^{(\beta+\omega)x}$$

We want to choose the coefficients $\kappa_1, \kappa_2 \in \mathbb{C}$ in such a way, that the boundary conditions are satisfied. Since the boundary conditions are homogeneous, the resulting linear system for κ_1, κ_2 is also homogeneous. In order to permit a non-trivial solution u, the system matrix must have a non-trivial kernel. Otherwise both of the coefficients would vanish. Therefore the determinant of the system matrix must vanish and this delivers the determining equation for the admissible values that the eigenvalue λ can attain.

D) Dirichlet BC at $\xi \in \{0,1\}$: $\quad u(\xi) \;=\; 0 \qquad \Leftrightarrow$

$$\kappa_1 \mathrm{e}^{(\beta-\omega)\xi} \;+\; \kappa_2 \mathrm{e}^{(\beta+\omega)\xi} \;=\; 0 \qquad \Leftrightarrow$$

$$\begin{pmatrix} 1 & 1 \\ \mathrm{e}^{\beta-\omega} & \mathrm{e}^{\beta+\omega} \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\det \begin{pmatrix} 1 & 1 \\ \mathrm{e}^{\beta-\omega} & \mathrm{e}^{\beta+\omega} \end{pmatrix} \;=\; \mathrm{e}^{\beta+\omega} \;-\; \mathrm{e}^{\beta-\omega} \;=\; 0 \quad \Leftrightarrow \quad \boxed{\mathrm{e}^{2\omega} = 1}$$

N) Neumann BC at $\xi \in \{0,1\}$: $\quad \partial_x u(\xi) \;=\; 0 \qquad \Leftrightarrow$

$$\kappa_1(\beta-\omega)\,\mathrm{e}^{(\beta-\omega)\xi} \;+\; \kappa_2(\beta+\omega)\,\mathrm{e}^{(\beta+\omega)\xi} \;=\; 0 \qquad \Leftrightarrow$$

$$\begin{pmatrix} (\beta-\omega) & (\beta+\omega) \\ (\beta-\omega)\mathrm{e}^{\beta-\omega} & (\beta+\omega)\mathrm{e}^{\beta+\omega} \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\det \begin{pmatrix} (\beta-\omega) & (\beta+\omega) \\ (\beta-\omega)\mathrm{e}^{\beta-\omega} & (\beta+\omega)\mathrm{e}^{\beta+\omega} \end{pmatrix} \;=\; (\beta^2-\omega^2)(\mathrm{e}^{\beta+\omega} - \mathrm{e}^{\beta-\omega}) \;=\; 0$$

$$\Leftrightarrow \qquad \boxed{\omega^2 = \beta^2} \quad \text{or} \quad \boxed{\mathrm{e}^{2\omega} = 1}$$

R) Robin-type BC at $\xi \in \{0,1\}$: $\;a\,u(\xi) \;-\; \nu\,\partial_x u(\xi) \;=\; 0$
Dividing by ν and using $\frac{a}{\nu} = 2\beta$ yields: $\;2\beta\,u(\xi) \;-\; \partial_x u(\xi) \;=\; 0 \quad \Leftrightarrow$

$$\kappa_1\big[2\beta f_1(\xi) - \partial_x f_1(\xi)\big] \;+\; \kappa_2\big[2\beta f_2(\xi) - \partial_x f_2(\xi)\big] \;=\; 0 \qquad \Leftrightarrow$$

$$\begin{pmatrix} 2\beta-(\beta-\omega) & 2\beta-(\beta+\omega) \\ \big[2\beta-(\beta-\omega)\big]\mathrm{e}^{\beta-\omega} & \big[2\beta-(\beta+\omega)\big]\mathrm{e}^{\beta+\omega} \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\det \begin{pmatrix} \beta+\omega & \beta-\omega \\ (\beta+\omega)\,\mathrm{e}^{\beta-\omega} & (\beta-\omega)\,\mathrm{e}^{\beta+\omega} \end{pmatrix} \;=\; (\beta^2-\omega^2)(\mathrm{e}^{\beta+\omega} - \mathrm{e}^{\beta-\omega}) \;=\; 0$$

$$\Leftrightarrow \qquad \boxed{\omega^2 = \beta^2} \quad \text{or} \quad \boxed{\mathrm{e}^{2\omega} = 1}$$

All three cases lead to the condition $\mathrm{e}^{2\omega} = 1$, that is equivalent to $\omega = \mathrm{i}\pi n$ with $n \in \mathbb{Z}\setminus\{0\}$. The case $n=0$ is excluded, since we assume $\zeta_1 \neq \zeta_2$ implying $\omega \neq 0$. Solving

$$\omega = \sqrt{\tfrac{a^2}{4\nu^2} - \tfrac{\lambda}{\nu} + \tfrac{c}{\nu}} = \mathrm{i}\pi n$$

for λ yields the claimed formula (A.3). Note, that we get for $\pm n$ the same λ, so we can assume $n \in \mathbb{N}$.

In order to compute the associated eigenfunctions, we have to determine the coefficients κ_1, κ_2. Recall $\mathrm{e}^{\omega} = \mathrm{e}^{\mathrm{i}\pi n} = (-1)^n$.

