On the dynamical low-rank numerical method for kinetic equations

– Stability analysis and application to inverse problems –

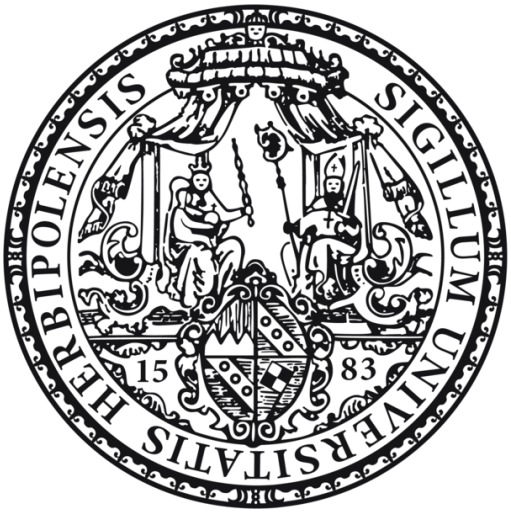
– Doctoral Thesis –

submitted by

Lena Baumann

under the supervision of

Prof. Dr. Christian Klingenberg



Julius-Maximilians-Universität Würzburg
Faculty of Mathematics and Computer Science

Würzburg, September 2025

Abstract

The numerical solution of kinetic partial differential equations (PDEs) usually exhibits high computational costs and memory requirements. This problem can be overcome when using numerical reduction techniques such as dynamical low-rank approximation (DLRA). Its main idea consists in representing and evolving the solution to a given equation on a low-rank manifold, thereby splitting up the solution of one high-dimensional problem into lower-dimensional subproblems. Efficient fully discrete DLRA schemes must be carefully constructed in order to account for the underlying structure of the problem and to ensure numerical stability.

The first part of this thesis is devoted to the derivation of stable fully discrete DLRA schemes for different linear PDEs. For the thermal radiative transfer equations (RTEs) with Su-Olson closure, a provably energy stable and mass conservative DLRA algorithm is proposed. For its construction an implicit coupling of particle density and internal energy as well as a rank-adaptive augmented low-rank integrator and a suitable conservative truncation strategy are used. In certain settings, a multiplicative splitting of the kinetic distribution function is advantageous for the construction of an efficient DLRA scheme. We first reconsider the thermal RTEs with Su-Olson closure with a multiplicative splitting of the distribution function, giving rise to additional complexities in the proof of energy stability and mass conservation for the DLRA scheme. In a second step, the gained insights are transferred to the linear Boltzmann-Bhatnagar-Gross-Krook (BGK) equation. Being different in structure, a distinct notion of numerical stability is required and new ideas for basis augmentations and an appropriate truncation strategy are introduced into the mathematically rigorous proof of stability. Various numerical experiments confirm the efficiency and the accuracy of the proposed DLRA schemes and validate the theoretical results.

In the second part of this thesis, the method of DLRA is applied to parameter identification inverse problems. For the reconstruction of the scattering coefficient in the RTE, a PDE constrained optimization problem together with a gradient-based iterative update scheme is formulated. The optimization procedure requires the solution of the forward and the adjoint kinetic equations in each step of the algorithm, rendering numerical computations especially in higher dimensions extremely expensive. For the reduction of computational demands a DLRA approach is applied to the fully discrete forward and adjoint equations. Its efficiency is further enhanced by using an adaptive choice of the optimization step size and of the DLRA truncation tolerance. Numerical test examples underline the applicability of DLRA to inverse problems and confirm the efficiency of the proposed method.

Zusammenfassung

Die numerische Lösung kinetischer partieller Differentialgleichungen (PDEs) erfordert in der Regel umfangreichen Rechenaufwand und hohen Speicherbedarf. Eine Methode, die zur Verringerung des numerischen Aufwandes eingesetzt werden kann, ist das Konzept der dynamischen Niedrigrang-Approximation (DLRA). Dessen Hauptidee besteht darin, die Lösung einer gegebenen Gleichung auf eine Niedrigrang-Mannigfaltigkeit zu projizieren und dort in der Zeit weiterzuentwickeln. Dadurch reduziert sich die Lösung eines hochdimensionalen Problems auf mehrere niedrigdimensionale Teilprobleme. Effiziente, vollständig diskretisierte DLRA Verfahren müssen jedoch sehr sorgfältig konstruiert werden, damit die zugrundeliegende Struktur des Problems berücksichtigt und numerische Stabilität garantiert werden kann.

Der erste Teil dieser Dissertation beschäftigt sich mit der Herleitung stabiler, vollständig diskretisierter DLRA Verfahren für verschiedene lineare PDEs. Zunächst wird ein nachweislich energiestabiler and massenerhaltender DLRA Algorithmus für die thermischen Strahlungstransportgleichungen (RTEs) mit Su-Olson-Abschluss konstruiert. Hierfür sind eine implizite Kopplung von Teilchendichte und innerer Energie sowie ein Rang-adaptiver erweiterter Niedrigrang-Integrator und eine geeignete massenerhaltende Strategie zum Abschneiden der Lösung auf einen bestimmten Rang unerlässlich. Unter gewissen Voraussetzungen kann auch ein multiplikatives Aufspalten der kinetischen Verteilungsfunktion von Vorteil sein, um ein effizientes DLRA Verfahren zu erhalten. Aus diesem Grund werden die thermischen RTEs mit Su-Olson-Abschluss nochmals mit multiplikativer Struktur der Verteilungsfunktion untersucht. Dies führt zu zusätzlichen Herausforderungen im Beweis der Energiestabilität und Massenerhaltung des DLRA Verfahrens. Mithilfe der aus dieser Arbeit gewonnenen Einblicke wird daraufhin die lineare Boltzmann-Bhatnagar-Gross-Krook (BGK)-Gleichung betrachtet. Für diese werden aufgrund ihrer Struktur ein anderer Stabilitätsbegriff und neue Ideen zum Beweis der numerischen Stabilität benötigt. Die Effizienz und Genauigkeit der hergeleiteten DLRA Verfahren sowie die Ergebnisse aus den theoretischen Betrachtungen werden in zahlreichen numerischen Testbeispielen bestätigt.

Im zweiten Teil dieser Dissertation wird die DLRA Methode auf inverse Probleme zur Identifizierung von Modellparametern angewendet. Hierfür betrachten wir das Problem der Rekonstruktion des Streuungskoeffizienten in der RTE, welches als restringiertes Optimierungsproblem formuliert wird und mit einem Gradienten-basierten iterativen Verfahren gelöst werden soll. Jeder Schritt des Verfahrens benötigt sowohl die Lösung der Vorwärtsgleichungen als auch der adjungierten Gleichungen. Dies führt vor allem in höheren Dimensionen zu erheblichem numerischem Aufwand. Um diesen zu reduzieren, wird ein DLRA Ansatz auf die vollständig diskretisierten Vorwärts-Gleichungen und adjungierten Gleichungen angewendet. Dessen Effizienz wird durch eine adaptive Wahl der Optimierungsschrittweite und der Toleranz zum Abschneiden der DLRA Lösung auf einen gewissen Rang noch einmal gesteigert. Numerische Testbeispiele bestätigen die Anwendbarkeit von DLRA Methoden auf inverse Probleme und die Effizienz des betrachteten Verfahrens.

Acknowledgements

This work would not have been possible without the support of many people and institutions. First and foremost, my great gratitude goes to my supervisor Prof. Dr. Christian Klingenberg for his constant and pleasant guidance through the years, for providing interesting research topics to work on, and for all the opportunities he offered for my personal and academic development.
I am also deeply thankful for the research collaborations with Prof. Dr. Lukas Einkemmer and Prof. Dr. Jonas Kusch. It was a great pleasure working together, thereby learning from their immense expertise and deepening my understanding of various mathematical concepts. This cooperation significantly enriched my research.
My research also profited from the financial support that I obtained from the PhD scholarship of the Stiftung der Deutschen Wirtschaft (Foundation of German Business) as well as the WMCCI (Würzburg Mathematics Center for Collaboration and Interaction), which allowed for fully concentrating on my work and for several scientific travels.
In general, I greatly enjoyed going to international conferences, meeting other researchers and broadening my mathematical mind. In particular, I would like to thank Prof. Dr. Kui Ren and Prof. Dr. Marlies Pirner for fruitful discussions and for excellent ideas.
I consider myself fortunate being part of such a lively and warmhearted work group in Würzburg. I truly enjoyed attending the seminar talks, welcoming our international guests and having all the cake meetings together.
I also received a lot of support during the process of preparing this thesis. Especially, I would like to thank Dr. Chiara Piazzola, Dr. Claudius Birke, Dr. Kathrin Hellmuth and Dr. Sandra Warnecke for providing helpful material. Additionally, I would like to express my gratitude to everyone who proofread this thesis.
I am deeply thankful to my parents and to my sister Andrea. It is an invaluable gift to have grown up in a loving family which still unconditionally supports me in everything I do. My final heartfelt thanks go to Claudius for his love, emotional support, and for being part of my life.

Lena Baumann

Contents

List of algorithms

1 Introduction

Many natural phenomena can be mathematically modeled by partial differential equations (PDEs). Classic examples are for instance bacterial movements [KS70], medical treatments such as radiation therapy [HMA81], radiation transport [Cha60], astrophysical phenomena [Alf42], heat transfer [CN47], wave equations [d'A47] or gas dynamics [Bol72]. The description of the underlying problem can hereby be given at different physical scales [Gra49, Deg04, Son19]. The most detailed *microscopic* description traces each particle of the considered medium individually, usually leading to a huge system of equations which is expensive to solve. The *mesoscopic* or *kinetic* description makes use of a distribution function that is based on the statistical repartition of the particles in phase space and can be considered as a probability density. The *macroscopic* regime contains less details. It depends only on macroscopic measurable quantities such as density, mean velocity, temperature, pressure, energy etc. This thesis focuses on the intermediate perspective, the class of kinetic equations.

Numerical solution of kinetic equations. For many systems described by kinetic PDEs, the computation of an analytical solution, if existing, is highly involved. In such cases, numerical approximations come into play, requiring a discretization of the system in the time variable $t \in \mathbb{R}_0^+$, the space variable $\mathbf{x} \in \Omega_{\mathbf{x}} \subseteq \mathbb{R}^{d_x}$, and the velocity variable $\mathbf{v} \in \Omega_{\mathbf{v}} \subseteq \mathbb{R}^{d_v}$. Depending on the number of space dimensions d_x and d_v, respectively, and the general complexity of the problem, the numerical solution of a kinetic equation can be computationally expensive. To speed up simulations and save computational effort and memory requirements, model reduction techniques such as *dynamical low-rank approximation (DLRA)* [KL07] can be used.

Basic principles of DLRA. For the application of the DLRA method to a kinetic equation, the distribution function f is approximated by a low-rank representation of the form

$$f(t, \mathbf{x}, \mathbf{v}) \approx \sum_{i,j=1}^{r} X_i(t, \mathbf{x}) S_{ij}(t) V_j(t, \mathbf{v}), \tag{1.1}$$

where $\{X_i : i = 1, ..., r\}$ denotes the set of orthonormal basis functions in space and $\{V_j : j = 1, ..., r\}$ the set of orthonormal basis functions in velocity. The matrix $\mathbf{S} = (S_{ij}) \in \mathbb{R}^{r \times r}$ is called the *coefficient* or *coupling matrix* and r the *rank* of the approximation. This splitting approach can be understood as a continuous analogue to the singular value decomposition (SVD) of a matrix. However, the matrix $\mathbf{S}$ is not required to be diagonal. The main idea of DLRA is to project the solution to a manifold of low-rank functions of the form (1.1) and to constrain the solution dynamics there. Special time integrators that are able to update the low-rank factors while not suffering from the curvature of the low-rank manifold exist [LO14, CL22, CKL22, CKL24]. This approach reduces the solution of a high-dimensional problem to solving lower-dimensional subproblems.

Research contributions. The construction of fully discretized numerical schemes for solving kinetic PDEs is challenging and requires careful consideration. Among the essential properties of a numerical scheme is its stability, meaning that approximation errors do not increase unrestrictedly over time so that a reasonable solution of the underlying physical problem is ensured. In this thesis, the concept of energy stability is used. This approach gives bounds to the energy of a system, hereby making the numerical behavior of a scheme predictable. Another difficulty arising from the application of reduction techniques is the preservation of physical properties inherent in the underlying equations such as for instance conservation laws. If physical conservation properties cannot be guaranteed, the reconstruction of the solution provided by the scheme may be inconsistent with the governing physical principles and, consequently, be considerably less applicable to realistic settings. In the first part of this thesis, we present DLRA schemes that explicitly accomplish numerical stability and conservation properties.

The first research contribution presented in this thesis, which is published in [BEKK24a], concerns the *thermal radiative transfer equations (RTEs)*. These equations form a system of two coupled PDEs that models radiation particles moving through and interacting with a background material. The Su-Olson closure is applied to obtain a linearized internal energy model, called the *Su-Olson problem*. We derive an energy stable DLRA scheme for the Su-Olson problem and provide a mathematically rigorous proof of energy stability under a certain hyperbolic Courant-Friedrichs-Lewy (CFL) condition. The conducted analysis allows for an optimal choice of the time step size, enhancing the computational performance of the algorithm. For the derivation of the DLRA scheme the basis augmentation step proposed by the rank-adaptive augmented basis update & Galerkin (BUG) integrator [CKL22] is implemented and adjusted in a way that together with a conservative truncation strategy as described in [EOS23] mass conservation can be ensured. Numerical experiments confirm the derived theoretical results.

It has been shown, for instance in [EHY21] for the non-linear isothermal Boltzmann-Bhatnagar-Gross-Krook (BGK) equation, that for the construction of efficient DLRA schemes a multiplicative splitting of the distribution function can offer advantages in reducing the computational effort. To investigate these schemes from an analytical perspective, we reconsider the Su-Olson problem together with a multiplicative splitting of the distribution function. The multiplicative structure poses additional challenges for the

construction of an energy stable DLRA scheme for the Su-Olson problem. For instance, careful consideration must be given to the discretization of the spatial derivatives and additional basis augmentations are required to ensure the exactness of the projection operators in the mathematically rigorous proof of energy stability. Mass conservation can be guaranteed similarly to the Su-Olson problem without multiplicative splitting when using a suitable low-rank integrator and a conservative truncation strategy. Numerical test examples confirm the properties of the derived DLRA scheme. The corresponding results can be found in [BEKK25b].
To extend the gained insights to more complicated problems such as the non-linear isothermal Boltzmann-BGK equation considered in [EHY21], further investigations on the multiplicative structure of the distribution function are conducted. In [BEKK24b], a multiplicative DLRA scheme for the linear isothermal Boltzmann-BGK equation is proposed. Within an appropriate stability framework, we perform a mathematically rigorous stability analysis and derive a concrete hyperbolic CFL condition. To ensure the analytical correctness, additional basis augmentations in the rank-adaptive augmented BUG integrator as well as a specifically designed truncation strategy are required. Various numerical experiments underline the theoretical results and the efficiency of the proposed DLRA scheme.
Computational advantages of the DLRA method can be especially observed in higher-dimensional settings. A classic problem requiring the solution of a considerable number of potentially high-dimensional kinetic equations, is the parameter identification in inverse problems. The second part of this thesis is devoted to the application of DLRA for the reconstruction of searched-for parameters in inverse problems. Following the presentation in [BEKK25a], we consider the RTE with a spatially dependent scattering coefficient. For its reconstruction, a PDE constrained optimization procedure with gradient-based update formula is applied, requiring the solution of the forward as well as of the adjoint equations in each step of the iterative scheme. For the numerical reconstruction, we propose a DLRA solver for the forward and the adjoint equations. An adaptive choice of the step size in the gradient-based iterative scheme and of the DLRA rank truncation tolerance lead to further improvements in efficiency. Numerical test examples show promising results for the combination of DLRA methods and parameter identification inverse problems.
Altogether, this thesis covers two main topics to which it contributes new results:

(i) Stability analysis (and conservation properties) for (multiplicative) DLRA schemes.

(ii) Application of DLRA to parameter identification inverse problems.

Structure of the thesis. After the introduction in Chapter 1, we review some fundamentals on kinetic theory in Chapter 2. These include a detailed overview of established possibilities for describing processes on different physical scales with a focus on the kinetic perspective as well as the important Boltzmann and simplified Boltzmann-BGK equation. Chapter 3 provides an introduction to the topic of numerical discretization in the space, time and velocity variable and recalls important concepts for numerical stability. Chapter 4 is devoted to the method of DLRA, which is at the center of this thesis. It formalizes

the basic idea of DLRA, explains different exact and robust time integrators and reviews results concerning stability and conservation properties of DLRA schemes. Part I of this thesis provides rigorous stability results for DLRA schemes applied to different problems. In Chapter 5 an energy stable and mass conservative DLRA scheme for the Su-Olson problem is proposed. Chapter 6 reconsiders the Su-Olson problem with a multiplicative splitting of the distribution function, posing further challenges in the construction of an appropriate energy stable and mass conservative DLRA scheme. Chapter 7 is devoted to the derivation of a provably stable DLRA scheme for the linear Boltzmann-BGK equation. Part II concerns the application of the DLRA method to inverse problems. Chapter 8 provides basic information for the numerical solution of parameter identification inverse problems, including their definition, the formulation of a corresponding optimization problem and techniques for its efficient solution. In Chapter 9 the application of an adaptive DLRA solver for the reconstruction of the scattering coefficient in the RTE is presented. Chapter 10 draws a short conclusion and provides an outlook for future research.

2

Fundamentals on kinetic theory

When modeling natural phenomena by PDEs, different physical scales containing more or less details may be considered. In Section 2.1 we provide a short overview of common approaches. Section 2.2 focuses on the *kinetic description* and provides basic principles, including a formal definition of the distribution function and the general form of a kinetic equation. Section 2.3 is devoted to the *Boltzmann equation*, which is a crucial equation in kinetic theory that continues to be actively studied [UA82, Per90, BGL00, FHJ12]. Section 2.4 introduces a simplification of the Boltzmann collision operator, namely the *BGK collision operator*. The following sections rely mainly on introductory work on kinetic theory and on the Boltzmann equation, in particular [Cer88, CIP94, CC90, Gra49, DP14].

2.1 PDE models in different physical regimes

Depending on the level of accuracy required in describing physical processes, different models are available. Since the models are intended to represent real-world applications, we use three-dimensional (3D) *Cartesian coordinates*, i.e. we consider $\mathbf{x} \in \Omega_{\mathbf{x}} \subseteq \mathbb{R}^3$ and $\mathbf{v} \in \Omega_{\mathbf{v}} \subseteq \mathbb{R}^3$.

Microscopic description. In the *microscopic* regime each particle of the corresponding medium is considered individually. The fundamental principles of particle dynamics in classical mechanics can be derived from Newton's laws of motion given in [New87]. Let $\mathbf{x}_i$ for $i = 1, ..., N$ be the position of the i-th particle of a medium consisting of N such particles and $\mathbf{v}_i$ its velocity. The time evolution of this particle is determined by Newton's equations

$$\dot{\mathbf{x}}_i(t) = \mathbf{v}_i \qquad \text{and} \qquad m_i \dot{\mathbf{v}}_i(t) = \mathbf{F}_i\left(\mathbf{x}_1, ..., \mathbf{x}_N\right), \tag{2.1}$$

where m_i denotes the particle mass and $\mathbf{F}_i$ the force acting on the i-th particle. In general, the force term $\mathbf{F}_i$ includes both the force that is exerted on the i-th particle by other particles as well as external forces, such as gravity. Determining the time evolution of an N-particle system using equations (2.1) requires the solution of $6N$ differential equations. For a huge number of particles N (as for instance a number of $\mathcal{O}\left(10^{23}\right)$ particles such as described by the Avogadro constant to be contained in one mole of a gas) this approach is usually infeasible.

Kinetic description. A less detailed description is given in the *mesoscopic* or *kinetic* regime. The idea of using a particle density distribution function instead of tracing each single particle of a rarefied monatomic gas goes back to Boltzmann. He presented the famous Boltzmann equation in [Bol72]. More information on the Boltzmann equation and on the basic assumptions for its derivation from the microscopic description can be found in Section 2.3. His work was inspired by previous considerations made by Maxwell [Max67] who gave a heuristic derivation of the particle density distribution function for a gas in thermodynamic equilibrium, the so-called *Maxwell-Boltzmann* or *Maxwellian distribution*. Important historical contributions for the solution of the Boltzmann equation were also made by Hilbert [Hil12], Chapman [Cha16] and Enskog [Ens17]. The idea of using a distribution function spread from the field of rarefied gas dynamics to other areas of research such as radiative transfer, neutron transport or quantum effects in gases [CC90].

Macroscopic description. Under certain limiting assumptions (see for instance [BGL91, Deg04, EP04]), it is possible to derive *macroscopic* or *fluid* equations from the kinetic regime. Historically, those equations go further back than kinetic ones as they only rely on observable macroscopic quantities such as density, mean velocity, temperature, pressure, energy etc., which are measurable quantities in experiments. Important macroscopic systems of PDEs are for instance the Euler equations [Eul57] and their more general extension to the Navier-Stokes equations, which include effects of viscosity [Nav22, Nav27, Sto45]. The Euler equations constitute a system of hyperbolic conservation laws, for which we provide a general definition.

Definition 2.1 (Conservation law, [LeV92])**.** Let $\mathbf{u}(t,\mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}^m$ denote an m-dimensional vector of conserved quantities. The *differential form of a conservation law* is given by

$$\partial_t \mathbf{u}(t,\mathbf{x}) + \nabla_{\mathbf{x}} \cdot \mathcal{F}(\mathbf{u}(t,\mathbf{x})) = 0, \tag{2.2}$$

where $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3)^\top \in \mathbb{R}^{3m}$ denotes the flux vector containing the flux functions $\mathcal{F}_i : \mathbb{R}^m \to \mathbb{R}^m$ for $i = 1, 2, 3$. For $m = 1$, equation (2.2) is called a *scalar conservation law*, for $m \geq 2$ a *system of conservation laws*.

Note that we do not specify the regularity of a solution to (2.2) here. For more information on different solution concepts of conservation laws the reader is referred to literature such as [Eva10, LeV02, Daf16, Mar21]. Equation (2.2) can be rewritten in the quasi-linear

form

$$\partial_t \mathbf{u}(t,\mathbf{x}) + \sum_{i=1}^{3} \mathcal{A}_i \partial_{x_i} \mathbf{u}(t,\mathbf{x}) = 0,$$

where $\mathcal{A}_i := \nabla_{\mathbf{u}} \mathcal{F}_i(\mathbf{u}) \in \mathbb{R}^{m\times m}$ denotes the flux Jacobian matrices. We focus on a special class of equations, the hyperbolic equations.

Definition 2.2 (Hyperbolicity, [LeV92])**.** The conservation law (2.2) is called *hyperbolic* if the matrix $\mathcal{A} := \sum_{i=1}^{3} \alpha_i \mathcal{A}_i$ with $\alpha_i \in \mathbb{R}$ has only real eigenvalues $\lambda_1, ..., \lambda_m$ and is diagonalizable, i.e. a full set of m linearly independent eigenvectors exists. If $\mathcal{A}$ has m distinct eigenvalues, (2.2) is called *strictly hyperbolic*.

Using equation (2.2), the time evolution of the function $\mathbf{u}(t,\mathbf{x})$ can be determined. Adding information in the form of a suitable initial condition

$$\mathbf{u}(0,\mathbf{x}) = \mathbf{u}^0(\mathbf{x}) \tag{2.3}$$

complements the solution. Equation (2.2) together with (2.3) is called a *Cauchy* or *initial value problem (IVP)*.

Choice of a suitable description. Figure 2.1 provides an overview of the descriptions introduced for PDE models on different physical scales. This illustration is inspired by the one given in [War22]. The decision on a model of appropriate accuracy can be challenging. A helpful indicator can be the Knudsen number Kn of the particular system. The Knudsen number Kn represents the ratio of the mean free path, i.e. the distance that a particle travels on average until it collides with another one, and a characteristic length scale of the corresponding system. In [Str05] a rough classification for an appropriate description depending on the Knudsen number is provided. For our purpose, a microscopic description is clearly infeasible as the considered systems consist of large numbers of particles. A macroscopic description potentially loses too much information as it only accounts for velocity-averaged quantities. For this reason, we focus on kinetic hyperbolic models in this thesis.

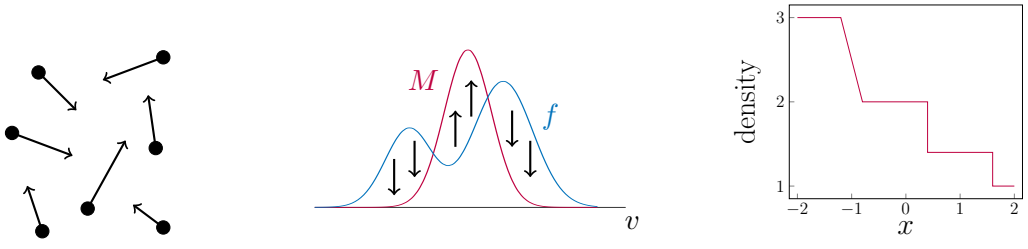


Figure 2.1: Possibilities for the description of natural phenomena on differently detailed physical scales. **Left:** Microscopic description: The trajectory of each single particle is considered individually. **Middle:** Kinetic description: Statistical repartition of the particles using a distribution function f, which tends to a Maxwellian distribution M in equilibrium. This illustration is described more precisely in Figure 2.2. **Right:** Macroscopic description: Only measurable quantities such as the density are available.

2.2 Basic principles for the kinetic description

When determining the time evolution of a system consisting of a large number of particles, a statistical description based on a distribution function can be useful. The concept of using distribution functions in kinetic theory arises from probability theory and we refer to [Cer88] for further reading. We provide the following definition.

Definition 2.3 (Distribution function and phase space, [CC90, Pir18])**.** An integrable function $f : \mathbb{R}_0^+ \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}_0^+, (t, \mathbf{x}, \mathbf{v}) \mapsto f(t, \mathbf{x}, \mathbf{v})$ is called a *distribution function* if and only if $f \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{v}$ is the probable number of particles which are situated in the volume element $(\mathbf{x}, \mathbf{x} + \mathrm{d}\mathbf{x})$ and have velocities in $(\mathbf{v}, \mathbf{v} + \mathrm{d}\mathbf{v})$ at time t. The set of all possible physical states $(\mathbf{x}, \mathbf{v})$ of a particle at time t is called the *phase space*.

Under the common assumption of a normalization to one, the distribution function $f(t, \mathbf{x}, \mathbf{v})$ describes the probability of finding a particle at the position $(\mathbf{x}, \mathbf{v})$ in phase space at time t. We use this concept to introduce the general form of a kinetic equation.

Definition 2.4 (General kinetic equation, [DP14])**.** Let $f(t, \mathbf{x}, \mathbf{v})$ be a distribution function. The *general form of a kinetic equation* is given by

$$\partial_t f(t, \mathbf{x}, \mathbf{v}) + \mathbf{v} \cdot \nabla_{\mathbf{x}} f(t, \mathbf{x}, \mathbf{v}) + \frac{\mathbf{F}(t, \mathbf{x})}{m} \cdot \nabla_{\mathbf{v}} f(t, \mathbf{x}, \mathbf{v}) = \mathcal{Q}[f](t, \mathbf{x}, \mathbf{v}), \tag{2.4}$$

where m denotes the particle mass, $\mathbf{F}$ the effects of external forces, and $\mathcal{Q}[f]$ the collision operator describing the effects of internal forces due to particle interactions.

Note that in this thesis only problems without external forces, i.e. $\mathbf{F} = 0$, are considered. Given the distribution function, important macroscopic quantities can be derived by taking moments in the velocity variable $\mathbf{v}$.

Definition 2.5 (Macroscopic quantities, [Pir18])**.** Let $f(t, \mathbf{x}, \mathbf{v})$ be a distribution function and the subsequent integrands be in $L^1(\mathrm{d}\mathbf{v})$. Then the following *macroscopic quantities* are defined.

(i) Let m denote the particle mass. The functions

$$n(t, \mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}_0^+, \qquad (t, \mathbf{x}) \mapsto \int_{\mathbb{R}^3} f(t, \mathbf{x}, \mathbf{v}) \, \mathrm{d}\mathbf{v} \qquad \text{and}$$
$$\rho(t, \mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}_0^+, \qquad (t, \mathbf{x}) \mapsto m \int_{\mathbb{R}^3} f(t, \mathbf{x}, \mathbf{v}) \, \mathrm{d}\mathbf{v}$$

are called the *number density* and the *mass density*, respectively, and it holds $\rho(t, \mathbf{x}) = mn(t, \mathbf{x})$.

(ii) We define the function

$$n(t, \mathbf{x}) \, \bar{\mathbf{u}}(t, \mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}^3, \qquad (t, \mathbf{x}) \mapsto \int_{\mathbb{R}^3} f(t, \mathbf{x}, \mathbf{v}) \, \mathbf{v} \, \mathrm{d}\mathbf{v}$$

and, for $n(t, \mathbf{x}) > 0$, call $\bar{\mathbf{u}}(t, \mathbf{x})$ the *mean velocity*.

(iii) The function

$$E(t,\mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}_0^+, \qquad (t,\mathbf{x}) \mapsto \frac{m}{2}\int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\,|\mathbf{v}|^2\,\mathrm{d}\mathbf{v}$$

is called the *energy density*.

(iv) The energy density can be split into two parts, the *kinetic energy* $E_{\mathrm{kin}}(t,\mathbf{x}) = \frac{1}{2}\rho(t,\mathbf{x})\,|\bar{\mathbf{u}}(t,\mathbf{x})|^2$ and the *internal energy*

$$\begin{aligned} e(t,\mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}_0^+, \qquad (t,\mathbf{x}) &\mapsto E(t,\mathbf{x}) - E_{\mathrm{kin}}(t,\mathbf{x}) \\ &= \frac{m}{2}\int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\,|\mathbf{v}-\bar{\mathbf{u}}(t,\mathbf{x})|^2\,\mathrm{d}\mathbf{v}. \end{aligned}$$

(v) Using the ideal gas law, for $n(t,\mathbf{x}) > 0$, the *temperature* can be derived as

$$\begin{aligned} T(t,\mathbf{x}) : \mathbb{R}_0^+ \times \mathbb{R}^3 \to \mathbb{R}_0^+, \quad (t,\mathbf{x}) &\mapsto \frac{2e(t,\mathbf{x})}{3n(t,\mathbf{x})\,k_{\mathrm{B}}} \\ &= \frac{m}{3n(t,\mathbf{x})\,k_{\mathrm{B}}}\int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\,|\mathbf{v}-\bar{\mathbf{u}}(t,\mathbf{x})|^2\,\mathrm{d}\mathbf{v}, \end{aligned}$$

where k_{B} denotes the Boltzmann constant. The *pressure* is obtained as $p(t,\mathbf{x}) = n(t,\mathbf{x})\,k_{\mathrm{B}}T(t,\mathbf{x})$.

Having $\mathbf{F} = 0$ in (2.4), the collision operator $\mathcal{Q}[f]$ should, in general, be designed to preserve the conservation properties of the physical system, i.e. it should satisfy

$$\int_{\mathbb{R}^3} \mathcal{Q}[f](t,\mathbf{x},\mathbf{v})\,\varphi(\mathbf{v})\,\mathrm{d}\mathbf{v} = 0, \tag{2.5}$$

where $\varphi(\mathbf{v}) = 1, \mathbf{v}, |\mathbf{v}|^2$ characterizes the conservation of mass, momentum and energy, respectively. Under this assumption, integration of (2.4) against $\varphi(\mathbf{v})$ with respect to $\mathbf{v}$ yields the following system of *local conservation laws*

$$\partial_t \int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\,\varphi(\mathbf{v})\,\mathrm{d}\mathbf{v} + \int_{\mathbb{R}^3} \mathbf{v}\cdot\nabla_{\mathbf{x}} f(t,\mathbf{x},\mathbf{v})\,\varphi(\mathbf{v})\,\mathrm{d}\mathbf{v} = 0. \tag{2.6}$$

If it holds

$$\partial_t \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} f(t,\mathbf{x},\mathbf{v})\,\varphi(\mathbf{v})\,\mathrm{d}\mathbf{v}\,\mathrm{d}\mathbf{x} = 0, \tag{2.7}$$

the corresponding *global conservation laws* are fulfilled.

System (2.6) is not closed since the second term depends on higher order moments in $\mathbf{v}$. For the derivation of a closed set of equations, an additional assumption on the collision operator involving the Maxwellian equilibrium distribution is made.

Definition 2.6 (Maxwellian distribution, [CC90])**.** A distribution function of the form

$$M[f](t,\mathbf{x},\mathbf{v}) = \frac{n(t,\mathbf{x})}{\left(2\pi\frac{k_\mathrm{B}T(t,\mathbf{x})}{m}\right)^{3/2}} \exp\left(-\frac{|\mathbf{v}-\bar{\mathbf{u}}(t,\mathbf{x})|^2}{2\frac{k_\mathrm{B}T(t,\mathbf{x})}{m}}\right) \tag{2.8}$$

is called *Maxwellian distribution*.

The Maxwellian distribution describes a system that is in thermodynamic equilibrium. This relation shall be incorporated in the collision operator so that it fulfills

$$\mathcal{Q}[f] = 0 \quad \text{if and only if} \quad f = M[f]. \tag{2.9}$$

Both the Boltzmann collision operator $\mathcal{Q}_\mathrm{Bol}[f]$ and the simplified BGK collision operator $\mathcal{Q}_\mathrm{BGK}[f]$, which are introduced in the next sections, accomplish the two properties (2.5) and (2.9). Note that from now on the particle mass m as well as the Boltzmann constant k_B are set to one.

2.3 Boltzmann equation

The Boltzmann equation may be considered the most important equation in kinetic theory. It describes the time evolution of a perfect monatomic dilute gas. For its derivation from the microscopic description a considerable number of assumptions are made. A well-structured overview is provided in [Vil02]. First of all, only binary collisions of identical gas particles are assumed, meaning that interactions involving more than two particles are neglected. Second, the collisions are supposed to be purely local and instantaneous in time, i.e. they happen at a given time t and a given position $\mathbf{x}$ and have a very short duration compared to the typical time scale of the system. Third, the collisions are assumed to be elastic. This means that for two particles with velocities $\mathbf{v}'$ and $\mathbf{v}'_*$ before the collision and velocities $\mathbf{v}$ and $\mathbf{v}_*$ after the collision the following corresponding relations for the microscopic conservation of momentum and energy shall hold:

$$\begin{aligned} \mathbf{v}' + \mathbf{v}'_* &= \mathbf{v} + \mathbf{v}_* && \text{(conservation of momentum)}, \\ |\mathbf{v}'|^2 + |\mathbf{v}'_*|^2 &= |\mathbf{v}|^2 + |\mathbf{v}_*|^2 && \text{(conservation of energy)}. \end{aligned}$$

Fourth, the collisions shall be reversible in time at a microscopic level. This assumption implies that changing the velocities from $(\mathbf{v}', \mathbf{v}'_*)$ to $(\mathbf{v}, \mathbf{v}_*)$ is as probable as changing the velocities from $(\mathbf{v}, \mathbf{v}_*)$ to $(\mathbf{v}', \mathbf{v}'_*)$. The fifth assumption is referred to as Boltzmann's *molecular chaos assumption* or *Stosszahlansatz*. It states that the velocities of the particles that are about to collide are statistically uncorrelated. Indeed, the velocities of the particles that have just collided are statistically correlated. This asymmetry bridges the gap between the time reversible microscopic and the time irreversible kinetic and macroscopic description. We can now present the following formulation of the *Boltzmann equation*.