D) $$\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \ker \begin{pmatrix} 1 & 1 \\ (-1)^n\mathrm{e}^{\beta} & (-1)^n\mathrm{e}^{\beta} \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \operatorname{span}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

N) $$\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \ker \begin{pmatrix} (\beta-\omega) & (\beta+\omega) \\ (\beta-\omega)(-1)^n\mathrm{e}^{\beta} & (\beta+\omega)(-1)^n\mathrm{e}^{\beta} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \operatorname{span}\begin{pmatrix} \omega+\beta \\ \omega-\beta \end{pmatrix}$$

R) $$\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \ker \begin{pmatrix} (\beta+\omega) & (\beta-\omega_n) \\ (\beta+\omega)(-1)^n\mathrm{e}^{\beta} & (\beta-\omega)(-1)^n\mathrm{e}^{\beta} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \operatorname{span}\begin{pmatrix} \omega-\beta \\ \omega+\beta \end{pmatrix}$$

Thus we obtain by resubstituting: $\zeta_{1,2} = \beta \mp \omega,\ \beta = \frac{a}{2\nu},\ \omega = \mathrm{i}\pi n.$

$$\begin{aligned} u_D(x) \;&\propto\; \mathrm{e}^{\zeta_1 x} \;-\; \mathrm{e}^{\zeta_2 x} \;=\; \mathrm{e}^{\beta x}\left[\mathrm{e}^{-\omega x} \;-\; \mathrm{e}^{\omega x}\right] \\ &=\; -\mathrm{e}^{\beta x}\left[\mathrm{e}^{\omega x} - \mathrm{e}^{-\omega x}\right] \;=\; -2\mathrm{i}\,\mathrm{e}^{\beta x}\,\sin(\pi n x) \\ &=\; -2\mathrm{i}\,\exp(\tfrac{a}{2\nu}x)\,\sin(\pi n x) \end{aligned}$$

$$\begin{aligned}
u_N(x) &\propto (\omega_n + \beta)\,\mathrm{e}^{\zeta_1 x} \;+\; (\omega - \beta)\,\mathrm{e}^{\zeta_2^+ x} \\
&= \mathrm{e}^{\beta x}\left[\,(\omega + \beta)\,\mathrm{e}^{-\omega x} \;+\; (\omega - \beta)\,\mathrm{e}^{\omega x}\,\right] \\
&= \mathrm{e}^{\beta x}\left[\,\omega\,(\mathrm{e}^{\omega x} + \mathrm{e}^{-\omega x}) \;-\; \beta\,(\mathrm{e}^{\omega x} - \mathrm{e}^{\omega x})\,\right] \\
&= \mathrm{e}^{\beta x}\left[\,2\mathrm{i}\pi n\,\cos(\pi n x) \;-\; 2\mathrm{i}\beta\,\sin(\pi n x)\,\right] \\
&= 2\mathrm{i}\,\exp(\tfrac{a}{2\nu}x)\left[\,\pi n\,\cos(\pi n x) - \tfrac{a}{2\nu}\,\sin(\pi n x)\right]
\end{aligned}$$

$$\begin{aligned}
u_R(x) &\propto (\omega_n - \beta)\,\mathrm{e}^{\zeta_1 x} \;+\; (\omega + \beta)\,\mathrm{e}^{\zeta_2} \\
&= \mathrm{e}^{\beta x}\left[\,(\omega\beta)\,\mathrm{e}^{-\omega x} \;+\; (\omega + \beta)\,\mathrm{e}^{\omega x}\,\right] \\
&= \mathrm{e}^{\beta x}\left[\,\omega\,(\mathrm{e}^{\omega x} + \mathrm{e}^{-\omega x}) \;+\; \beta\,(\mathrm{e}^{\omega x} - \mathrm{e}^{-\omega x})\,\right] \\
&= \mathrm{e}^{\beta x}\left[\,2\mathrm{i}\pi n\,\cos(\pi n x) \;+\; 2\mathrm{i}\beta\,\sin(\pi n x)\,\right] \\
&= 2\mathrm{i}\,\exp(\tfrac{a}{2\nu}x)\left[\,\pi n\,\cos(\pi n x) + \tfrac{a}{2\nu}\,\sin(\pi n x)\right]
\end{aligned}$$

So we recover the claimed eigenfunctions that are uniquely determined up to a scalar factor.

In the Neumann and Robin case we additionally obtain the condition $\omega^2 = b^2$ that transforms into $\lambda = c$ thus implying $\omega = \beta$ and $\zeta_1 = 0$, $\zeta_2 = 2\beta$. The coefficients for the Neumann case are then determined by

$$\begin{pmatrix} 0 & 2\beta \\ 0 & 2\beta\,\mathrm{e}^{2\beta} \end{pmatrix}\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad .$$

Obviously $\kappa_2 = 0$ while we may choose $\kappa_1 = 1$. So we find as eigenfunction $\kappa_1\mathrm{e}^{\zeta_1 x} + \kappa_1\mathrm{e}^{\zeta_2 x} = \mathrm{e}^{0x} = 1$, which turns out to be u_N^0. For the Robin case the coefficients must satisfy

$$\begin{pmatrix} 2\beta & 0 \\ 2\beta\,\mathrm{e}^{2\beta} & 0 \end{pmatrix}\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad .$$

This time we are in the inverse situation. So we obtain with $\kappa_1 = 0$ and $\kappa_2 = 1$ the following eigenfunction $\kappa_1\mathrm{e}^{\zeta_1 x} + \kappa_2\mathrm{e}^{\zeta_2 x} = \mathrm{e}^{2\beta x} = \mathrm{e}^{\frac{a}{\nu}x}$, which is exactly u_R^0.