Definition 2.7 (Boltzmann equation, [Gol05, Cer88])**.** In terms of a distribution function f, the *Boltzmann equation* reads

$$\partial_t f(t,\mathbf{x},\mathbf{v}) + \mathbf{v}\cdot\nabla_{\mathbf{x}} f(t,\mathbf{x},\mathbf{v}) = \mathcal{Q}_{\mathrm{Bol}}[f](t,\mathbf{x},\mathbf{v}), \tag{2.10}$$

with $\mathcal{Q}_{\mathrm{Bol}}[f]$ being the quadratic *Boltzmann collision operator*

$$\mathcal{Q}_{\mathrm{Bol}}[f](t,\mathbf{x},\mathbf{v}) = \int_{\mathbb{R}^3}\int_{\mathcal{S}^2} \left(f\left(t,\mathbf{x},\mathbf{v}'\right) f\left(t,\mathbf{x},\mathbf{v}'_*\right) - f(t,\mathbf{x},\mathbf{v})\, f(t,\mathbf{x},\mathbf{v}_*)\right) \overline{B}(\mathbf{v}-\mathbf{v}_*,\eta)\,\mathrm{d}\eta\,\mathrm{d}\mathbf{v}_*,$$

where $\eta \in \mathcal{S}^2$ denotes an arbitrary unit vector contained in the 3D unit sphere $\mathcal{S}^2$ and $\overline{B}(\mathbf{v}-\mathbf{v}_*,\eta)$ is called the *collision kernel*.

The exact form of the collision kernel $\overline{B}$ depends on the considered setting. Common assumptions for the particle interactions are hard sphere collisions or interactions due to a central force in a smooth potential. In both cases, the collision kernel can be explicitly stated. More information on the choice of an appropriate collision kernel can be found in [Vil02, Gol05, Cer88]. In this thesis, we follow [Gol05] and assume $\overline{B}$ to be locally integrable on $\mathbb{R}^3\times\mathcal{S}^2$. In addition, we assume the distribution function f to be continuous with compact support in the velocity variable. Then the Boltzmann collision operator $\mathcal{Q}_{\mathrm{Bol}}[f]$ can be rewritten as

$$\mathcal{Q}_{\mathrm{Bol}}[f](t,\mathbf{x},\mathbf{v}) = \mathcal{Q}^+_{\mathrm{Bol}}[f](t,\mathbf{x},\mathbf{v}) - \mathcal{Q}^-_{\mathrm{Bol}}[f](t,\mathbf{x},\mathbf{v}),$$

where

$$\mathcal{Q}^+_{\mathrm{Bol}}[f](t,\mathbf{x},\mathbf{v}) = \int_{\mathbb{R}^3}\int_{\mathcal{S}^2} f\left(t,\mathbf{x},\mathbf{v}'\right) f\left(t,\mathbf{x},\mathbf{v}'_*\right)\overline{B}(\mathbf{v}-\mathbf{v}_*,\eta)\,\mathrm{d}\eta\,\mathrm{d}\mathbf{v}_* \qquad \text{and}$$
$$\mathcal{Q}^-_{\mathrm{Bol}}[f](t,\mathbf{x},\mathbf{v}) = \int_{\mathbb{R}^3}\int_{\mathcal{S}^2} f(t,\mathbf{x},\mathbf{v})\, f(t,\mathbf{x},\mathbf{v}_*)\,\overline{B}(\mathbf{v}-\mathbf{v}_*,\eta)\,\mathrm{d}\eta\,\mathrm{d}\mathbf{v}_*$$

are called the *gain term* and the *loss term*, respectively. The gain term accounts for particles having velocities $(\mathbf{v}',\mathbf{v}'_*)$ that change their velocities to $(\mathbf{v},\mathbf{v}_*)$ after the collision. In this sense, particles with velocity $\mathbf{v}$ are gained in the volume element $\mathrm{d}\mathbf{v}$ centered around $\mathbf{v}$. The loss term instead accounts for particles having velocities $(\mathbf{v},\mathbf{v}_*)$ that change their velocities to $(\mathbf{v}',\mathbf{v}'_*)$ after the collision. In this sense, particles with velocity $\mathbf{v}$ are lost in the volume element $\mathrm{d}\mathbf{v}$ centered around $\mathbf{v}$.

The Boltzmann collision operator fulfills important physical properties. For instance, it guarantees the conservation of mass, momentum and energy.

Theorem 2.8 (Conservation properties for the Boltzmann equation, [Gol05])**.** *Let* $f = f(\mathbf{v}) \in C_c\left(\mathbb{R}^3\right)$ *and* $\overline{B} \in L^1_{loc}\left(\mathbb{R}^3\times\mathcal{S}^2\right)$*. Then, for* $i = 1,2,3$*, the Boltzmann collision*

operator $\mathcal{Q}_{Bol}[f]$ *fulfills*

$$\int_{\mathbb{R}^3} \mathcal{Q}_{Bol}[f](t,\mathbf{x},\mathbf{v})\,\mathrm{d}\mathbf{v} = 0 \qquad \textit{(conservation of mass)},$$
$$\int_{\mathbb{R}^3} \mathcal{Q}_{Bol}[f](t,\mathbf{x},\mathbf{v})\,v_i\,\mathrm{d}\mathbf{v} = 0 \qquad \textit{(conservation of momentum)},$$
$$\int_{\mathbb{R}^3} \mathcal{Q}_{Bol}[f](t,\mathbf{x},\mathbf{v})\,|\mathbf{v}|^2\,\mathrm{d}\mathbf{v} = 0 \qquad \textit{(conservation of energy)}.$$

Proof. See for instance [Gol05, Cer88]. □

Remark 2.9. Note that the assumption $f = f(\mathbf{v}) \in C_c\left(\mathbb{R}^3\right)$ in Theorem 2.8 is quite strong and can be weakened under additional assumptions [Gol05].

According to the second law of thermodynamics, the entropy of a thermodynamical system is non-decreasing, explaining that such processes are irreversible in time. This behavior is mirrored in *Boltzmann's H-theorem*.

Theorem 2.10 (Boltzmann's H-theorem, [Gol05])**.** *Let* $f = f(\mathbf{v}) \in C\left(\mathbb{R}^3\right)$ *be positive and rapidly decaying at infinity and* $\overline{B} \in L^1_{loc}\left(\mathbb{R}^3 \times \mathcal{S}^2\right)$*. Further, assume that there exists* $m > 0$ *such that*

$$\int_{\mathcal{S}^2} \overline{B}(\mathbf{v},\eta)\,\mathrm{d}\eta + |\ln f(\mathbf{v})| = \mathcal{O}\left(|\mathbf{v}|^m\right) \qquad \textit{as} \quad |\mathbf{v}| \to +\infty.$$

Then the following inequality holds

$$\int_{\mathbb{R}^3} \mathcal{Q}_{Bol}[f](t,\mathbf{x},\mathbf{v}) \ln f(t,\mathbf{x},\mathbf{v})\,\mathrm{d}\mathbf{v} \leq 0,$$

with equality if and only if f *is a Maxwellian distribution.*

Proof. See for instance [Gol05, Cer88]. □

The proof of Boltzmann's H-theorem also reveals the special Maxwellian structure of the equilibrium.

Corollary 2.11 (Structure of the equilibrium, [Gol05])**.** *Let the assumptions of the previous theorem hold. Then the Boltzmann collision operator satisfies*

$$\mathcal{Q}_{Bol}[f](t,\mathbf{x},\mathbf{v}) = 0 \quad \textit{for all} \quad \mathbf{v} \in \mathbb{R}^3 \quad \textit{if and only if} \quad f(t,\mathbf{x},\mathbf{v}) = M[f](t,\mathbf{x},\mathbf{v}).$$

2.4 Boltzmann-BGK equation

For practical implementations, the solution of the full Boltzmann equation (2.10) is, in general, computationally expensive. Simplifications of the Boltzmann collision operator that maintain its key properties while being numerically less demanding are sought. A

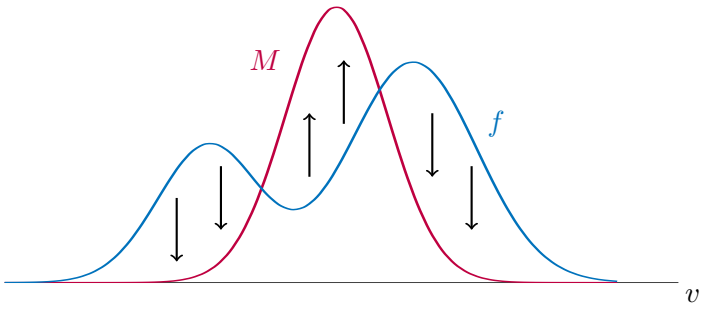


Figure 2.2: Relaxation of the distribution function f towards the Maxwellian distribution M in the space homogeneous case. The arrows depict the time derivatives of f with respect to the evolution equation (2.11).

widely used collision model is the *Bhatnagar-Gross-Krook (BGK) collision operator*. It was proposed by Bhatnagar, Gross and Krook [BGK54] and is sometimes referred to as the *Krook operator*. We replace the Boltzmann collision operator $\mathcal{Q}_{\mathrm{Bol}}[f]$ in (2.10) by the BGK collision operator $\mathcal{Q}_{\mathrm{BGK}}[f]$ and obtain the *Boltzmann-BGK equation*.

Definition 2.12 (Boltzmann-BGK equation, [BGK54])**.** The *Boltzmann-BGK equation* reads

$$\partial_t f(t,\mathbf{x},\mathbf{v}) + \mathbf{v}\cdot\nabla_{\mathbf{x}} f(t,\mathbf{x},\mathbf{v}) = \mathcal{Q}_{\mathrm{BGK}}[f](t,\mathbf{x},\mathbf{v}), \tag{2.11}$$

where the *BGK collision operator* is given by

$$\mathcal{Q}_{\mathrm{BGK}}[f](t,\mathbf{x},\mathbf{v}) = \sigma(t,\mathbf{x})\left(M[f](t,\mathbf{x},\mathbf{v}) - f(t,\mathbf{x},\mathbf{v})\right), \tag{2.12}$$

and $\sigma = \sigma(t,\mathbf{x}) \geq 0$ denotes a prescribed collision frequency.

The Boltzmann-BGK equation (2.11) describes the relaxation of the distribution function f towards the corresponding Maxwellian distribution M when the system is close to thermodynamic equilibrium. For this reason, the BGK collision operator (2.12) is also called a *relaxation operator*. Figure 2.2 illustrates this behavior. The BGK collision operator (2.12) maintains important properties of the full Boltzmann collision operator.

Theorem 2.13 (Properties of the BGK collision operator, [Pir18])**.** *Let $f = f(\mathbf{v}) \in C(\mathbb{R}^3)$ be positive and rapidly decaying at infinity. Then the BGK collision operator $\mathcal{Q}_{BGK}[f]$ has the same main properties as the Boltzmann collision operator $\mathcal{Q}_{Bol}[f]$, namely the conservation of mass, momentum and energy, the H-theorem, and the structure of the equilibrium.*

Proof. See for instance [Pir18, Str05]. □

The existence and uniqueness of solutions to the Boltzmann-BGK equation (2.11) has been proven in [Per89, PP93]. The Boltzmann-BGK equation is a simplified model of the Boltzmann equation that is widely used in research in its original as well as in modified forms, allowing, for example, for the reproduction of a correct Prandtl number [Hol66, ATPP00] or for velocity-dependent collision frequencies [Str97, HHK+21].

3

Discretization and numerical stability

In order to obtain a numerical solution of PDEs, their continuous formulation must first be transformed into a fully discretized representation. For deriving the theoretical concepts of discretization and numerical stability, we follow the explanations given in [LeV92] and restrict our considerations to one-dimensional (1D) linear advection equations of the form

$$\partial_t u(t,x) + a\partial_x u(t,x) = 0, \tag{3.1}$$

where $u(t,x) : [0,T] \times \Omega_x \to \mathbb{R}$ is assumed to be sufficiently regular and $a \in \mathbb{R}$ denotes a constant value. Combined with an appropriate initial condition $u(0,x) = u^0(x)$, the solution of (3.1) admits the explicit form $u(t,x) = u^0(x - at)$. Consequently, the solution $u(t,x)$ at any fixed point (t,x) solely depends on the initial condition evaluated at $x - at$ and is constant along each *characteristic curve* described by $x^0 = x - at$. The set

$$\mathcal{D}(t,x) = \{x - at\} \tag{3.2}$$

is called the *true domain of dependence* of the PDE. In Figure 3.1 the characteristic curves as well as the solution to the linear advection equation (3.1) are sketched for $a > 0$.
Section 3.1 presents techniques for the spatial and temporal discretization commonly used for problems of the form (3.1). Section 3.2 is devoted to stability considerations. Section 3.3 relates kinetic equations to the previously discussed methods by introducing an appropriate discretization in the velocity variable. A selection of introductory literature used for Sections 3.1 and 3.2 can be found in [Tho95, LeV92, LeV07, Str04, RM67].

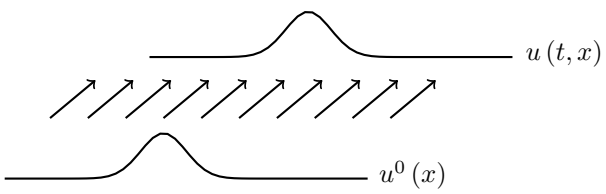


Figure 3.1: Illustration of the initial condition $u^0(x)$, the solution to the linear advection equation $u(t,x)$ and the characteristic curves depicted by arrows for $a > 0$.

3.1 Discretization in space and time

The linear advection equation (3.1) depends continuously on space and time. For the derivation of a numerical scheme, a discretization in both the spatial variable x and the temporal variable t must be performed. We begin with a discretization in the spatial variable in Section 3.1.1, leading to a *semi-discrete* representation, before the system is rendered *fully discrete* through a temporal discretization presented in Section 3.1.2.

3.1.1 Spatial discretization

Regarding the discretization of the spatial domain Ω_x, we construct a uniform *spatial grid* consisting of a finite number of grid cells $N_x \in \mathbb{N}$. The grid points $x_0, x_1, ..., x_{N_x} \in \Omega_x$ are assumed to be uniformly distributed with equidistant spacing $\Delta x = \frac{1}{N_x}$. An approximate solution $\mathbf{u}(t) \in \mathbb{R}^{N_x}$ to (3.1) on the spatial grid is obtained by evaluating $u(t,x)$ at the end point of each grid cell, i.e. by computing

$$u_j(t) \approx u(t, x_j) \qquad \text{for} \quad j = 1, ..., N_x.$$

The linear advection equation (3.1) involves spatial derivatives. Different approaches for their numerical approximation are available. In this thesis, we focus on *centered finite difference (FD)* schemes.

Numerical differentiation. For the derivation of a centered FD scheme for equation (3.1) we consider the following Taylor expansions of its continuous solution:

$$u(t, x + \Delta x) = u(t,x) + \Delta x \partial_x u(t,x) + \frac{(\Delta x)^2}{2} \partial_{xx} u(t,x) + \mathcal{O}\left((\Delta x)^3\right), \tag{3.3}$$

$$u(t, x - \Delta x) = u(t,x) - \Delta x \partial_x u(t,x) + \frac{(\Delta x)^2}{2} \partial_{xx} u(t,x) - \mathcal{O}\left((\Delta x)^3\right). \tag{3.4}$$

Subtracting equation (3.4) from equation (3.3) allows for the formulation

$$\partial_x u(t,x) = \frac{u(t, x+\Delta x) - u(t, x-\Delta x)}{2\Delta x} + \mathcal{O}\left((\Delta x)^2\right).$$

This approach is used for the approximation of first-order derivatives as

$$\partial_x u(t,x)\big|_{x=x_j} \approx \frac{u_{j+1}(t) - u_{j-1}(t)}{2\Delta x}. \tag{3.5}$$

At the grid points x_{N_x} and x_1, formula (3.5) involves the evaluation of $u_{N_x+1}(t)$ and $u_0(t)$, respectively. *Periodic boundary conditions* can be introduced, determining

$$u_0(t) := u_{N_x}(t) \qquad \text{and} \qquad u_{N_x+1}(t) := u_1(t). \tag{3.6}$$

Then the following semi-discrete time-continuous form of the advection equation (3.1) can be derived as

$$\dot{u}_j(t) = -\frac{a}{2\Delta x}\left(u_{j+1}(t) - u_{j-1}(t)\right). \tag{3.7}$$

Expression (3.7) can be reformulated as

$$\dot{\mathbf{u}}(t) = -a\mathbf{D}^x\mathbf{u}(t),$$

where the matrix $\mathbf{D}^x \in \mathbb{R}^{N_x \times N_x}$ with entries

$$D^x_{j,j\pm1} = \frac{\pm 1}{2\Delta x}, \qquad D^x_{1,N_x} = \frac{-1}{2\Delta x}, \qquad D^x_{N_x,1} = \frac{1}{2\Delta x} \tag{3.8}$$

incorporates a centered FD approximation for first-order spatial derivatives ∂_x as well as periodic boundary conditions.

Extension to second-order derivatives. Although equation (3.1) involves no spatial derivatives beyond the first order, adding a second-order diffusion term to the spatial discretization can be beneficial for its numerical stability [LeV02]. For the derivation of a centered FD approximation for second-order derivatives, the Taylor expansions given in (3.3) and (3.4) are reconsidered. Adding these equations leads to the expression

$$\partial_{xx}u(t,x) = \frac{u(t,x+\Delta x) - 2u(t,x) + u(t,x-\Delta x)}{(\Delta x)^2} + \mathcal{O}\left((\Delta x)^2\right).$$

Similarly to the first-order case, approximations of second-order derivatives can be obtained as

$$\partial_{xx}u(t,x)\big|_{x=x_j} \approx \frac{u_{j+1}(t) - 2u_j(t) + u_{j-1}(t)}{(\Delta x)^2}. \tag{3.9}$$

Instead of (3.7), an advection-diffusion equation of the form

$$\dot{u}_j(t) = -\frac{a}{2\Delta x}\left(u_{j+1}(t) - u_{j-1}(t)\right) + \frac{|a|}{(\Delta x)^2}\left(u_{j+1}(t) - 2u_j(t) + u_{j-1}(t)\right) \tag{3.10}$$

is numerically solved. Expression (3.10) can be reformulated as

$$\dot{\mathbf{u}}(t) = -a\mathbf{D}^x\mathbf{u}(t) + |a|\,\mathbf{D}^{xx}\mathbf{u}(t),$$

where the matrix $\mathbf{D}^{xx} \in \mathbb{R}^{N_x \times N_x}$ with entries

$$D^{xx}_{j,j} = -\frac{2}{(\Delta x)^2}, \qquad D^{xx}_{j,j\pm1} = \frac{1}{(\Delta x)^2}, \qquad D^{xx}_{1,N_x} = D^{xx}_{N_x,1} = \frac{1}{(\Delta x)^2} \tag{3.11}$$

incorporates a centered FD approximation for second-order spatial derivatives ∂_{xx} as well as periodic boundary conditions.

Properties of the differentiation matrices. The spatial stencil matrices $\mathbf{D}^x$ and $\mathbf{D}^{xx}$ exhibit useful properties. For instance, the discrete counterpart of the continuous integration by parts method can be shown.

Lemma 3.1 (Summation by parts)**.** *Let $\mathbf{y}, \mathbf{z} \in \mathbb{R}^{N_x}$ be vectors with indices $i, j = 1, ..., N_x$. In addition, we define $y_0 := y_{N_x}$ and $y_{N_x+1} := y_1$ as well as $z_0 := z_{N_x}$ and $z_{N_x+1} := z_1$ to account for periodic boundary conditions. Then the stencil matrices $\mathbf{D}^x$ and $\mathbf{D}^{xx}$ fulfill the following properties:*

$$\sum_{i,j=1}^{N_x} y_j D^x_{ji} z_i = -\sum_{i,j=1}^{N_x} z_j D^x_{ji} y_i, \qquad \sum_{i,j=1}^{N_x} z_j D^x_{ji} z_i = 0, \qquad \sum_{i,j=1}^{N_x} y_j D^{xx}_{ji} z_i = \sum_{i,j=1}^{N_x} z_j D^{xx}_{ji} y_i.$$

Moreover, consider the stencil matrix $\mathbf{D}^+ \in \mathbb{R}^{N_x \times N_x}$, defined by its entries

$$D^+_{j,j} = \frac{-1}{\Delta x}, \qquad D^+_{j,j+1} = \frac{1}{\Delta x}, \qquad D^+_{N_x,1} = \frac{1}{\Delta x}.$$

Then,

$$\sum_{i,j=1}^{N_x} z_j D^{xx}_{ji} z_i = -\sum_{j=1}^{N_x} \left(\sum_{i=1}^{N_x} D^+_{ji} z_i \right)^2.$$

Proof. The assertions follow directly by inserting the definitions of the spatial stencil matrices and by properly rearranging the expressions. In detail, we obtain

$$\begin{aligned}
\sum_{i,j=1}^{N_x} y_j D^x_{ji} z_i &= \frac{1}{2\Delta x} \sum_{j=1}^{N_x} y_j \left(z_{j+1} - z_{j-1} \right) = -\frac{1}{2\Delta x} \sum_{j=1}^{N_x} z_j \left(y_{j+1} - y_{j-1} \right) = -\sum_{i,j=1}^{N_x} z_j D^x_{ji} y_i, \\
\sum_{i,j=1}^{N_x} z_j D^x_{ji} z_i &= -\sum_{i,j=1}^{N_x} z_j D^x_{ji} z_i = 0, \\
\sum_{i,j=1}^{N_x} y_j D^{xx}_{ji} z_i &= -\frac{2}{(\Delta x)^2} \sum_{j=1}^{N_x} y_j z_j + \frac{1}{(\Delta x)^2} \sum_{j=1}^{N_x} y_j \left(z_{j+1} + z_{j-1} \right) \\
&= -\frac{2}{(\Delta x)^2} \sum_{j=1}^{N_x} z_j y_j + \frac{1}{(\Delta x)^2} \sum_{j=1}^{N_x} z_j \left(y_{j+1} + y_{j-1} \right) = \sum_{i,j=1}^{N_x} z_j D^{xx}_{ji} y_i, \\
\sum_{i,j=1}^{N_x} z_j D^{xx}_{ji} z_i &= -\frac{2}{(\Delta x)^2} \sum_{j=1}^{N_x} z_j^2 + \frac{1}{(\Delta x)^2} \sum_{j=1}^{N_x} z_j \left(z_{j+1} + z_{j-1} \right) \\
&= -\frac{1}{(\Delta x)^2} \sum_{j=1}^{N_x} \left(z_j^2 - 2 z_j z_{j+1} + z_{j+1}^2 \right) = -\frac{1}{(\Delta x)^2} \sum_{j=1}^{N_x} \left(z_j - z_{j+1} \right)^2 \\
&= -\sum_{j=1}^{N_x} \left(\sum_{i=1}^{N_x} D^+_{ji} z_i \right)^2.
\end{aligned}$$

□

Extension to two dimensions. The theoretical stability analysis in this thesis is conducted in one spatial dimension, represented by the spatial variable x. However, numerical experiments will be performed in two-dimensional (2D) spatial settings involving dependencies on the variable $\mathbf{x} = (x, y)^\top \in \mathbb{R}^2$. Therefore, we extend the centered FD framework to equations of the form

$$\partial_t u(t, x, y) + a\partial_x u(t, x, y) + b\partial_y u(t, x, y) = 0, \tag{3.12}$$

equipped with an appropriate initial condition

$$u(0, x, y) = u^0(x, y),$$

where the function $u(t, x, y) : [0, T] \times \Omega_x \times \Omega_y \to \mathbb{R}$ is assumed to be sufficiently regular and $a, b \in \mathbb{R}$ denote constant scalar values. The spatial domain Ω_y is discretized analogously to Ω_x, i.e. by construction of a uniform spatial grid with a finite number of grid cells $N_y \in \mathbb{N}$ and equidistant spacing $\Delta y = \frac{1}{N_y}$. A spatially discretized approximation to (3.12) is obtained by evaluating $u(t, x, y)$ at the grid points $(x_j, y_i) \in \Omega_x \times \Omega_y$ and setting

$$u_{ji}(t) \approx u(t, x_j, y_i) \qquad \text{for } \; j = 1, ..., N_x, \; i = 1, ..., N_y.$$

For the approximation of first-order spatial derivatives we perform a dimensional splitting as proposed in [LeV02] and apply a centered FD scheme to each spatial direction, i.e. we approximate

$$\begin{aligned} \partial_x u(t, x, y)\big|_{(x,y)=(x_j,y_i)} &\approx \frac{u_{j+1,i}(t) - u_{j-1,i}(t)}{2\Delta x} \qquad \text{and} \\ \partial_y u(t, x, y)\big|_{(x,y)=(x_j,y_i)} &\approx \frac{u_{j,i+1}(t) - u_{j,i-1}(t)}{2\Delta y}. \end{aligned}$$

This leads to the semi-discrete time-continuous reformulation of equation (3.12) as

$$\dot{u}_{ji}(t) = -\frac{a}{2\Delta x}\left(u_{j+1,i}(t) - u_{j-1,i}(t)\right) - \frac{b}{2\Delta y}\left(u_{j,i+1}(t) - u_{j,i-1}(t)\right).$$

Analogously to the 1D case, the spatial stencil matrices $\mathbf{D}^x$ and $\mathbf{D}^y$ approximating first-order spatial derivatives ∂_x and ∂_y, respectively, and incorporating periodic boundary conditions can be constructed. To account for numerical stability, second-order spatial stencil matrices $\mathbf{D}^{xx}$ and $\mathbf{D}^{yy}$, which are derived similar to the 1D setting, can be added.

Alternatives to the centered FD method. Alternative methods for the spatial discretization of PDEs exist. Concerning the FD method, the one-sided *forward* or *backward FD approximation* shall be mentioned. Compared to the second-order accurate centered FD scheme, these approximations are only first-order accurate and therefore not considered in this thesis. Frequently used approaches are also *finite volume methods* and *finite element methods*, for which the reader is referred to standard textbooks such as [LeV02] and [ZTZ13].

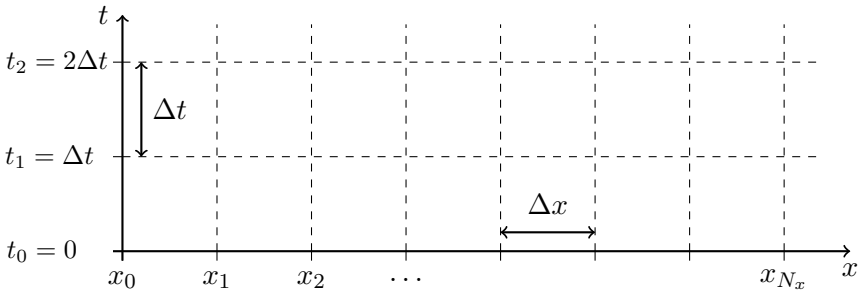


Figure 3.2: Space-time grid with N_x grid cells of width Δx in the spatial variable and N_t grid cells of width Δt in the temporal variable.

3.1.2 Temporal discretization

The derivations presented thus far have yielded a semi-discrete time-continuous form of the hyperbolic advection equation (3.1). To retrieve established theoretical concepts for PDEs in the next section, a discretization in the temporal variable is required. For the discretization of the time interval $[0, T]$ we construct a uniform *temporal grid* with a finite number of grid cells N_t and prescribed grid size Δt. The number of equidistant grid cells N_t is determined from the relation $N_t = \left\lceil \frac{T}{\Delta t} \right\rceil$. An approximation of the semi-discrete time-continuous solution $u_j(t)$ on the temporal grid is obtained by evaluating

$$u_j^n \approx u(t_n, x_j) \qquad \text{for} \quad t_n = n\Delta t, \qquad n = 0, ..., N_t.$$

The resulting space-time grid is illustrated in Figure 3.2. The linear advection equation (3.1) involves temporal derivatives. In this thesis, we focus on *Euler methods* for their approximation.

Explicit forward Euler method. An elementary strategy is the application of the *forward Euler* method. Its basic idea was first introduced in [Eul68] and relies on a Taylor expansion in time. For the linear advection equation (3.1) the fully discrete update formula

$$u_j^{n+1} = u_j^n - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right) \tag{3.13}$$

can be derived from the centered FD discretization given in (3.7). This is a first-order accurate scheme in time and a second-order accurate scheme in space. Note that the time update from time t_n to time $t_{n+1} = t_n + \Delta t$ described in (3.13) only requires knowledge of quantities evaluated at time t_n. Such schemes are called *explicit* methods.

Implicit and implicit-explicit methods. Numerical schemes, on the other hand, that for a time update require the evaluation of quantities at time t_{n+1} (as e.g. the *backward Euler* method) are called *implicit* methods. They are often used to handle a potential *stiffness*, which can introduce numerical instabilities, leading to unacceptably small time step sizes. Implicit methods usually exhibit an improved stability behavior for stiff problems but generally require the numerical solution of more complicated systems, which can involve

coupled dependencies [HW96]. Also the combination of explicit and implicit methods is possible. Such schemes are called *implicit-explicit (IMEX)* methods.

3.2 Numerical stability

This section is devoted to the concept of numerical stability for fully discrete FD schemes. We provide a definition of stability in Section 3.2.1 and introduce necessary and sufficient conditions for the stability of FD schemes in Sections 3.2.2 and 3.2.3. Section 3.2.4 explains the concept of *strong stability*, which is often related to the energy of the system.

3.2.1 Consistency, convergence and stability

We consider one-step FD schemes of the form

$$u_j^{n+1} = \mathcal{N}_{\Delta t,\Delta x} u_j^n, \tag{3.14}$$

where $\mathcal{N}_{\Delta t,\Delta x}$ denotes an FD update operator associated with a given temporal and spatial grid. The indices $\Delta t, \Delta x$ refer to the fixed grid sizes.

Example 3.2. The FD update operator associated with the fully discrete scheme (3.13) for the linear advection equation (3.1) applied to a sufficiently regular function ϕ has the form

$$\mathcal{N}_{\Delta t,\Delta x}\left[\phi\right](t_n, x_j) = \phi^n - \frac{a\Delta t}{2\Delta x}\left(\phi_{j+1}^n - \phi_{j-1}^n\right).$$

Consistency. An important property of a numerical scheme is its consistency with the differential equation. This means that the numerical update operator shall approximate the solution to the continuous equation well locally. This behavior can be quantified by the *local truncation error*.

Definition 3.3 (Local truncation error, [LeV92])**.** For one-step FD schemes associated with an update operator $\mathcal{N}_{\Delta t,\Delta x}$ such as given in (3.14) the quantity $\tau^n = \left(\tau_j^n\right) \in \mathbb{R}^{N_x}$ with

$$\tau_j^n = \frac{1}{\Delta t}\left(u\left(t_{n+1}, x_j\right) - \mathcal{N}_{\Delta t,\Delta x}\left[u\right](t_n, x_j)\right)$$

is called the *local truncation error*.

The local truncation error is used to define the *consistency* of a numerical scheme.

Definition 3.4 (Consistency, [LeV92, Tho95])**.** A one-step FD scheme associated with an update operator $\mathcal{N}_{\Delta t,\Delta x}$ such as given in (3.14) is called *consistent* if, in an appropriate norm $\|\cdot\|$, it holds

$$\|\tau^n\| \to 0 \qquad \text{as} \quad \Delta t, \Delta x \to 0. \tag{3.15}$$

Condition (3.15) provides no information on the rate of convergence. More details are included in the concept of *accuracy*.

Definition 3.5 (Accuracy, [LeV92, Tho95])**.** A one-step FD scheme associated with an update operator $\mathcal{N}_{\Delta t,\Delta x}$ such as given in (3.14) is called *accurate of order p in time and of order q in space* if for any sufficiently regular solution $u(t,x)$ with compactly supported initial condition $u^0(x)$ it holds

$$\|\tau^n\| = \mathcal{O}\left((\Delta t)^p\right) + \mathcal{O}\left((\Delta x)^q\right).$$

Remark 3.6. Accurate schemes of order $p, q \geq 1$ are consistent [Tho95].

Convergence. Another important property for the construction of numerical schemes is the pointwise *convergence* of the numerical solution to the true solution of the PDE as the grid sizes become arbitrarily small.

Definition 3.7 (Convergence, [Str04])**.** A one-step FD scheme associated with an update operator $\mathcal{N}_{\Delta t,\Delta x}$ such as given in (3.14) is called *convergent* if for any solution to the continuous PDE $u(t,x)$ and solutions to the FD scheme u_j^n, such that u_j^0 converges to $u^0(x)$ as $j\Delta x$ converges to x, then u_j^n converges to $u(t,x)$ as $(n\Delta t, j\Delta x)$ converges to (t,x) as $\Delta t, \Delta x$ converge to zero.

Stability. The first part of this thesis focuses on the *stability* of numerical schemes. This concept ensures that errors, which are for instance introduced in the initial condition, do not increase uncontrollably over time and dominate the true behavior of the solution.

Definition 3.8 (Stability, [Tho95, LeV02])**.** A one-step FD scheme associated with an update operator $\mathcal{N}_{\Delta t,\Delta x}$ such as given in (3.14) is *stable* in a stability region Λ with respect to an appropriate norm $\|\cdot\|$ if, for each time T, there is a constant $C_T > 0$ such that

$$\left\|\mathcal{N}_{\Delta t,\Delta x}^n\right\| \leq C_T \qquad \text{for all} \quad 0 \leq n \leq N_t \qquad \text{with} \quad (\Delta t, \Delta x) \in \Lambda. \tag{3.16}$$

Relation between the concepts. For linear FD schemes approximating linear PDEs, for which the corresponding IVP is well-posed, the following relation between the above concepts exist. A rigorous definition of well-posedness of an IVP can be found in [Str04].