2nd case: Let us now assume $\omega = 0 \Leftrightarrow \lambda = \frac{a^2}{4\nu} + c$. Then we have only one root ζ of the characteristic polynomial. A fundamental system is given by $f_1(x) := \mathrm{e}^{\zeta x}$ and $f_2(x) := x\mathrm{e}^{\zeta x}$.

D) Dirichlet BC at $\xi \in \{0, 1\}$:

$$\kappa_1 f_1(\xi) + \kappa_2 f_2(\xi) = 0 \quad\Leftrightarrow\quad \kappa_1\mathrm{e}^{\beta\xi} + \kappa_2\xi\mathrm{e}^{\beta\xi} = 0 \quad\Leftrightarrow$$

$$\begin{pmatrix} 1 & 0 \\ \mathrm{e}^{\beta} & \mathrm{e}^{\beta} \end{pmatrix}\begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \det\begin{pmatrix} 1 & 0 \\ \mathrm{e}^{\beta} & \mathrm{e}^{\beta} \end{pmatrix} = \mathrm{e}^{\beta} \neq 0$$

Since the system matrix is regular, there remains only the trivial solution for κ_1, κ_2. But this means, that the equation (A.2) has for $\lambda = \frac{a^2}{4\nu} + c$ only the trivial solution in $\mathcal{C}_D$ and therefore $\frac{a^2}{4\nu} + c$ is no eigenvalue.

N) Neumann BC at $\xi \in \{0,1\}$:

$$\kappa_1 f_1'(\xi) + \kappa_2 f_2'(\xi) = 0 \quad \Leftrightarrow \quad \kappa_1 \beta \mathrm{e}^{\beta\xi} + \kappa_2(\mathrm{e}^{\beta\xi} + x\beta\mathrm{e}^{\beta\xi}) \quad \Leftrightarrow$$

$$\begin{pmatrix} \beta & 1 \\ \beta\mathrm{e}^{\beta} & \mathrm{e}^{\beta} + \beta\mathrm{e}^{\beta} \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \det \begin{pmatrix} \beta & 1 \\ \beta\mathrm{e}^{\beta} & \mathrm{e}^{\beta} + \beta\mathrm{e}^{\beta} \end{pmatrix} = \beta^2 \mathrm{e}^{\beta}$$

Obviously, the system matrix is regular except for $\beta = 0$. Therefore an eigenvalue can only be found in this case, implying $a = 0$ and $\lambda = c$. The associated fundamental system is $f_1(x) = 1$ and $f_2(x) = x$. The coefficients have to satisfy the equation $\left(\begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} \kappa_1 \\ \kappa_2 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$. So $\kappa_2 = 0$ while we can choose $\kappa_1 = 1$. Hence we find again the constant function as eigenfunction of the eigenvalue c.

R) Robin-type BC at $\xi \in \{0,1\}$:

$$\begin{aligned}
&\kappa_1\big[af_1(\xi) - \nu f_1'(\xi)\big] \;+\; \kappa_2\big[af_2(\xi) - \nu f_2'(\xi)\big] \;=\; 0 \quad \Leftrightarrow \\
&\kappa_1\big[2\beta f_1(\xi) - f_1'(\xi)\big] \;+\; \kappa_2\big[2\beta f_2(\xi) - f_2'(\xi)\big] \;=\; 0 \quad \Leftrightarrow \\
&\kappa_1\big[2\beta\mathrm{e}^{\beta\xi} - \beta\mathrm{e}^{\beta\xi}\big] \;-\; \big[2\beta\xi\mathrm{e}^{\beta\xi} \;-\; \mathrm{e}^{\beta\xi} - \xi\beta\mathrm{e}^{\beta\xi}\big] \;=\; 0 \quad \Leftrightarrow
\end{aligned}$$

$$\begin{pmatrix} \beta & 1 \\ \beta\mathrm{e}^{\beta} & \mathrm{e}^{\beta} - \beta\mathrm{e}^{\beta} \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \det \begin{pmatrix} \beta & 1 \\ \beta\mathrm{e}^{\beta} & \mathrm{e}^{\beta} - \beta\mathrm{e}^{\beta} \end{pmatrix} = -\beta^2 \mathrm{e}^{\beta}$$

Like above, the determinant of the system matrix vanishes only for $\beta = 0$. In this case the Robin-type boundary conditions reduce to the Neumann boundary conditions. But this is in agreement with the fact that $u_R^0 = \mathrm{e}^{\frac{a}{\nu}x}$ turns into $u_N^0 = 1$ for $a = 0$.

Since we have not found any new eigenvalues in the second case and since there is no other case to be considered, all eigenvalues have already been captured in the first case. ■

Remark A.1. (Interpretation of Robin condition) The quantity $av - \nu\partial_x v$ is called *flux*. If $c = 0$ (no production or destruction of v) and if the flux is equal at the left and right boundary, then the integral of v remains constant:

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}t}\int_0^1 v(t,x)\,\mathrm{d}x \;&=\; \int_0^1 \partial_t v\,\mathrm{d}x \;=\; \int_0^1 \big(\nu\partial_x^2 v - a\partial_x v\big)\,\mathrm{d}x \\
&=\; \big[\nu\partial_x v(t,1) - av(t,1)\big] \;-\; \big[\nu\partial_x v(t,0) - av(t,0)\big] \\
&=\; 0 \quad \text{due to the assumed equality of the fluxes}
\end{aligned}$$

In particular, equality is provided if the flux is required to vanish at the boundary points by imposing the homogeneous Robin condition. Associating with v a concentration for instance, this boundary condition models an impermeable wall preventing both outflow and inflow. Therefore this boundary condition is also referred to as the *natural* boundary condition.