Theorem 3.9 (Lax-Richtmeyr equivalence theorem, [Str04])**.** *A consistent and linear FD scheme for a linear PDE, for which the corresponding IVP is well-posed, is convergent if and only if it is stable.*

Proof. See for instance [LR56, Str04]. □

Briefly summarized, the Lax-Richtmeyr equivalence theorem states that for linear PDEs and linear FD methods it holds

$$\text{consistency} \quad + \quad \text{stability} \quad \Longleftrightarrow \quad \text{convergence}.$$

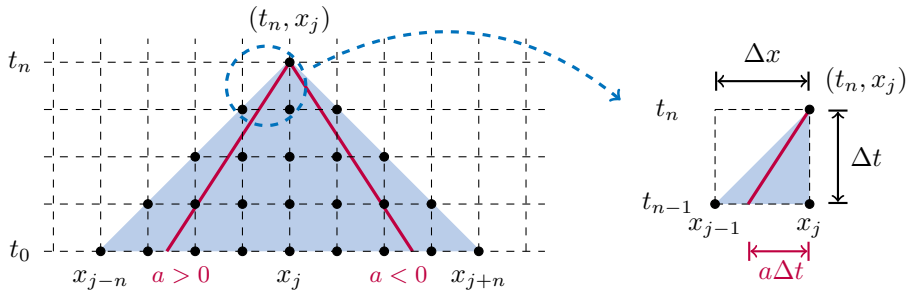


Figure 3.3: Visualization of the numerical domain of dependence and of the CFL condition for the linear advection equation (3.1). **Left:** The numerical domain of dependence for a time-explicit three-point method is depicted exemplarily for the point (t_n, x_j) in blue. The characteristic curves for the linear advection equation (3.1) are added in violet-red for $a > 0$ and $a < 0$. For the FD scheme to satisfy the CFL condition, the characteristic curves must lie inside the blue numerical domain of dependence. **Right:** Zoom into one grid cell, illustrating the concrete CFL condition given in (3.17).

3.2.2 CFL condition

For many schemes showing boundedness of the FD update operator as described in (3.16) is a non-trivial problem. A necessary condition for the stability of FD schemes, which is usually easier to derive, was discovered by Courant, Friedrichs and Lewy in [CFL28]. Analogously to the analytical true domain of dependence, given for instance in (3.2) for the linear advection (3.1), the *numerical domain of dependence* can be defined. For a fixed grid point (t_n, x_j), it contains all grid points x_j at the initial time $t = 0$ for which u_j^0 has an impact on the solution u_j^n. The *Courant-Friedrichs-Lewy (CFL) condition* relates the true and the numerical domain of dependence.

Theorem 3.10 (CFL condition, [LeV07])**.** *A numerical method can only be stable (and hence convergent) if its numerical domain of dependence contains the true domain of dependence of the PDE, at least in the limit as Δt and Δx go to zero.*

Proof. See for instance [CFL28, Str04]. □

In Figure 3.3 the numerical domain of dependence for a time-explicit three-point FD scheme is displayed. The computation of u_j^{n+1} requires knowledge of u_{j-1}^n, u_j^n and u_{j+1}^n. For the linear advection equation (3.1) together with an explicit FD scheme, the CFL condition translates to

$$C_{\mathrm{CFL}} = \left| \frac{a\Delta t}{\Delta x} \right| \leq 1, \tag{3.17}$$

where C_{CFL} is called the *Courant number*. This concrete condition is also illustrated in Figure 3.3. With the insights gained from the derivation of the CFL condition the following theorem was first proven in [CFL28].

Theorem 3.11. *There are no explicit, consistent, unconditionally stable FD schemes for the solution of hyperbolic PDEs.*

Proof. See for instance [CFL28]. □

Note that the CFL condition is a necessary condition for stability (and convergence). To guarantee numerical stability, a thorough stability analysis is still required.

3.2.3 Von Neumann stability

A further method for deriving necessary or even sufficient conditions for the stability of linear FD schemes, which is generally more feasible than the criterion given in (3.16), relies on the application of Fourier analysis. This approach goes back to [CN47, CFvN50] and is commonly referred to as the *von Neumann stability analysis*. The basic idea of Fourier analysis, introduced in [Fou08, Fou22], consists in expanding generally complicated functions in terms of simpler trigonometric expressions. Around this concept an entire theoretical framework has been constructed. More information, for instance on the *continuous Fourier transform*, can be found in standard textbooks such as [Yos95, Tit48]. In this thesis, we restrict our considerations to grid functions $\mathbf{u} = (..., u_{-1}, u_0, u_1, ...)^\top \in \ell^2$.

Definition 3.12 (Fourier transform of a grid function, [Tho95])**.** The *Fourier transform* of a grid function $\mathbf{u} \in \ell^2$ is the 2π-periodic function $\widehat{u} \in L^2[-\pi, \pi]$ defined by

$$\widehat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} e^{-ij\xi} u_j \qquad \text{for } \xi \in [-\pi, \pi],$$

where $i \in \mathbb{C}$ denotes the imaginary unit.

Given the Fourier transform $\widehat{u} \in L^2[-\pi, \pi]$, the original grid function $\mathbf{u} \in \ell^2$ can be uniquely recovered.

Proposition 3.13 (Fourier inversion formula, [Tho95])**.** *Let $\mathbf{u} \in \ell^2$ and $\widehat{u} \in L^2[-\pi, \pi]$ be its Fourier transform. Then,*

$$u_j = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{ij\xi} \widehat{u}(\xi) \, \mathrm{d}\xi. \tag{3.18}$$

Proof. See for instance [Tho95, Tit48]. □

A fundamental result, motivating to work within the L^2-space in Fourier analysis, is Parseval's identity.

Proposition 3.14 (Parseval's identity, [Tho95])**.** *Let $\mathbf{u} \in \ell^2$ and $\widehat{u} \in L^2[-\pi, \pi]$ be its Fourier transform. Then,*

$$\|\widehat{u}\|_2 = \|\mathbf{u}\|_2 .$$

Proof. See for instance [Tho95, Tit48]. □

Stability of linear FD schemes. Concerning stability considerations of linear one-step FD schemes, we insert a Fourier approach as proposed in (3.18) into the update formula

given in (3.14), which is assumed to be extended to infinitely many indices $-\infty < j < \infty$. This leads to an equation of the form

$$\widehat{u}^{n+1}(\xi) = g(\xi)\,\widehat{u}^{n}(\xi),$$

where the scalar value $g(\xi)$ is called the *amplification factor*. The obtained representation is decoupled from all other Fourier modes [RM67].

Theorem 3.15 (von Neumann condition, [Tho95, MM05])**.** *Let us consider a linear FD scheme approximating a linear PDE. A necessary condition for its stability in a stability region Λ is that there exists a constant K such that*

$$|g(\xi)| \leq 1 + K\Delta t \qquad \text{for all } \xi \in [-\pi, \pi] \qquad \text{and } (\Delta t, \Delta x) \in \Lambda. \tag{3.19}$$

For a linear system of equations the amplification factor takes the form of a matrix, which is called the amplification matrix. The von Neumann condition translates to

$$|\lambda_i| \leq 1 + K\Delta t \qquad \text{for all } i,$$

where λ_i are the eigenvalues of the amplification matrix $\mathcal{G}(\xi)$.

Proof. See for instance [Tho95, MM05]. □

Remark 3.16. It can be shown that a discretization of the linear advection equation (3.1) obtained from a centered FD method in the spatial variable and an explicit forward Euler step in the temporal variable is not von Neumann stable [Str04].

The von Neumann condition is a necessary but generally not sufficient condition for the numerical stability of linear FD schemes. However, under certain assumptions it becomes sufficient to ensure stability. The following two results are taken from [RM67].

Theorem 3.17. *If the amplification matrix $\mathcal{G}(\xi)$ is a normal matrix, the von Neumann condition is a sufficient condition for the stability of linear FD schemes.*

Proof. See for instance [RM67, MM05]. □

Corollary 3.18. *In particular, a linear one-step FD scheme approximating a linear scalar PDE with constant coefficients such as the linear advection equation (3.1) that satisfies the von Neumann condition (3.19) is stable.*

Note that the concepts presented in this section have formally been derived for grid functions $\mathbf{u}$ with unrestricted index. For practical applications, limitations to finite sets of grid points are imposed. The above results equivalently translate to this setting [Tho95, Str04].

3.2.4 Energy stability

Another approach for showing the stability of FD schemes, which is also applicable to problems with variable instead of constant coefficients or problems with non-periodic

boundary conditions, is the concept of *energy stability*. Its essential idea consists in deriving a suitable norm for the solution vector $\mathbf{u} \in \mathbb{R}^{N_x}$ so that the norm of the solution stays uniformly bounded over time. In general, the identification of an appropriate norm is a challenging task. In many cases the physical energy associated with the system provides a natural candidate. A comprehensive introduction to energy methods as well as possible generalizations and further results can be found in [RM67, GKO13]. We begin with the definition of *strong stability* of an FD scheme.

Definition 3.19 (Strong stability, [RM67])**.** Let $\mathcal{H}_{\Delta t,\Delta x}$ be an operator acting on $\mathbf{u}$ and K_1 and K_2 be some fixed positive constants. An FD scheme is called *strongly stable* if the following conditions hold:

(i) For every fixed Δt the operator $\mathcal{H}_{\Delta t,\Delta x}$ is well-defined and it holds

$$\frac{1}{K_1} \|\mathbf{u}\|^2 \leq \|\mathbf{u}\|_{\mathcal{H}}^2 \leq K_1 \|\mathbf{u}\|^2 ,$$

where $\|\mathbf{u}\|_{\mathcal{H}}^2 = \sum_{j=1}^{N_x} u_j \mathcal{H}_{\Delta t,\Delta x} u_j$.

(ii) The solution of the FD scheme satisfies

$$\left\|\mathbf{u}^{n+1}\right\|_{\mathcal{H}} \leq (1 + K_2 \Delta t) \left\|\mathbf{u}^n\right\|_{\mathcal{H}} . \tag{3.20}$$

It can easily be seen that strong stability implies the classic stability given in Definition 3.8. Since it explicitly depends on the $\mathcal{H}$-norm, which is typically associated with the energy of the system, we also refer to it as *energy stability*.
For one-step FD schemes associated with an update operator $\mathcal{N}_{\Delta t,\Delta x}$ such as given in (3.14) the condition imposed in (3.20) translates to

$$\|\mathcal{N}_{\Delta t,\Delta x}\|_{\mathcal{H}} \leq 1 + K\Delta t,$$

which is consistent with the boundedness required for stability in (3.16) introduced in Definition 3.8. In general, it can be shown that for problems with constant coefficients and periodic boundary conditions Definition 3.19 is equivalent to Definition 3.8 [RM67]. Further important contributions using the method of energy stability can be found in [Fri54, Lee60, Lax61, Kre63]. In more recent work such as [SN14] for example summation by parts schemes for non-periodic boundary conditions are studied using the energy method.

3.3 Discretization in velocity

In contrast to the macroscopic linear advection equation given in (3.1), kinetic equations as proposed in (2.4) exhibit an additional velocity dependence. To obtain fully discrete numerical schemes for kinetic equations, a discretization in the velocity variable is required.

We restrict our considerations to a 1D setting and to equations of the form

$$\partial_t f(t,x,v) + v\partial_x f(t,x,v) = \mathcal{Q}[f](t,x,v),$$

where $f(t,x,v) : [0,T] \times \Omega_x \times \Omega_v \to \mathbb{R}_0^+$ denotes a distribution function. Section 3.3.1 introduces a *nodal* approach making use of a pointwise approximation of the solution. Section 3.3.2 is devoted to a *modal* approach, expanding the distribution function in terms of orthogonal basis functions. Having performed a discretization in the velocity variable, a discretization in the spatial and the temporal variable as well as stability considerations can be conducted as described in the previous sections.

3.3.1 Nodal approach

For applying the *nodal* approach we discretize the velocity domain Ω_v by constructing a *velocity grid* with a finite number of grid points $N_v \in \mathbb{N}$. An approximation of the distribution function in the 1D velocity variable v is obtained by evaluating $f(t,x,v)$ at each grid point $v_1, ..., v_{N_v} \in \Omega_v$, i.e. by computing

$$f_k(t,x) \approx f(t,x,v_k) \qquad \text{for} \quad k = 1, ..., N_v.$$

Numerical integration. According to Definition 2.5, macroscopic quantities such as the density, mean velocity or temperature are obtained by taking moments of the distribution function. This process involves the evaluation of integrals with respect to the velocity variable v. Let $a, b \in \mathbb{R}$ and let us consider integrals over the interval $[a,b] \subseteq \Omega_v$ of the form

$$I(f) := \int_a^b \omega(v) f(v) \, \mathrm{d}v,$$

where $\omega(v)$ is a given non-negative *weight function* on $[a,b]$. Note that this interval may also be infinite. In accordance to [SB02, Atk89], the weight function $\omega : [a,b] \to \mathbb{R}_0^+$ must accomplish the following properties:

(i) $\omega(v)$ is measurable on the finite or infinite interval $[a,b]$.

(ii) All moments $\int_a^b v^n \omega(v) \, \mathrm{d}v$ exist and are finite for all $n \geq 0$.

(iii) Suppose that

$$\int_a^b \omega(v) g(v) \, \mathrm{d}v = 0$$

for some non-negative continuous function $g(v)$. Then $g(v) \equiv 0$ on $[a,b]$.

For the approximation of the integral $I(f)$ with an appropriate weight function $\omega(v)$, a

quadrature rule of the form

$$I(f) \approx \sum_{k=1}^{N_v} \omega_k f_k \tag{3.21}$$

is sought. The points $v_1, ..., v_{N_v}$ are called the *quadrature nodes* and $\omega_1, ..., \omega_{N_v}$ the associated *quadrature weights*. There are multiple options for the distribution of the quadrature nodes in the scalar case. An intuitive approach is to consider equidistant spacing, analogously to the spatial grid constructed in Section 3.1.1. Examples of this include the *Newton-Cotes formulae* such as the *trapezoidal rule* or *Simpson's rule*. More information on this topic is provided in [Atk89, IK66]. However, more accurate approximations of the integral $I(f)$ can be obtained by allowing the quadrature nodes to be non-uniformly distributed. This leads to *Gaussian quadrature rules* for which the nodes are determined as the roots of orthogonal polynomials. This choice ensures that polynomials up to degree $2n-1$ can be exactly computed [SB02]. Depending on the interval $[a, b]$ and on the weight function $\omega(v)$, different orthogonal polynomials are considered. An overview of common choices can be found in [DR84], a tabular list of numerical values in [AS72].

Gauss-Hermite quadrature. In this thesis, we are interested in integrals evaluated over the whole real line $\mathbb{R}$ of the form

$$I(f) = \int_{\mathbb{R}} e^{-v^2} f(v) \, \mathrm{d}v$$

with weight function $\omega(v) = e^{-v^2}$. The set of orthogonal polynomials associated with this special weight function are the *Hermite polynomials* $\{H_n\}_{n \in \mathbb{N}_0}$, which are defined as

$$H_n(v) = (-1)^n e^{v^2} \frac{\mathrm{d}^n}{\mathrm{d}v^n} e^{-v^2}.$$

Let N_v be the desired quadrature order of the numerical scheme. Then the quadrature nodes $v_1, ..., v_{N_v}$ are determined as the roots of the Hermite polynomial H_{N_v} and the corresponding quadrature weights are obtained by

$$\omega_k = \frac{2^{N_v+1} N_v! \sqrt{\pi}}{\left[H_{N_v+1}(v_k)\right]^2} \qquad \text{for} \quad k = 1, ..., N_v.$$

This choice for the approximation of the integral $I(f)$ in the form as given in (3.21) is called the *Gauss-Hermite quadrature rule*.

Extension to two velocity dimensions. For numerical experiments performed in 2D settings also the approximation of 2D integrals of the form

$$I(f) = \int_{\mathbb{R}\times\mathbb{R}} e^{-|\mathbf{v}|^2} f(\mathbf{v}) \, \mathrm{d}\mathbf{v} = \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\left(v^2+w^2\right)} f(v, w) \, \mathrm{d}v \, \mathrm{d}w \tag{3.22}$$

with $\mathbf{v} = (v, w)^\top \in \mathbb{R}$ is relevant. Regarding the discretization of the velocity domain $\Omega_v \times \Omega_w$, we construct a velocity grid with $N_v \in \mathbb{N}$ grid points in the v direction and $N_w \in \mathbb{N}$ grid points in the w direction. The quadrature nodes in each single velocity dimension is assumed to be derived from the 1D Gauss-Hermite quadrature rule. An approximation of the distribution function $f(t, x, v, w)$ at each grid point $(v_k, w_\ell) \in \Omega_v \times \Omega_w$ is obtained by evaluating

$$f_{k\ell}(t, x) \approx f(t, x, v_k, w_\ell) \qquad \text{for} \quad k = 1, ..., N_v, \ \ell = 1, ..., N_w.$$

Concerning the approximation of the integral (3.22), we follow the ideas presented in [Jäc05] and compute

$$I(f) \approx \sum_{k=1}^{N_v} \sum_{\ell=1}^{N_w} \omega_k \omega_\ell f_{k\ell},$$

where ω_k, ω_ℓ are the corresponding Gauss-Hermite quadrature weights. More information on multivariate Gauss quadrature can be found in [DR84, Str71].