Periodic boundary conditions

Let us now treat the case of periodic boundary conditions. For this the following function space is introduced.

$$\mathcal{C}_P := \Big\{ w \in \mathcal{C}^2\big([0,1],\mathbb{R}\big) : \partial_x^j w(0) = \partial_x^j w(1) \;\text{ for } j = 0,\,1,\,2 \Big\} \tag{A.6}$$

One should be aware, that not only w but also its first and second derivative are required to be continuous periodic functions[1]. A time separation ansatz leads to the subsequent eigenvalue problem (P):

$$\big(-\nu\,\partial_x^2 \;+\; a\,\partial_x \;+\; c\big)\,u \;=\; \lambda\,u \quad \text{with} \quad u \in \mathcal{C}_P \quad . \tag{A.7}$$

Analogously to theorem A.1 we obtain:

Theorem A.2. *The eigenvalue problem (P) has the following* complex-valued*, simple eigenvalues:*

$$\lambda_n \;=\; 4\pi^2 n^2 \nu \;+\; 2\mathrm{i}\pi n a \;+\; c \qquad n \in \mathbb{Z}. \tag{A.8}$$

The associated eigenfunctions are given by $u_P^n(x) = \mathrm{e}^{2\mathrm{i}\pi n x}$.

Proof: As the characteristic polynomial is not affected by the boundary conditions, we can use the notation defined in (A.4) and (A.5). First we suppose that the roots of the characteristic polynomial are different i.e. $\omega \neq 0$. Then we obtain the following system for the coefficients κ_1, κ_2.

Periodic BC: $u(0) = u(1)$ and $\partial_x u(0) = \partial_x u(1) \quad \Leftrightarrow$

$$\begin{cases} \kappa_1 \mathrm{e}^{(\beta-\omega)0} + \kappa_2 \mathrm{e}^{(\beta+\omega)0} = \kappa_1 \mathrm{e}^{(\beta-\omega)1} + \kappa_2 \mathrm{e}^{(\beta+\omega)1} \\ \kappa_1(\beta-\omega)\mathrm{e}^{(\beta-\omega)0} + \kappa_2(\beta+\omega)\mathrm{e}^{(\beta+\omega)0} = \kappa_1(\beta-\omega)\mathrm{e}^{(\beta-\omega)1} + \kappa_2(\beta+\omega)\mathrm{e}^{(\beta+\omega)1} \end{cases}$$

$$\Leftrightarrow \quad \begin{pmatrix} 1-\mathrm{e}^{\beta-\omega} & 1-\mathrm{e}^{\beta+\omega} \\ (\beta-\omega)(1-\mathrm{e}^{\beta-\omega}) & (\beta+\omega)(1-\mathrm{e}^{\beta+\omega}) \end{pmatrix} \begin{pmatrix} \kappa^- \\ \kappa^+ \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\det \begin{pmatrix} 1-\mathrm{e}^{\beta-\omega} & 1-\mathrm{e}^{\beta+\omega} \\ (\beta-\omega)(1-\mathrm{e}^{\beta-\omega}) & (\beta+\omega)(1-\mathrm{e}^{\beta+\omega}) \end{pmatrix} \;=\; 2\omega\,(1-\mathrm{e}^{\beta-\omega})(1-\mathrm{e}^{\beta+\omega}) = 0$$

$$\Leftrightarrow \quad \boxed{1-\mathrm{e}^{\beta-\omega}=0} \quad \text{or} \quad \boxed{1-\mathrm{e}^{\beta+\omega}=0} \quad \text{or} \quad \boxed{\omega = 0}$$

1) Let $n \in \mathbb{Z}$: $\mathrm{e}^{\beta-\omega} = 1 \quad \Leftrightarrow \quad \beta - \omega \;=\; 2\mathrm{i}\pi n \quad \Leftrightarrow \quad \lambda \;=\; 4\pi^2 n^2 \nu + 2\mathrm{i}\pi n a + c$

   Note, that $n = 0$ is only admissible if assuming $\beta \neq 0 \;\Leftrightarrow\; a \neq 0$.
   Otherwise $0 - \omega = 2\mathrm{i}\pi 0 \;\Rightarrow\; \omega = 0$, yielding the excluded double zero.
   Computing the kernel of the system matrix for $1 - \mathrm{e}^{\beta-\omega} = 0$, leads to the eigenfunctions:

$$\begin{pmatrix} 0 & 1-\mathrm{e}^{\beta+\omega} \\ 0 & (\beta+\omega)(1-\mathrm{e}^{\beta+\omega}) \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \operatorname{span} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\Rightarrow \quad u_1^n(x) \;=\; \mathrm{e}^{\zeta_1 x} \;=\; \mathrm{e}^{(\beta-\omega)x} \;=\; \mathrm{e}^{[\beta-(\beta-2\mathrm{i}\pi n)]x} \;=\; \mathrm{e}^{2\mathrm{i}\pi n x}$$

[1]The periodic boundary conditions suggest, that we treat 0 and 1 practically as the same point just as if we bend the interval $[0, 1]$ to a circle and glue it together at the endpoints $0, 1$. Mathematically periodic boundary conditions are obtained by considering the quotient space of $\mathbb{R}$ with respect to 1, abbreviated by $\mathbb{R}/1 =: \mathbb{T}^1$. This manifold is frequently referred to as the one-dimensional torus being equivalent to the one-dimensional sphere S^1. Instead of defining $\mathcal{C}_P$ where periodic boundary conditions must be specified explicitly, one could also write $\mathcal{C}^2(\mathbb{T}^1, \mathbb{R})$.