3.3.2 Modal approach

In a *modal* framework the distribution function is expanded in the velocity variable in terms of orthogonal polynomials, which represent the *modes* of the solution. We restrict our considerations to the interval $[-1, 1]$. This is a common choice for radiative transfer problems to which frequently the modal P_N *method* is applied. More information (also on alternatives such as the nodal discrete ordinates S_N *method*) can be found in [BG70, CZ67]. The P_N method makes use of the orthogonal *Legendre polynomials* $\left\{\widetilde{P}_n\right\}_{n \in \mathbb{N}_0}$, which are defined as

$$\widetilde{P}_n(v) = \frac{(-1)^n}{2^n n!} \frac{\mathrm{d}^n}{\mathrm{d}v^n} \left(1 - v^2\right)^n.$$

Together with their standard normalization in $L^2[-1, 1]$, they satisfy an orthogonality condition of the form

$$\int_{-1}^{1} \widetilde{P}_m(v) \widetilde{P}_n(v) \, \mathrm{d}v = \gamma_n^2 \delta_{mn} \qquad \text{with} \quad \gamma_n^2 = \frac{2}{(2n+1)}. \tag{3.23}$$

In addition, the Legendre polynomials fulfill the recurrence relation

$$(n+1) \widetilde{P}_{n+1}(v) = (2n+1) v \widetilde{P}_n(v) - n \widetilde{P}_{n-1}(v).$$

In this thesis, we perform a rescaling of the Legendre polynomials by setting $P_n = \frac{\widetilde{P}_n}{\gamma_n}$ to translate the orthogonality condition (3.23) into an orthonormality condition given by

$$\int_{-1}^{1} P_m(v) P_n(v) \, \mathrm{d}v = \delta_{mn}. \tag{3.24}$$

This rescaling implies a rescaled Legendre polynomial of degree zero $P_0 = \frac{1}{\sqrt{2}}$ as well as a rescaled recurrence relation given as

$$vP_n(v) = \frac{(n+1)\gamma_{n+1}}{(2n+1)\gamma_n} P_{n+1} + \frac{n\gamma_{n-1}}{(2n+1)\gamma_n} P_{n-1}. \tag{3.25}$$

Then the P_N method employs of a finite expansion of the distribution function $f(t,x,v)$ in the velocity variable v with N_v expansion coefficients $u_n(t,x)$, called the *moments*, of the form

$$f(t,x,v) \approx f_{N_v}(t,x,v) = \sum_{n=0}^{N_v-1} u_n(t,x) P_n(v),$$

and relies on this expansion to derive the evolution equations for the moments.

Numerical integration. For the computation of integrals with respect to the velocity variable v, the orthonormality (3.24) of the rescaled Legendre polynomials as well as the recurrence relation (3.25) are used. In preparation for later chapters, we introduce the matrix $\mathbf{A} \in \mathbb{R}^{N_v \times N_v}$ with entries

$$A_{mn} := \int_{-1}^{1} vP_m(v) P_n(v) \, \mathrm{d}v. \tag{3.26}$$

Note that the matrix $\mathbf{A}$ is symmetric and diagonalizable in the form $\mathbf{A} = \mathbf{Q}\mathbf{M}\mathbf{Q}^\top$ with $\mathbf{Q}$ being orthogonal and $\mathbf{M} = \mathrm{diag}(\sigma_0, ..., \sigma_{N_v-1})$. Further we define $|\mathbf{A}| = \mathbf{Q}|\mathbf{M}|\mathbf{Q}^\top$.

Extension to two angular dimensions. Besides the 1D analysis presented in this thesis, numerical experiments are performed in higher dimensions. We consider a velocity vector $\mathbf{v} \in \mathbb{R}^3$ and perform the splitting

$$\mathbf{v} = |\mathbf{v}| \, \mathbf{\Omega},$$

where $|\mathbf{v}|$ denotes the absolute value of the velocity and $\mathbf{\Omega} \in \mathcal{S}^2$ a unit vector in the direction of motion. The unit vector $\mathbf{\Omega}$ is usually given in *spherical coordinates*, depending on the polar angle $\theta \in [0, \pi]$ as well as the azimuthal angle $\varphi \in [0, 2\pi)$. We refer to this expression as the *2D angular representation*. The corresponding 3D Cartesian coordinates can be derived from the angular representation as

$$\Omega_x = \sin\theta\cos\varphi, \qquad \Omega_y = \sin\theta\sin\varphi, \qquad \Omega_z = \cos\theta.$$

For an application of the P_N method to an equation including a 2D angular variable $\mathbf{\Omega} = (\theta, \varphi)$, we introduce the *associated Legendre polynomials* of degree $n \in \mathbb{N}_0$ and order $m = 0, ..., n$. They are defined as

$$\widetilde{P}_n^m(v) = (-1)^m \left(1 - v^2\right)^{m/2} \frac{\mathrm{d}^m}{\mathrm{d}v^m} \widetilde{P}_n(v) \qquad \text{with} \quad v \in [-1, 1].$$

This definition can be generalized to negative integers $-n \leq m < 0$ by

$$\widetilde{P}_n^{-m}(v) = (-1)^m \frac{(n-m)!}{(n+m)!} \widetilde{P}_n^m(v).$$

The associated Legendre polynomials satisfy the orthogonality condition

$$\int_{-1}^{1} \widetilde{P}_n^m(v)\, \widetilde{P}_{n'}^{m'}(v)\, \mathrm{d}v = \frac{2}{2n+1} \frac{(n+m)!}{(n-m)!} \delta_{nn'} \tag{3.27}$$

as well as the recurrence relations

$$(n-m+1)\, \widetilde{P}_{n+1}^m(v) = (2n+1)\, v \widetilde{P}_n(v) - (n+m)\, \widetilde{P}_{n-1}^m(v) \tag{3.28}$$

and

$$\left(v^2-1\right) \frac{\mathrm{d}}{\mathrm{d}v} \widetilde{P}_n^m(v) = nv \widetilde{P}_n^m(v) - (n+m)\, \widetilde{P}_{n-1}^m(v). \tag{3.29}$$

The associated Legendre polynomials are used to derive the complex-valued normalized *spherical harmonics*, which can be given as

$$Y_{nm}(\mathbf{\Omega}) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} \widetilde{P}_n^m(\cos\theta)\, e^{im\varphi}.$$

The spherical harmonics fulfill the orthonormality condition

$$\int_0^{2\pi} \int_{-1}^{1} Y_{nm}(\mathbf{\Omega})\, \overline{Y}_{n'm'}(\mathbf{\Omega})\, \mathrm{d}\cos\theta\, \mathrm{d}\varphi = \delta_{nn'}\delta_{mm'},$$

where $\overline{Y}_{nm}$ denotes the complex conjugate of Y_{nm}, which can be determined from the relation

$$Y_{n,-m}(\mathbf{\Omega}) = (-1)^m \overline{Y}_{nm}(\mathbf{\Omega}).$$

In the context of this thesis, similar to the approach used in [Kus20], we employ a real-valued spherical harmonics basic of the form

$$\widehat{Y}_{nm}(\mathbf{\Omega}) = \begin{cases} \frac{(-1)^m}{\sqrt{2}} \left(Y_{n,-|m|}(\mathbf{\Omega}) + (-1)^m Y_{n|m|}(\mathbf{\Omega})\right), & m < 0, \\ Y_{n0}(\mathbf{\Omega}), & m = 0, \\ \frac{(-1)^m}{\sqrt{2}i} \left(Y_{n,-|m|}(\mathbf{\Omega}) - (-1)^m Y_{n|m|}(\mathbf{\Omega})\right), & m > 0. \end{cases}$$

Then a finite spherical harmonics expansion of the distribution function $f(t, \mathbf{x}, \mathbf{\Omega})$ in the angular variable $\mathbf{\Omega}$ with $(N_{\mathbf{\Omega}}+1)^2$ expansion coefficients $\mathbf{u}_{nm}(t, \mathbf{x})$ is obtained by

$$f(t, \mathbf{x}, \mathbf{\Omega}) \approx f_{N_{\mathbf{\Omega}}}(t, \mathbf{x}, \mathbf{\Omega}) = \sum_{n=0}^{N_{\mathbf{\Omega}}} \sum_{m=-n}^{n} \mathbf{u}_{nm}(t, \mathbf{x})\, \widehat{Y}_{nm}(\mathbf{\Omega}).$$

Integrals with respect to the angular variable $\mathbf{\Omega}$ are evaluated using the orthogonality of the associated Legendre polynomials given in (3.27) as well as the recurrence relations stated in equations (3.28) and (3.29).

4

Dynamical low-rank approximation

The numerical solution of kinetic equations is computationally demanding due to their high dimensionality. An approach to reduce the computational costs and memory requirements is the method of *dynamical low-rank approximation (DLRA)* [KL07]. It provides accurate and efficient approximations of the solution to kinetic PDEs and has recently been applied in various fields of research. For instance, contributions on radiation transport [BEKK24a, FKP25, PMF20, YEHS24], radiation therapy [KS23], plasma physics [EL18, EOP20, EOS23], chemical kinetics [EMP24, PEL23] or Boltzmann type transport equations [BEKK24b, EHY21, DL21, HW22] are available. The review article [EKK+25] provides an overview of recent developments on low-rank methods in kinetic theory.

In Section 4.1 the basic idea of DLRA is explained in a semi-discrete time-continuous matrix setting. Section 4.2 provides an overview of frequently used time integrators, which accomplish the important properties of being exact and robust to small singular values. In Section 4.3 the idea of DLRA is reformulated in a fully continuous setting as the order of discretizing and applying the DLRA method may affect theoretical and numerical results. Section 4.4 is devoted to linear stability results for DLRA schemes and the conservation of physical invariants.

4.1 Basic idea of DLRA

We follow the explanations in [KL07], where the concept of DLRA has been introduced in a semi-discrete time-dependent matrix setting. Let $\mathbf{f}(t) \in \mathbb{R}^{N_x \times N_v}$, depending smoothly on the time parameter t, be the solution to the matrix differential equation

$$\dot{\mathbf{f}}(t) = \mathbf{F}(t, \mathbf{f}(t)), \qquad \mathbf{f}(t_0) = \mathbf{f}^0, \qquad t \geq t_0, \tag{4.1}$$

for which the right-hand side is denoted by $\mathbf{F}(t, \mathbf{f}(t)) : [0, T] \times \mathbb{R}^{N_x \times N_v} \to \mathbb{R}^{N_x \times N_v}$. Then we seek an approximation $\mathbf{f}_r(t) \in \mathbb{R}^{N_x \times N_v}$ of rank r with $r \leq \min\{N_x, N_v\}$ of the matrix $\mathbf{f}(t) \in \mathbb{R}^{N_x \times N_v}$. The set of matrices of rank r constitutes a differentiable manifold, which

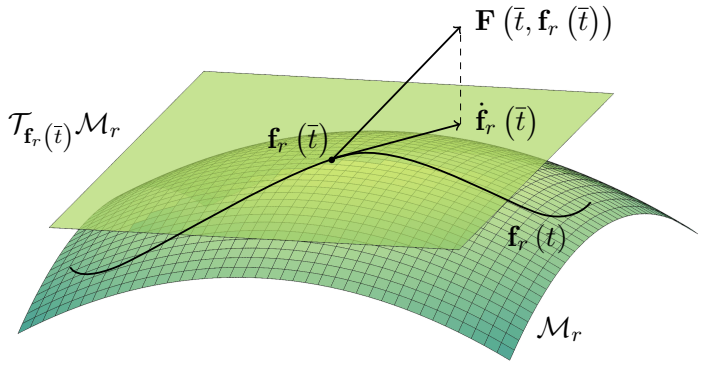


Figure 4.1: Illustration of the basic idea of DLRA. The low-rank manifold $\mathcal{M}_r$ containing a time-dependent low-rank function $\mathbf{f}_r\left(t\right)$ is depicted in dark green. The tangent plane $\mathcal{T}_{\mathbf{f}_r(\bar{t})}\mathcal{M}_r$ at $\mathbf{f}_r\left(\bar{t}\right)$ at some fixed time $\bar{t}$ is depicted in light green. The derivative $\dot{\mathbf{f}}_r\left(\bar{t}\right)$ is required to stay on the tangent plane. This behavior is ensured by an orthogonal projection of $\mathbf{F}\left(\bar{t}, \mathbf{f}_r\left(\bar{t}\right)\right)$ onto the tangent plane.

we denote by $\mathcal{M}_r$ [Pia19, Sch22]. Its corresponding tangent space at $\mathbf{f}_r(t)$ is denoted by $\mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r$. The searched-for approximation $\mathbf{f}_r(t) \in \mathcal{M}_r$ is determined such that at all times t the minimization problem

$$\min_{\dot{\mathbf{f}}_r(t) \in \mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r} \left\| \dot{\mathbf{f}}_r\left(t\right) - \mathbf{F}\left(t, \mathbf{f}_r\left(t\right)\right) \right\|_F \tag{4.2}$$

is fulfilled. Here, $\|\cdot\|_F$ denotes the Frobenius norm. The low-rank approximation $\mathbf{f}_r(t)$ is complemented with an initial condition $\mathbf{f}_r\left(t_0\right) = \mathbf{f}_r^0$, which ideally satisfies $\mathbf{f}_r^0 = \mathbf{f}^0$. If this is not the case, $\mathbf{f}_r^0$ is usually computed as a low-rank approximation of $\mathbf{f}^0$ using a truncated SVD. Following [KL07], the minimization constraint (4.2) on the tangent space is equivalent to determining

$$\dot{\mathbf{f}}_r\left(t\right) = \mathcal{P}\left(\mathbf{f}_r\left(t\right)\right) \mathbf{F}\left(t, \mathbf{f}_r\left(t\right)\right), \tag{4.3}$$

where $\mathcal{P}$ denotes the orthogonal projector onto the tangent space $\mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r$. This approach is visualized in Figure 4.1, which has been created similarly to [Pia19]. Each matrix $\mathbf{f}_r(t) \in \mathcal{M}_r$ can be decomposed into low-rank factors as

$$\mathbf{f}_r\left(t\right) = \mathbf{X}\left(t\right) \mathbf{S}\left(t\right) \mathbf{V}\left(t\right)^\top, \tag{4.4}$$

where the matrices $\mathbf{X}\left(t\right) \in \mathbb{R}^{N_x \times r}$ and $\mathbf{V}\left(t\right) \in \mathbb{R}^{N_v \times r}$ have r orthonormal columns, i.e.

$$\mathbf{X}\left(t\right)^\top \mathbf{X}\left(t\right) = \mathbf{I} \qquad \text{and} \qquad \mathbf{V}\left(t\right)^\top \mathbf{V}\left(t\right) = \mathbf{I},$$

where $\mathbf{I} \in \mathbb{R}^{r \times r}$ denotes the identity matrix. The matrix $\mathbf{X}\left(t\right)$ contains the orthonormal basis functions in space and $\mathbf{V}\left(t\right)$ the orthonormal basis functions in velocity. The matrix $\mathbf{S}\left(t\right) \in \mathbb{R}^{r \times r}$ is assumed to be nonsingular and called the *coefficient* or *coupling matrix*, containing the coefficients of the approximation. Note that $\mathbf{S}\left(t\right)$ is not required to be diagonal and that the representation given in (4.4) is not unique. The orthogonal matrices

$\mathbf{X}(t)$ and $\mathbf{V}(t)$ are contained in the following manifold.

Definition 4.1 (Stiefel manifold and its tangent space, [AMS08, Bou23])**.** The set of matrices with orthonormal columns

$$\mathcal{V}_{N_x,r} = \left\{\mathbf{X} \in \mathbb{R}^{N_x \times r} : \mathbf{X}^\top \mathbf{X} = \mathbf{I}\right\}$$

constitutes an embedded submanifold of $\mathbb{R}^{N_x \times r}$ and is called *Stiefel manifold*. Its tangent space at $\mathbf{X} \in \mathcal{V}_{N_x,r}$ is given as

$$\mathcal{T}_{\mathbf{X}}\mathcal{V}_{N_x,r} = \left\{\dot{\mathbf{X}} \in \mathbb{R}^{N_x \times r} : \dot{\mathbf{X}}^\top \mathbf{X} + \mathbf{X}^\top \dot{\mathbf{X}} = 0\right\}.$$

It can be shown that for time-dependent orthogonal matrices $\mathbf{X}(t) \in \mathcal{V}_{N_x,r}$ their corresponding time derivative is contained in the tangent space $\mathcal{T}_{\mathbf{X}(t)}\mathcal{V}_{N_x,r}$ at $\mathbf{X}(t)$ [Sch22]. Let $\dot{\mathbf{X}}(t) \in \mathcal{T}_{\mathbf{X}(t)}\mathcal{V}_{N_x,r}$ be the time derivative of $\mathbf{X}(t)$ and $\dot{\mathbf{V}}(t) \in \mathcal{T}_{\mathbf{V}(t)}\mathcal{V}_{N_v,r}$ be the time derivative of $\mathbf{V}(t)$, respectively. If for representation (4.4) the additional orthogonality constraints

$$\mathbf{X}(t)^\top \dot{\mathbf{X}}(t) = 0 \qquad \text{and} \qquad \mathbf{V}(t)^\top \dot{\mathbf{V}}(t) = 0$$

are imposed, the elements $\dot{\mathbf{f}}_r(t) \in \mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r$ are uniquely determined and of the form

$$\begin{aligned}\mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r = \{&\dot{\mathbf{f}}_r(t) \in \mathbb{R}^{N_x \times N_v} : \\ &\dot{\mathbf{f}}_r(t) = \dot{\mathbf{X}}(t)\mathbf{S}(t)\mathbf{V}(t)^\top + \mathbf{X}(t)\dot{\mathbf{S}}(t)\mathbf{V}(t)^\top + \mathbf{X}(t)\mathbf{S}(t)\dot{\mathbf{V}}(t)^\top \\ &\text{with } \dot{\mathbf{S}} \in \mathbb{R}^{r\times r}, \dot{\mathbf{X}} \in \mathcal{T}_{\mathbf{X}}\mathcal{V}_{N_x,r}, \dot{\mathbf{V}} \in \mathcal{T}_{\mathbf{V}}\mathcal{V}_{N_v,r} \text{ and } \mathbf{X}^\top\dot{\mathbf{X}} = 0, \mathbf{V}^\top\dot{\mathbf{V}} = 0\}.\end{aligned}$$

It has been proven in [KL07] that deriving a low-rank approximation $\mathbf{f}_r(t) \in \mathcal{M}_r$ of the form given in (4.4), which fulfills the minimization constraint on the tangent space stated in (4.2), is equivalent to determining $\dot{\mathbf{f}}_r(t) \in \mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r$, for which the corresponding low-rank factors are evolved according to

$$\dot{\mathbf{X}}(t) = \left(\mathbf{I} - \mathbf{X}(t)\mathbf{X}(t)^\top\right)\mathbf{F}(t, \mathbf{f}_r(t))\mathbf{V}(t)\mathbf{S}(t)^{-1}, \tag{4.5a}$$

$$\dot{\mathbf{V}}(t) = \left(\mathbf{I} - \mathbf{V}(t)\mathbf{V}(t)^\top\right)\mathbf{F}(t, \mathbf{f}_r(t))^\top\mathbf{X}(t)\mathbf{S}(t)^{-\top}, \tag{4.5b}$$

$$\dot{\mathbf{S}}(t) = \mathbf{X}(t)^\top\mathbf{F}(t, \mathbf{f}_r(t))\mathbf{V}(t). \tag{4.5c}$$

This leads to unique low-rank factors $\mathbf{X}(t), \mathbf{S}(t)$ and $\mathbf{V}(t)$ and a unique representation of the expression given in (4.4). Then the orthogonal projector $\mathcal{P}$ onto $\mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r$ for the solution of (4.3) can be explicitly derived as

$$\begin{aligned}\mathcal{P}(\mathbf{f}_r(t))\mathbf{F}(t, \mathbf{f}_r(t)) = \mathbf{X}\mathbf{X}^\top\mathbf{F}(t, \mathbf{f}_r(t)) - \mathbf{X}\mathbf{X}^\top\mathbf{F}(t, \mathbf{f}_r(t))\mathbf{V}\mathbf{V}^\top \\ + \mathbf{F}(t, \mathbf{f}_r(t))\mathbf{V}\mathbf{V}^\top,\end{aligned} \tag{4.6}$$

where we omit the time-dependency of $\mathbf{X}(t)$ and $\mathbf{V}(t)$ for a better readability.
In the case of very small singular values the matrix $\mathbf{S}(t)$ becomes nearly singular. When solving the evolution equations (4.5) with standard numerical integrators (as e.g. Runge-

Kutta methods) the inversion of the matrix $\mathbf{S}(t)$ in (4.5a) and (4.5b) leads to thorough computational challenges, imposing severe step size restrictions and rendering the algorithm highly unstable. To overcome this problem, different exact and robust time integrators which are able to evolve the low-rank solution on the manifold $\mathcal{M}_r$ while not suffering from potentially small singular values have been introduced. An explanation of two of them is provided in the upcoming section.