2) Let $n \in \mathbb{Z}$: $\mathrm{e}^{\beta+\omega} = 1 \quad \Leftrightarrow \quad \beta + \omega \ = \ 2\mathrm{i}\pi n \quad \Leftrightarrow \quad \lambda \ = \ 4\pi^2 n^2 \nu - 2\mathrm{i}\pi n a + c$

   Again $n = 0$ is only admissible, if we assume $\beta \neq 0 \ \Leftrightarrow \ a \neq 0$.
   Computing the kernel of the system matrix for $1 - \mathrm{e}^{\beta+\omega} = 0$, leads to the eigenfunctions:

$$\begin{pmatrix} 1 - \mathrm{e}^{\beta-\omega} & 0 \\ (\beta-\omega)(1-\mathrm{e}^{\beta-\omega}) & 0 \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} \in \mathrm{span} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\Rightarrow \quad u_2^n(x) \ = \ \mathrm{e}^{\zeta_2 x} \ = \ \mathrm{e}^{(\beta+\omega)x} \ = \ \mathrm{e}^{[\beta+(-\beta+2\mathrm{i}\pi n)]x} \ = \ \mathrm{e}^{2\mathrm{i}\pi n x}$$

Obviously case 1) and 2) lead to the same eigenvalues and associated eigenfunctions, which correspond to those given by the theorem. For $n = 0$ we obtain the constant function as eigenfunction; note that this is also true if $a = 0$, what we had to exclude in order to avoid $\omega = 0$.
The special situation with the characteristic polynomial having a double root, i.e. $\zeta_1 = \zeta_2 \ \Leftrightarrow \ \omega = 0$ has to be treated separately with a different fundamental system. However, as in the proof of the previous theorem, no additional eigenvalues or eigenfunctions are found. ■

Remark A.2. (Degeneracy for $a = 0$) It should be noticed that the spectrum is real if $a = 0$, since the operator becomes then symmetric. In this case the eigenvalues are equal for n and $-n$, $n \in \mathbb{N}$. With exception of $n = 0$, all eigenvalues are two-fold degenerated with the $x \mapsto \cos(2\pi n x)$ and $x \mapsto \sin(2\pi n x)$ spanning the eigenspace. As soon as $a \neq 0$ the degeneracy disappears; the eigenvalues become complex-valued and emerge for $n \geq 1$ in complex conjugated pairs.

Remark A.3. (Elimination of advection by Galilean transformations) Real-valued solutions of the advection diffusion equation with periodic boundary conditions are obtained in the following manner:

$$\begin{aligned} v^m(t,x) \ &:= \ \tfrac{1}{2}\Big[\mathrm{e}^{-\lambda_m t}\, u_P^m(x) \ + \ \mathrm{e}^{-\lambda_{-m} t}\, u_P^{-m}(x)\Big] \\ &= \ \tfrac{1}{2}\Big[\mathrm{e}^{-(4\pi^2 m^2 \nu + 2\mathrm{i}\pi m a + c)t}\, \mathrm{e}^{2\mathrm{i}\pi m x} \ + \ \mathrm{e}^{-(4\pi^2 m^2 \nu - 2\mathrm{i}\pi m a + c)t}\, \mathrm{e}^{-2\mathrm{i}\pi m x}\Big] \\ &= \ \mathrm{e}^{-(4\pi^2 m^2 \nu + c)t}\, \tfrac{1}{2}\Big[\mathrm{e}^{2\mathrm{i}\pi m(x-at)} \ + \ \mathrm{e}^{-2\mathrm{i}\pi m(x-at)}\Big] \\ &= \ \mathrm{e}^{-(4\pi^2 m^2 \nu + c)t}\, \cos\big(2\pi m(x-at)\big) \\ w^m(t,x) \ &:= \ \tfrac{1}{2\mathrm{i}}\Big[\mathrm{e}^{-\lambda_m t}\, u_P^m(x) \ - \ \mathrm{e}^{-\lambda_{-m} t}\, u_P^{-m}(x)\Big] \\ &= \ \mathrm{e}^{-(4\pi^2 m^2 \nu + c)t}\, \sin\big(2\pi m(x-at)\big) \end{aligned}$$

The sinusoidal profile of the solution is independent of the advection velocity a, which determines only how fast the profile is shifted to the right ($a > 0$) or left ($a < 0$).
As the unit interval with periodic boundary conditions can be identified with the quotient space $\mathbb{R}/1\mathbb{R}$, the addition of an arbitrary real number represents an invertible transformation (translation) from $\mathbb{R}/1\mathbb{R}$ onto $\mathbb{R}/1\mathbb{R}$. So, the term $a\partial_x$ being responsible for the advection can be eliminated by virtue of the time-dependent transformation $z = x - at$. Formulating the equation with respect to the new space

variable z, ∂_t is replaced by $\partial_t - a\partial_z$ whereas the spatial derivatives do not alter. Therefore $a\partial_z$ is canceled by the transformation of the time derivative. Physically, this means, that the inertial system is changed, such that the original (uniform) advection with velocity a appears as rest.

Concluding remarks

Remark A.4. (Completeness of eigenfunctions) In order to prove alternatively, that there are no other eigenfunctions than those provided by theorem A.1 and A.2, one might argue that the set of eigenfunctions forms a complete basis of a suitable Hilbert space. In particular it has to be shown that the set formed by all possible linear combinations of the given eigenfunctions lies dense in this Hilbert space.

Remark A.5. (Length scaling of the space interval) So far, the eigenvalue problems have been considered on the unit interval $[0,1]$. How are the eigenvalues and eigenfunctions affected if the space interval $[0,1]$ is replaced by an interval $[0,L]$ of arbitrary length $L>0$? Using the transformation $x \mapsto Lx$ mapping $[0,1]$ onto $[0,L]$, we can avoid to repeat the computations of the preceding paragraphs. Composing the eigenfunction $u : [0,1] \to \mathbb{R}$ with the inverse transformation $x \mapsto x/L$ we obtain a function $v : [0,L] \to \mathbb{R}$.

$$v(x) := u(x/L) \quad \Rightarrow \quad v'(x) = \tfrac{1}{L}u'(x/L) \qquad \text{and thus} \quad \left\{ \begin{array}{rcl} \partial_x u &=& L\partial_x v \\ \partial_x^2 u &=& L^2\partial_x^2 v \end{array} \right.$$

Taking into account the relation between the derivatives of u and v, we see that v is also eigenfunction of an advection-diffusion-reaction operator where, however, ν and a are substituted by νL^2 and aL.

$$-\nu\partial_x^2 u + a\partial_x u + cu = \lambda u \quad \Rightarrow \quad -\nu L^2\partial_x^2 v + aL\partial_x v + cv = \lambda v$$

If u satisfies periodic boundary conditions or homogeneous boundary conditions of Dirichlet or Neumann type at 0 and 1, then v fulfills the same conditions at 0 and L. In the case of homogeneous Robin conditions we get a slight modification as it involves the function and its derivative simultaneously.

$$(a - \nu\partial_x)u\big|_{x\in\{0,1\}} = 0 \quad \Rightarrow \quad \left\{ \begin{array}{rr} & (a - \nu L\partial_x)v\big|_{x\in\{0,L\}} = 0 \\ \Leftrightarrow & (aL - \nu L^2\partial_x)v\big|_{x\in\{0,L\}} = 0 \end{array} \right.$$

In order to interpret the new coefficients νL^2 and aL as diffusivity and advection velocity, we introduce

$$\tilde\nu := \nu L^2, \qquad \tilde a := aL.$$

Substituting a and ν by $\tilde a$ and $\tilde\nu$ in the explicit representation of $v_R^n(x) := u_R^n(x/L)$, we get

$$\begin{aligned} v_R^n(x) &= u_R^n(x/L) = \exp(\tfrac{a}{2\nu}x/L)\big[\pi n\cos(\pi nx/L) + \tfrac{a}{2\nu}\sin(\pi nx/L)\big] \\ &= \exp(\tfrac{aL}{2\nu L^2}Lx/L)\big[\pi n\cos(\pi nx/L) + \tfrac{aL}{2\nu L^2}L\sin(\pi nx/L)\big] \\ &= \exp(\tfrac{\tilde a}{2\tilde\nu}x)\big[\pi n\cos(\pi nx/L) + \tfrac{\tilde a}{2\tilde\nu}L\sin(\pi nx/L)\big] \end{aligned}$$

Proceeding similarly with the corresponding eigenvalue yields

$$\lambda_n = \frac{a^2}{4\nu} + \pi^2 n^2 \nu + c = \frac{\tilde{a}^2}{4\tilde{\nu}} + \pi^2 n^2 \tilde{\nu}/L^2 + c =: \tilde{\lambda}_n$$

so that we obtain finally

$$\big(-\tilde{\nu}\partial_x^2 + \tilde{a}\partial_x + c\big) v_R^n = \tilde{\lambda}_n v_R^n \qquad \big(\tilde{a} - \tilde{\nu}\partial_x\big) v_R^n\big|_{x\in\{0,L\}} = 0 \ .$$

Of course, the tildas could be removed now.
Summarizing, we see now how the eigenvalue and eigenfunction of the advection-diffusion-reaction operator $-\nu\partial_x^2 + a\partial_x + c$ (with homogeneous Robin conditions) depend on the length L of the spatial domain:

$$\begin{aligned} \text{eigenvalue:} \quad & \lambda_n = \tfrac{a^2}{4\nu} + \pi^2 n^2 \nu/L^2 + c\,, \\ \text{eigenfunction:} \quad & x \mapsto \exp(\tfrac{a}{2\nu}x)\big[\pi n \cos(\pi n x/L) + \tfrac{a}{2\nu} L \sin(\pi n x/L)\big] \ . \end{aligned}$$

Analogously the other boundary conditions can be handled.

Bibliography

[1] M.K. Banda, W.-A. Yong, and A. Klar. A stability notion for lattice-Boltzmann equations. *SISC*, 27,(6):2098–2111, 2006.

[2] C. Bardos, F. Golse, and D. Levermore. Fluid dynamic limits of kinetic equations 1, formal derivations. *J. Stat. Phys.*, 63:323–344, 1991.