4.2 Exact and robust time integrators

For an exact and robust DLRA scheme the implementation of a suitable time integrator is essential. Different such integrators are available [LO14, CL22, CKL22, CKL24]. Section 4.2.1 focuses on the *projector-splitting integrator* introduced in [LO14], whereas Section 4.2.2 is devoted to *BUG integrators* [CL22, CKL22, CKL24], especially the *rank-adaptive augmented BUG integrator* presented in [CKL22].

4.2.1 Projector-splitting integrator

In [LO14] the *projector-splitting integrator* is introduced. Rather than solving the evolution equations (4.5) directly, it relies on the orthogonal projection (4.6) onto the tangent space $\mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_r$. Its main idea is based on the application of splitting methods and the subsequent solution of three subprojections, each of which constitutes a simpler problem than that posed by the original equation. A first-order Lie-Trotter splitting is proposed in [KL07] but also higher-order extensions (e.g. to a second-order Strang splitting scheme) are possible using standard splitting techniques as described in [HLW06].

The projector-splitting integrator evolves the low-rank factors as given in decomposition (4.4) for the solution of the minimization problem (4.2) in the following alternating way: In the first step, the velocity basis $\mathbf{V}$ is fixed while the spatial basis $\mathbf{X}$ and the coefficient matrix $\mathbf{S}$ are updated forward in time. In the second step, the coefficient matrix $\mathbf{S}$ is updated backwards in time with fixed updated spatial basis $\mathbf{X}$ and fixed prior velocity basis $\mathbf{V}$. In the third step, the updated spatial basis $\mathbf{X}$ is fixed while the velocity basis $\mathbf{V}$ and again the coefficient matrix $\mathbf{S}$ are updated forwards in time. In detail, the projector-splitting integrator evolves the low-rank solution from $\mathbf{f}_r^n = \mathbf{X}^n\mathbf{S}^n\mathbf{V}^{n,\top}$ at time t_n to $\mathbf{f}_r^{n+1} = \mathbf{X}^{n+1}\mathbf{S}^{n+1}\mathbf{V}^{n+1,\top}$ at time $t_{n+1} = t_n + \Delta t$ as follows:

***K*-Step**: We fix the velocity basis $\mathbf{V}^n$ at time t_n, denote $\mathbf{K}(t) = \mathbf{X}(t)\mathbf{S}(t) \in \mathbb{R}^{N_x \times r}$, and solve the PDE

$$\dot{\mathbf{K}}(t) = \mathbf{F}\left(t, \mathbf{K}(t)\mathbf{V}^{n,\top}\right)\mathbf{V}^n, \qquad \mathbf{K}(t_n) = \mathbf{X}^n\mathbf{S}^n.$$

Then the spatial basis $\mathbf{X}^n$ is updated to $\mathbf{X}^{n+1} \in \mathbb{R}^{N_x \times r}$ with orthonormal columns by a factorization of $\mathbf{K}(t_{n+1}) = \mathbf{X}^{n+1}\widetilde{\mathbf{S}}^{n+1}$, where $\widetilde{\mathbf{S}}^{n+1} \in \mathbb{R}^{r \times r}$, e.g. by QR-decomposition.

***S*-step**: We fix the spatial basis $\mathbf{X}^{n+1}$ at time t_{n+1}, the velocity basis $\mathbf{V}^n$ at time t_n, and

solve the ordinary differential equation (ODE)

$$\dot{\mathbf{S}}(t) = -\mathbf{X}^{n+1,\top}\mathbf{F}\left(t, \mathbf{X}^{n+1}\mathbf{S}(t)\mathbf{V}^{n,\top}\right)\mathbf{V}^{n}, \qquad \mathbf{S}(t_n) = \widetilde{\mathbf{S}}^{n+1}.$$

Then the coefficient matrix $\widetilde{\mathbf{S}}^{n+1}$ is updated to $\widetilde{\widetilde{\mathbf{S}}}^{n} \in \mathbb{R}^{r\times r}$ by setting $\widetilde{\widetilde{\mathbf{S}}}^{n} = \mathbf{S}(t_{n+1})$.
***L*-Step**: We fix the spatial basis $\mathbf{X}^{n+1}$ at time t_{n+1}, denote $\mathbf{L}(t) = \mathbf{V}(t)\mathbf{S}(t)^{\top} \in \mathbb{R}^{N_v\times r}$, and solve the PDE

$$\dot{\mathbf{L}}(t) = \mathbf{F}\left(t, \mathbf{X}^{n+1}\mathbf{L}(t)^{\top}\right)^{\top}\mathbf{X}^{n+1}, \qquad \mathbf{L}(t_n) = \mathbf{V}^{n}\widetilde{\widetilde{\mathbf{S}}}^{n,\top}.$$

Then the velocity basis $\mathbf{V}^n$ is updated to $\mathbf{V}^{n+1} \in \mathbb{R}^{N_v\times r}$ with orthonormal columns by a factorization of $\mathbf{L}(t_{n+1}) = \mathbf{V}^{n+1}\mathbf{S}^{n+1,\top}$, where $\mathbf{S}^{n+1} \in \mathbb{R}^{r\times r}$, e.g. by QR-decomposition. Altogether, the update of $\mathbf{f}_r^n = \mathbf{X}^n\mathbf{S}^n\mathbf{V}^{n,\top}$ after one time step is given by $\mathbf{f}_r^{n+1} = \mathbf{X}^{n+1}\mathbf{S}^{n+1}\mathbf{V}^{n+1,\top}$.
The proposed projector-splitting integrator has favorable properties compared to the direct solution of the evolution equations (4.5). One of these is the following exactness result for matrices $\mathbf{f}(t)$ of rank r and right-hand sides $\mathbf{F} = \mathbf{F}(t)$.

Theorem 4.2 (Exactness property of the projector-splitting integrator, [LO14])**.** *Let $\mathbf{f}(t) \in \mathbb{R}^{N_x\times N_v}$ be a matrix of rank r for $t_n \le t \le t_{n+1}$, so that $\mathbf{f}(t)$ has a factorization $\mathbf{f}(t) = \mathbf{X}(t)\mathbf{S}(t)\mathbf{V}(t)^{\top}$ as given in (4.4) and let $\mathbf{X}(t_{n+1})^{\top}\mathbf{X}(t_n)$ and $\mathbf{V}(t_{n+1})^{\top}\mathbf{V}(t_n)$ be invertible. With the initial value $\mathbf{f}_r^n = \mathbf{f}(t_n)$, the projector-splitting integrator for $\dot{\mathbf{f}}_r(t) = \mathcal{P}(\mathbf{f}_r(t))\dot{\mathbf{f}}(t)$ with $\dot{\mathbf{f}}(t) = \mathbf{F}(t)$ is exact, i.e. it holds $\mathbf{f}_r^{n+1} = \mathbf{f}(t_{n+1})$.*

Proof. See [LO14]. □

When computing low-rank approximations only small singular values are allowed to be neglected in the approximation in order to prevent important information from getting lost and to retain a good accuracy of the approximation. Let us assume that for a prescribed truncation tolerance parameter ϑ all singular values smaller than ϑ are discarded. Then the smallest retained singular value cannot be expected to be much larger than the largest discarded one as a distinct gap in the singular value distribution cannot be generally assumed. This implies that the coefficient matrix $\mathbf{S}$ still contains entries of order $\mathcal{O}(\vartheta)$, where ϑ is potentially very small. In contrast to standard numerical integrators the projector-splitting integrator is insensitive to small singular values and the following robust error bound can be shown.

Theorem 4.3 (Robust error bound for the projector-splitting integrator, [KLW16])**.** *Let $\mathbf{f}(t) \in \mathbb{R}^{N_x\times N_v}$ be the solution to the matrix differential equation (4.1) and $\mathbf{f}_r^0 \in \mathcal{M}_r$ the initial value of the low-rank approximation. Assume further that the following conditions hold:*

(i) $\mathbf{F}$ *is Lipschitz continuous and bounded, i.e. there exist constants $L, B \ge 0$ such that for all $\mathbf{f}_r(t), \widetilde{\mathbf{f}}_r(t) \in \mathbb{R}^{N_x\times N_v}$ and $0 \le t \le T$ it holds*

$$\left\|\mathbf{F}(t, \mathbf{f}_r(t)) - \mathbf{F}\left(t, \widetilde{\mathbf{f}}_r(t)\right)\right\|_F \le L\left\|\mathbf{f}_r(t) - \widetilde{\mathbf{f}}_r(t)\right\|_F \qquad \text{and} \qquad \|\mathbf{F}(t, \mathbf{f}_r(t))\|_F \le B.$$

(ii) The non-tangential part of $\mathbf{F}(t, \mathbf{f}_r(t))$ *is* ε*-small, i.e. it holds*

$$\left\|\left(\mathbf{I} - \mathcal{P}(\mathbf{f}_r(t))\right)\mathbf{F}(t, \mathbf{f}_r(t))\right\|_F \leq \varepsilon \qquad \text{with} \quad \varepsilon > 0$$

for all $\mathbf{f}_r(t) \in \mathcal{M}_r$ *in a neighborhood of* $\mathbf{f}(t)$ *and* $0 \leq t \leq T$.

(iii) The error in the initial data is δ*-small, i.e. it holds*

$$\left\|\mathbf{f}_r^0 - \mathbf{f}^0\right\|_F \leq \delta \qquad \text{with} \quad \delta > 0.$$

Then the error of the projector-splitting integration scheme at time $t_n = n\Delta t$ *is bounded by*

$$\left\|\mathbf{f}_r^n - \mathbf{f}(t_n)\right\|_F \leq K_1\delta + K_2\varepsilon + K_3\Delta t \qquad \text{for} \quad t_n \leq T, \tag{4.7}$$

where the constants K_i *for* $i = 1, 2, 3$ *only depend on* L, B *and* T. *In particular, the constants* K_i *are independent of the singular values of the exact solution and its low-rank approximation.*

Proof. See [KLW16]. □

In addition, it is shown in [KLW16] that in the case of inexact solutions of the differential equations in the substeps of the splitting scheme, the overall error is bounded similarly as in (4.7). In particular, the bound is independent of the singular values.
Being exact and robust to small singular values are two important properties, distinguishing the projector-splitting integrator from other integration techniques. However, the integration backwards in time in the S-step can lead to numerical instabilities for strongly dissipative problems. Alternative integrators that avoid an integration backwards in time are available and introduced in the next section.

4.2.2 Rank-adaptive augmented basis update & Galerkin integrator

Other integrators that are frequently used for the DLRA approach are the *basis update & Galerkin (BUG) integrator* presented in [CL22] and the *rank-adaptive augmented BUG integrator* introduced in [CKL22]. They both compute all substeps forward in time. In addition, they update the spatial basis functions in the K- and the velocity basis functions in the L-step in parallel, enabling for enhanced parallelization structures and a faster computation of the solution. In contrast to the fixed-rank integrator described in [CKL22], the rank-adaptive augmented BUG integrator discussed in [CKL22] makes uses of certain basis augmentations, hereby allowing for an adaptive choice of the rank in each time step of the evolution. This procedure assists in overcoming the question of identifying a suitable fixed rank, which usually cannot be answered a priori. Also, the required rank may vary over time. Computing with a too small fixed rank leads to poor accuracy results, while for computations with a too large fixed rank too much information is carried and the computational performance deteriorates. In addition, the rank-adaptive

augmented BUG integrator is flexible to basis augmentations, facilitating for instance the implementation of conservation properties.

The rank-adaptive augmented BUG integrator will be used for the subsequently presented DLRA schemes. It evolves the low-rank factors as follows: In the first two steps, it updates and augments the spatial basis $\mathbf{X}$ and the velocity basis $\mathbf{V}$ in parallel, leading to an increase of rank from r to $2r$. Note that augmented quantities of rank $2r$ are denoted with hats. Having the augmented bases at hand, a Galerkin step for the coefficient matrix $\mathbf{S}$ is performed. In the last step, all quantities are truncated back to a new rank $r_{n+1} \leq 2r$, which is adaptively chosen depending on a prescribed error tolerance. In detail, the augmented BUG integrator evolves the low-rank solution from $\mathbf{f}_r^n = \mathbf{X}^n \mathbf{S}^n \mathbf{V}^{n,\top}$ at time t_n to $\mathbf{f}_r^{n+1} = \mathbf{X}^{n+1} \mathbf{S}^{n+1} \mathbf{V}^{n+1,\top}$ at time $t_{n+1} = t_n + \Delta t$ as follows:

***K*-Step**: We fix the velocity basis $\mathbf{V}^n$ at time t_n, denote $\mathbf{K}(t) = \mathbf{X}(t)\,\mathbf{S}(t) \in \mathbb{R}^{N_x \times r}$, and solve the PDE

$$\dot{\mathbf{K}}(t) = \mathbf{F}\left(t, \mathbf{K}(t)\,\mathbf{V}^{n,\top}\right)\mathbf{V}^n, \qquad \mathbf{K}(t_n) = \mathbf{X}^n \mathbf{S}^n.$$

Then the spatial basis $\mathbf{X}^n$ is updated to $\widehat{\mathbf{X}}^{n+1} \in \mathbb{R}^{N_x \times 2r}$ by determining $\widehat{\mathbf{X}}^{n+1}$ as an orthonormal basis of $[\mathbf{K}(t_{n+1}), \mathbf{X}^n] \in \mathbb{R}^{N_x \times 2r}$, e.g. by QR-decomposition. We compute and store $\widehat{\mathbf{M}} = \widehat{\mathbf{X}}^{n+1,\top} \mathbf{X}^n \in \mathbb{R}^{2r \times r}$.

***L*-Step**: We fix the spatial basis $\mathbf{X}^n$ at time t_n, denote $\mathbf{L}(t) = \mathbf{V}(t)\,\mathbf{S}(t)^\top \in \mathbb{R}^{N_v \times r}$, and solve the PDE

$$\dot{\mathbf{L}}(t) = \mathbf{F}\left(t, \mathbf{X}^n \mathbf{L}(t)^\top\right)^\top \mathbf{X}^n, \qquad \mathbf{L}(t_n) = \mathbf{V}^n \mathbf{S}^{n,\top}.$$

Then the velocity basis $\mathbf{V}^n$ is updated to $\widehat{\mathbf{V}}^{n+1} \in \mathbb{R}^{N_v \times 2r}$ by determining $\widehat{\mathbf{V}}^{n+1}$ as an orthonormal basis of $[\mathbf{L}(t_{n+1}), \mathbf{V}^n] \in \mathbb{R}^{N_v \times 2r}$, e.g. by QR-decomposition. We compute and store $\widehat{\mathbf{N}} = \widehat{\mathbf{V}}^{n+1,\top} \mathbf{V}^n \in \mathbb{R}^{2r \times r}$.

***S*-step**: We fix the updated spatial basis $\widehat{\mathbf{X}}^{n+1}$ and the updated velocity basis $\widehat{\mathbf{V}}^{n+1}$ at time t_{n+1}, respectively, and solve the ODE

$$\dot{\widehat{\mathbf{S}}}(t) = \widehat{\mathbf{X}}^{n+1,\top} \mathbf{F}\left(t, \widehat{\mathbf{X}}^{n+1} \widehat{\mathbf{S}}(t)\,\widehat{\mathbf{V}}^{n+1,\top}\right)\widehat{\mathbf{V}}^{n+1}, \qquad \widehat{\mathbf{S}}(t_n) = \widehat{\mathbf{M}} \mathbf{S}^n \widehat{\mathbf{N}}^\top.$$

Then the coefficient matrix $\mathbf{S}^n$ is updated to $\widehat{\mathbf{S}}^{n+1} \in \mathbb{R}^{2r \times 2r}$ by setting $\widehat{\mathbf{S}}^{n+1} = \widehat{\mathbf{S}}(t_{n+1})$.