[3] R. Benzi, S. Succi, and M. Vergassola. The lattice-Boltzmann equation: Theory and applications. *Phys. Rep.*, 222:145–197, 1992.

[4] H. Brezis. *Analyse fonctionelle – Théorie et applications.* Collection Mathématiques appliquées pour la maîtrise. Masson, 1983.

[5] A. Caiazzo. Analysis of lattice-Boltzmann initialization routines. *J. Stat. Phys.*, 121:37–48, 2005.

[6] A. Caiazzo. *Asymptotic analysis of lattice-Boltzmann methods for fluid-structure interaction problems.* PhD thesis, Technische Universität Kaiserslautern, Scuola Normale Superiore – Pisa, 2007.

[7] C. Cercignani. *Mathematical Methods in Kinetic Theory.* Plenum Press, 2nd edition, 1990.

[8] C. Cercignani, R. Illner, and M. Pulvirenti. *The Mathematical Theory of Dilute Gases.* Applied Mathematical Sciences 106. Springer, 1994.

[9] H. Chen, S. Chen, and W.H. Matthaeus. Recovery of the Navier-stokes equations using a lattice-gas Boltzmann method. *Phys. Rev. A*, 45:R5339–5342, 1992.

[10] S. Chen and G.D. Doolen. Lattice Boltzmann method for fluid flows. *Annu. Rev. Fluid Mech.*, 30:329–364, 1998.

[11] A. J. Chorin and J. E. Marsden. *A Mathematical Introduction to Fluid Mechanics.* Texts in Applied Mathematics. Springer, 3 edition, 1993.

[12] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations.* Springer, New York, 2002.

[13] D. d'Humières, I. Ginzburg, M. Krafczyk, P. Lallemand, and L.-S. Luo. Multiple-relaxation-time lattice boltzmann models in three dimensions. *Philosophical Transactions of Royal Society of London A*, 360(1972):437–451, 2002.

[14] G.D. Doolen, U. Frisch, B. Hasslacher, S. Orzag, and S. Wolfram, editors. *Lattice-gas methods for partial differential equations.* Addison-Wesley, Redwood City Ca, 1990.

[15] F. Dubois. Equivalent partial differential equations of a lattice-Boltzmann scheme. *ICMMES 05, preprint*, 2005.

[16] B. Elton, C. Levermore, and G. Rodrigue. Convergence of convective diffusive lattice Boltzmann methods. *SIAM J. Numer. Anal.*, 32:1327–1354, 1995.

[17] Lawrence C. Evans. *Partial Differential Equations.* Graduate Studies in Mathematics. American Mathematical Society, 1998.

[18] A. L. Fetter and J. D. Walecka. *Theoretical Mechanics of Particles and Continua.* International Series in Pure and Applied Physics. McGraw-Hill, 1980.

[19] U. Frisch, D. d'Humières, B. Hasslacher, P. Lallemand, Y. Pomeau, and J. P. Rivet. Lattice gas hydrodynamics in two and three dimensions. *Complex sys.*, 1:649–707, 1987.

[20] U. Frisch, B. Hasslacher, and Y. Pomeau. Lattice-gas automata for the Navier-Stokes equation. *Phys. Rev. Letter*, 56:1505, 1986.

[21] M. Gardner. On cellular automata, self-reproduction, the garden of eden and the game of "life". *Sci. Am.*, 224(2):112–117, 1971.

[22] J. Hardy, O. de Pazzis, and Y. Pomeau. Time evolution of a two-dimensional model system.i. invariant states and time correlation functions. *J. Matz. Phys.*, 14(12):1746–1759, 1973.

[23] J. Hardy, O. de Pazzis, and Y. Pomeau. Molecular dynamics of a classical lattice gas: transport properties and time correlation function. *Phys. Rev. A*, 13:1949–1961, 1976.

[24] X. He and L.-S. Luo. A priori derivation of the lattice Boltzmann equation. *Phys. Rev. E*, 55:6333–6336, 1997.

[25] X. He and L.-S. Luo. Lattice Boltzmann model for the incompressible Navier-Stokes equation. *J. Stat. Phys.*, 88:927–944, 1997.

[26] X. He and L.-S. Luo. Theory of the lattice Boltzmann method: From the Boltzmann equation to the lattice Boltzmann equation. *Phys. Rev. E*, 56:6811–6817, 1997.

[27] F. Higuera and J. Jiménez. Boltzmann approach to lattice gas simulations. *Europhys. Lett.*, 9:663, 1989.

[28] D. J. Holdych, D.R. Noble, J.G. Georgiadis, and R.O. Buckius. Truncation error analysis of lattice Boltzmann methods. *J. of Comp. Phys.*, 193:595–619, 2003.

[29] S. Jin and Z. Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Comm. Pure Appli. Math.*, 48:235–277, 1995.

[30] R.R. Nourgaliev T.N. Dinh T.G. Theofanous D. Joseph. The lattice-Boltzmann equation: Theoretical interpretation, numerics and implications. *Int. J. Multiphase Flow*, 29:4823–4842, 2003.

[31] M. Junk. On the construction of discrete equilibrium distributions for kinetic schemes. *ITWM report*, 14, 1999.

[32] M. Junk. A finite difference interpretation of the lattice-Boltzmann method. *Numer. Methods Partial Differ. Equations*, 17:383–402, 2001.

[33] M. Junk and A. Klar. Discretizations for the incompressible Navier-Stokes equations based on the lattice Boltzmann method. *SIAM. J. Sci. Comp.*, 22(1):1–19, 2000.

[34] M. Junk, A. Klar, and L.S. Luo. Asymptotic analysis of the lattice Boltzmann equation. *Journal Comp. Phys.*, 210:676–704, 2005.

[35] M. Junk and M. Rheinländer. Consistency and multiscale analysis of a lattice-Boltzmann method. *ICMMES 06, preprint*, 2006.

[36] M. Junk and Z. Yang. Analysis of lattice Boltzmann boundary conditions. *Proc. Appl. Math. Mech.*, 3:76–79, 2003.

[37] M. Junk and Z. Yang. Asymptotic analysis of finite difference methods. *Appl. Math. Comput.*, 158:267–301, 2004.

[38] M. Junk and Z. Yang. Asymptotic analysis of lattice Boltzmann boundary conditions. *J. Stat. Phys.*, 121:3–35, 2005.

[39] M. Junk and Z. Yang. One-point boundary condition for the lattice Boltzmann model. *Phy. Rev. E*, 72, 2005.

[40] M. Junk and W. A. Yong. Weighted l_2 stability of the lattice-Boltzmann method. *preprint*.

[41] M. Junk and W.-A. Yong. Rigorous navier-stokes limit of the lattice boltzmann equation. *Asymptotic Analysis*, 35:165–185, 2003.

[42] K. Königsberger. *Analysis 1*. Springer, 2 edition, 1992.

[43] P. Lallemand and L.S. Luo. Theory of the lattice-Boltzmann method: dispersion, dissipation, isotropy, Galilean invariance, and stability. *Physical Review E*, 61, 6:6546–6562, 2000.

[44] S. Lang. *Complex Analysis*. Springer, New York, 3rd edition, 1993.

[45] R. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser, Basel, 1992.

[46] L.-S. Luo. Some recent results on discrete velocity models and ramifications for lattice Boltzmann equation. *Computer Phys. Com.*, 129:63–74, 2000.

[47] G.R. McNamara and G. Zanetti. Use of Boltzmann equation to simulate lattice-gas automata. *Phys. Rev. L*, 61:2332, 1988.

[48] J.D. Murray. *Asymptotic Analysis*. Applied Mathematical Sciences. Springer, 1984.

[49] T. Platkowski and R. Illner. Discrete velocity models of the boltzmann equation: A survey on the mathematical aspects of the theory. *SIAM Rev.*, 30 No. 2:213–255, 1988.

[50] R. Plato. *Numerische Mathematik kompakt*. Vieweg, 2000.

[51] Y. Qian, D. d'Humières, and P. Lallemand. Lattice BGK models for Navier-stokes equation. *Europhys. Lett.*, 17:479–484, 1992.

[52] M. Reider and J. Sterling. Accuracy of discrete velocity BGK models for the simulation of the incompressible Navier-Stokes equations. *Computers and Fluids*, 24:459–467, 1995.

[53] Remmert. *Funktionentheorie 1*. Springer, Berlin Heidelberg, 3rd edition, 1992.

[54] M. Rheinländer. A consistent grid coupling method for lattice-Boltzmann schemes. *J. Stat. Phys.*, 121:49–74, 2005.

[55] M. Rheinländer. Stability and multiscale analysis of an advective lattice-boltzmann scheme. *ICMMES 06, preprint*, 2006.

[56] W. Rudin. *Principles of Mathematical Analysis*. McGraw Hill, 3 edition, 1976.

[57] V. Rutka. *Immersed interface methods for elliptic boundary value problems*. PhD thesis, Technische Universität Kaiserslautern, 2005.

[58] Houshang H. Sohrab. *Basic Real Analysis*. Birkhäuser, Boston, 2003.

[59] J. D. Sterling and S. Chen. Stability analysis of the lattice-Boltzmann method. *Journal of computational physics*, 123:196–206, 1996.

[60] Sauro Succi. *The Lattice Boltzmann equation. For Fluid Dynamics and Beyond*. Oxford University Press, 2001.

[61] Bertram Huppert. *Angewandte Lineare Algebra*. Walter de Gruyter, 1990.

[62] J. Weiß. *Numerical analysis of lattice-Boltzmann methods for the Heat equation on a bounded interval*. PhD thesis, Universität Karlsruhe, 2006.

[63] Dieter A. Wolf-Gladrow. *Lattice-Gas Cellular Automata and Lattice Boltzmann Models*. Lecture Notes in Mathematics. Springer, 2000.

[64] S. Wolfram. Statistical mechanics of cellular automata. *Rev. Mod. Phys.*, 55:601–644, 1983.

[65] S. Wolfram. Cellular automaton fluids 1: Basic theory. *J. Stat. Phys.*, 45:471, 1986.

[66] S. Wolfram. *Cellular Automata and Complexity.* Addison-Wesley, Reading, MA, 1994.

[67] Z. Yang. *Analysis of lattice-Boltzmann boundary conditions.* PhD thesis, Universität Konstanz, 2007.

[68] W.A. Yong and L.-S. Luo. Nonexistance of h theorems for the athermal lattice Boltzmann models with polynomial equilibria. *Phys. Rev. E*, 67(5):051105, 2003.

[69] W.A. Yong and L.-S. Luo. Nonexistance of h theorem for some lattice Boltzmann models. *J. Stat. Phys.*, 121(1/2):91–103, 2005.