Truncation: We compute $\widehat{\mathbf{P}} \boldsymbol{\Sigma} \widehat{\mathbf{Q}}^\top = \mathrm{svd}\left(\widehat{\mathbf{S}}^{n+1}\right)$ from an SVD, where $\widehat{\mathbf{P}}, \widehat{\mathbf{Q}} \in \mathbb{R}^{2r \times 2r}$ are orthogonal matrices and $\boldsymbol{\Sigma} \in \mathbb{R}^{2r \times 2r}$ is the diagonal matrix containing the singular values $\sigma_1, ..., \sigma_{2r}$. The new rank $r_{n+1} \leq 2r$ is determined such that

$$\left(\sum_{j=r_{n+1}+1}^{2r} \sigma_j^2\right)^{1/2} \leq \vartheta,$$

where ϑ denotes a prescribed tolerance parameter. Then we set $\mathbf{S}^{n+1} \in \mathbb{R}^{r_{n+1} \times r_{n+1}}$ to be the matrix containing the r_{n+1} largest singular values of $\widehat{\mathbf{S}}^{n+1}$ and the matrices $\mathbf{P}^{n+1}, \mathbf{Q}^{n+1} \in \mathbb{R}^{2r \times r_{n+1}}$ to contain the first r_{n+1} columns of $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{Q}}$, respectively. Finally,

we compute $\mathbf{X}^{n+1} = \widehat{\mathbf{X}}^{n+1}\mathbf{P}^{n+1} \in \mathbb{R}^{N_x \times r_{n+1}}$ and $\mathbf{V}^{n+1} = \widehat{\mathbf{V}}^{n+1}\mathbf{Q}^{n+1} \in \mathbb{R}^{N_v \times r_{n+1}}$.
Altogether, the update of $\mathbf{f}_r^n = \mathbf{X}^n\mathbf{S}^n\mathbf{V}^{n,\top}$ after one time step is given by $\mathbf{f}_r^{n+1} = \mathbf{X}^{n+1}\mathbf{S}^{n+1}\mathbf{V}^{n+1,\top}$. Note that we do not explicitly include the new rank r_{n+1} in the notation of the updated low-rank approximation $\mathbf{f}_r^{n+1}$.
The rank-adaptive augmented BUG integrator accomplishes the important property of exactness for matrices $\mathbf{f}(t)$ of rank r and right-hand sides $\mathbf{F} = \mathbf{F}(t, \mathbf{f}(t))$.

Theorem 4.4 (Exactness property of the rank-adaptive augmented BUG integrator, [CKL22]). *Let $\mathbf{f}(t) \in \mathbb{R}^{N_x \times N_v}$ be a matrix of rank r for $t_n < t < t_{n+1}$ so that a factorization $\mathbf{f}(t) = \mathbf{X}(t)\mathbf{S}(t)\mathbf{V}(t)^\top$ as in (4.4) exists and let $\mathbf{X}(t_{n+1})^\top \mathbf{X}(t_n)$ and $\mathbf{V}(t_{n+1})^\top \mathbf{V}(t_n)$ be invertible. Assume further that the truncation tolerance ϑ is smaller than the r-th singular value of $\mathbf{f}(t_{n+1})$. With the initial value $\mathbf{f}_r^n = \mathbf{f}(t_n)$, the rank-adaptive augmented BUG integrator for $\dot{\mathbf{f}}_r(t) = \mathcal{P}(\mathbf{f}_r(t))\dot{\mathbf{f}}(t)$ with $\dot{\mathbf{f}}(t) = \mathbf{F}(t, \mathbf{f}(t))$ is exact, i.e. it holds $\mathbf{f}_r^{n+1} = \mathbf{f}(t_{n+1})$.*

Proof. See [CKL22]. □

Beyond that, the rank-adaptive augmented BUG integrator is robust to small singular values and the following robust error bound can be given.

Theorem 4.5 (Robust error bound for the rank-adaptive augmented BUG integrator, [CKL22]). *Let $\mathbf{f}(t) \in \mathbb{R}^{N_x \times N_v}$ be the solution to the matrix differential equation (4.1) and $\mathbf{f}_r^0 \in \mathcal{M}_r$ the initial value of the low-rank approximation. Assume further that the following conditions hold:*

(i) $\mathbf{F}$ is Lipschitz continuous and bounded, i.e. there exist constants $L, B \geq 0$ such that for all $\mathbf{f}_r(t), \widetilde{\mathbf{f}}_r(t) \in \mathbb{R}^{N_x \times N_v}$ and $0 \leq t \leq T$ it holds

$$\left\|\mathbf{F}(t, \mathbf{f}_r(t)) - \mathbf{F}\left(t, \widetilde{\mathbf{f}}_r(t)\right)\right\|_F \leq L\left\|\mathbf{f}_r(t) - \widetilde{\mathbf{f}}_r(t)\right\|_F \quad \text{and} \quad \|\mathbf{F}(t, \mathbf{f}_r(t))\|_F \leq B.$$

(ii) The non-tangential part of $\mathbf{F}(t, \mathbf{f}_r(t))$ is ε-small at rank r_n for $\mathbf{f}_r(t)$ near $\mathbf{f}(t)$ and t near t_n, i.e. it holds

$$\|(\mathbf{I} - \mathcal{P}_{r_n}(\mathbf{f}_r(t)))\mathbf{F}(t, \mathbf{f}_r(t))\|_F \leq \varepsilon \qquad \text{with} \quad \varepsilon > 0$$

for all $\mathbf{f}_r(t) \in \mathcal{M}_{r_n}$ in a neighborhood of $\mathbf{f}(t)$ and t near t_n, where $\mathcal{P}_{r_n}$ denotes the orthogonal projector onto the tangent space $\mathcal{T}_{\mathbf{f}_r(t)}\mathcal{M}_{r_n}$ of the manifold $\mathcal{M}_{r_n}$ of matrices of rank r_n at $\mathbf{f}_r(t) \in \mathcal{M}_{r_n}$.

(iii) The error in the initial data is δ-small, i.e. it holds

$$\left\|\mathbf{f}_r^0 - \mathbf{f}^0\right\|_F \leq \delta \qquad \text{with} \quad \delta > 0.$$

Then the error of the rank-adaptive augmented BUG integration scheme at time $t_n = n\Delta t$ is bounded by

$$\|\mathbf{f}_r^n - \mathbf{f}(t_n)\|_F \leq K_1\delta + K_2\varepsilon + K_3\Delta t + K_4 n\vartheta \qquad \text{for} \quad t_n \leq T, \tag{4.8}$$

where the constants K_i for $i = 1, 2, 3, 4$ only depend on L, B and T. In particular, the constants K_i are independent of the singular values of the exact solution and its low-rank approximation.

Proof. See [CKL22]. □

In addition, it can be shown similarly as done in [KLW16] for the projector-splitting integrator that in the case of inexact solutions of the differential equations in the substeps of the splitting scheme, the overall error is bounded similarly as in (4.8). In particular, the bound is independent of the singular values [CKL22].
In [CKL24] the *parallel BUG integrator* has been presented. Its update strategy is similar to the rank-adaptive augmented BUG integrator but allows for a solution of all three substeps fully in parallel. Compared to the presented rank-adaptive augmented BUG integrator, it does not require the basis augmentations to rank $2r$ in the K- and L-step and the solution of a $2r \times 2r$ differential equation in the S-step of the scheme. The enhanced parallelization of all three substeps as well as the reduction from rank $2r$ to r renders this integrator even more efficient. However, the exactness property is not fulfilled but a first-order robust error bound can be established [CKL24].
Extensions to schemes with proven second-order robust error bounds have been proposed in [CEKL24] for the rank-adaptive augmented BUG and in [Kus25] for the parallel integrator.

4.3 DLRA in a fully continuous setting

Thus far, the concept of DLRA has been discussed in a semi-discrete time-dependent matrix framework, in which it was originally introduced in [KL07]. This means that, concerning the space and velocity discretization, a "first discretize, then low-rank" approach has been pursued. In contrast to that, the authors of [EL18] employ a "first low-rank, then discretize" strategy and derive the evolution equations for the low-rank factors in a fully continuous setting. We additionally present this approach as it is used in the subsequently presented work on the thermal RTEs with Su-Olson closure [BEKK24a].
Let the distribution function $f(t, \mathbf{x}, \mathbf{v}) : [0, T] \times \Omega_{\mathbf{x}} \times \Omega_{\mathbf{v}} \to \mathbb{R}_0^+$ be the solution to a given equation

$$\partial_t f(t, \mathbf{x}, \mathbf{v}) = F(t, f(t, \mathbf{x}, \mathbf{v})), \qquad f(t_0, \mathbf{x}, \mathbf{v}) = f^0(\mathbf{x}, \mathbf{v}), \qquad t \geq t_0.$$

We aim for a low-rank approximation of f of the form

$$f_r(t, \mathbf{x}, \mathbf{v}) = \sum_{i,j=1}^{r} X_i(t, \mathbf{x}) S_{ij}(t) V_j(t, \mathbf{v}), \tag{4.9}$$

where $\{X_i(t, \mathbf{x}) : i = 1, ..., r\}$ denotes the set of orthonormal basis functions in space and $\{V_j(t, \mathbf{v}) : j = 1, ..., r\}$ the set of orthonormal basis functions in velocity. They accomplish

the orthogonality relations

$$\langle X_i(t,\mathbf{x}), X_k(t,\mathbf{x})\rangle_{\mathbf{x}} = \delta_{ik} \qquad \text{and} \qquad \langle V_j(t,\mathbf{v}), V_\ell(t,\mathbf{v})\rangle_{\mathbf{v}} = \delta_{j\ell},$$

where $\langle\cdot,\cdot\rangle_{\mathbf{x}}$ and $\langle\cdot,\cdot\rangle_{\mathbf{v}}$ are the inner products on $L^2(\Omega_{\mathbf{x}})$ and $L^2(\Omega_{\mathbf{v}})$, respectively. As the representation given in (4.9) is not unique, the additional Gauge conditions

$$\langle \partial_t X_i(t,\mathbf{x}), X_k(t,\mathbf{x})\rangle_{\mathbf{x}} = 0 \qquad \text{and} \qquad \langle \partial_t V_j(t,\mathbf{v}), V_\ell(t,\mathbf{v})\rangle_{\mathbf{v}} = 0$$

are imposed. Then it can be derived that $\{X_i(t,\mathbf{x})\}$ and $\{V_j(t,\mathbf{v})\}$ are uniquely determined for invertible $\mathbf{S}(t) = (S_{ij}(t)) \in \mathbb{R}^{r\times r}$ [EL18]. This implies that we seek an approximation of f that for each time t lies in the manifold

$$\begin{aligned}\mathcal{M}_r = \Big\{ f_r \in L^2(\Omega_{\mathbf{x}}\times\Omega_{\mathbf{v}}) : f_r(\cdot,\mathbf{x},\mathbf{v}) = \sum_{i,j=1}^{r} X_i(\cdot,\mathbf{x})\, S_{ij}(\cdot)\, V_j(\cdot,\mathbf{v}) \ \text{ with invertible} \\ \mathbf{S} = (S_{ij}) \in \mathbb{R}^{r\times r}, X_i \in L^2(\Omega_{\mathbf{x}}), V_j \in L^2(\Omega_{\mathbf{v}}) \ \text{ and } \ \langle X_i, X_k\rangle_{\mathbf{x}} = \delta_{ik}, \\ \langle V_j, V_\ell\rangle_{\mathbf{v}} = \delta_{j\ell} \Big\}.\end{aligned}$$

Let $f_r(t,\cdot,\cdot)$ be a path on $\mathcal{M}_r$. A formal differentiation of f_r with respect to t leads to

$$\begin{aligned}\dot f_r(t,\cdot,\cdot) = \sum_{i,j=1}^{r} \Big(&\dot X_i(t,\cdot)\, S_{ij}(t)\, V_j(t,\cdot) + X_i(t,\cdot)\, \dot S_{ij}(t)\, V_j(t,\cdot) \\ &+ X_i(t,\cdot)\, S_{ij}(t)\, \dot V_j(t,\cdot) \Big).\end{aligned}$$

These functions restrict the solution dynamics to the low-rank manifold $\mathcal{M}_r$ and constitute the corresponding tangent space, which for fixed time t together with the Gauge conditions reads

$$\begin{aligned}\mathcal{T}_{f_r(t)}\mathcal{M}_r = \Big\{ \dot f_r \in L^2(\Omega_{\mathbf{x}}\times\Omega_{\mathbf{v}}) : \dot f_r(\cdot,\mathbf{x},\mathbf{v}) = \sum_{i,j=1}^{r} \Big(&\dot X_i(\cdot,\mathbf{x})\, S_{ij}(\cdot)\, V_j(\cdot,\mathbf{v}) \\ &+ X_i(\cdot,\mathbf{x})\, \dot S_{ij}(\cdot)\, V_j(\cdot,\mathbf{v}) + X_i(\cdot,\mathbf{x})\, S_{ij}(\cdot)\, \dot V_j(\cdot,\mathbf{v}) \Big) \ \text{with} \\ &\dot S_{ij} \in \mathbb{R}, \dot X_i \in L^2(\Omega_{\mathbf{x}}), \dot V_j \in L^2(\Omega_{\mathbf{v}}) \ \text{ and } \ \left\langle \dot X_i, X_k \right\rangle_{\mathbf{x}} = 0, \left\langle \dot V_j, V_\ell \right\rangle_{\mathbf{v}} = 0 \Big\}.\end{aligned}$$

Having defined the low-rank manifold and its tangent space, the next objective consists in determining $f(t,\cdot,\cdot) \in \mathcal{M}_r$ such that the minimization problem

$$\min_{\partial_t f_r(t,\cdot,\cdot)\in\mathcal{T}_{f_r(t)}\mathcal{M}_r} \|\partial_t f_r(t,\cdot,\cdot) - F(t, f_r(t,\cdot,\cdot))\|_{L^2(\Omega_{\mathbf{x}}\times\Omega_{\mathbf{v}})} \tag{4.10}$$

is solved. For the time evolution of the low-rank factors the following differential equations

can be derived [EL18]:

$$\sum_{i=1}^{r} S_{ij}\partial_t X_i = \langle V_j, F(t, f_r)\rangle_{\mathbf{v}} - \sum_{i=1}^{r} X_i \partial_t S_{ij},$$
$$\sum_{j=1}^{r} S_{ij}\partial_t V_j = \langle X_i, F(t, f_r)\rangle_{\mathbf{x}} - \sum_{j=1}^{r} \partial_t S_{ij} V_j,$$
$$\partial_t S_{ij} = \langle X_i V_j, F(t, f_r)\rangle_{\mathbf{x},\mathbf{v}}.$$

Note that we suppress the arguments for a better readability. Then the minimization constraint (4.10) can be reformulated as the problem of determining $f(t, \mathbf{x}, \mathbf{v})$ such that

$$\partial_t f_r(t, \mathbf{x}, \mathbf{v}) = \mathcal{P}(f_r(t, \mathbf{x}, \mathbf{v}))\, F(f_r(t, \mathbf{x}, \mathbf{v})),$$

where the orthogonal projector $\mathcal{P}$ onto the tangent space $\mathcal{T}_{f_r(t)}\mathcal{M}_r$ can be explicitly given as

$$\mathcal{P}(f_r) F(t, f_r) = \sum_{j=1}^{r} \langle V_j, F(t, f_r)\rangle_{\mathbf{v}} V_j - \sum_{i,j=1}^{r} X_i \langle X_i V_j, F(t, f_r)\rangle_{\mathbf{x},\mathbf{v}} V_j + \sum_{i=1}^{r} X_i \langle X_i, F(t, f_r)\rangle_{\mathbf{x}}.$$

The derivations of the continuous projector-splitting as well as of the continuous (rank-adaptive augmented) BUG integrator are straightforward. An explicit formulation can be found in [EKK+25].

4.4 Linear stability and conservation of physical invariants

A naturally arising question when considering DLRA schemes concerns their numerical stability as well as their behavior related to physical invariants. Section 4.4.1 is devoted to existing linear stability results. In Section 4.4.2 an overview of globally and locally conservative DLRA schemes is provided and a mass conservative truncation strategy is presented.

4.4.1 Linear stability

We begin with an analysis of linear stability of DLRA schemes. At this point, we do not distinguish between the "first discretize, then low-rank" and the "first low-rank, then discretize" approach. Even if the underlying PDE is linear in f, the coupled evolution equations for the low-rank factors, as for instance given in (4.5), are non-linear in $\mathbf{X}, \mathbf{S}$ and $\mathbf{V}$ and it is per se not clear if linear stability concepts can be applied. In [KEC23] it has been shown that the projector-splitting as well as the BUG integrator approximate

the non-linear evolution equations for $\mathbf{X}, \mathbf{S}$ and $\mathbf{V}$ as a series of linear equations due to the fact that in each substep all but one low-rank factor is fixed. In addition, it is known that QR-decompositions and possible truncation steps that are based on an SVD approach are stable in the L^2-norm [EKK+25], making the concepts of linear von Neumann stability analysis as described in Section 3.2.3 applicable. This enables us to derive the stability region of the corresponding DLRA scheme, allowing for a comparison of the DLRA stability region and the stability region of the full-rank problem and for the choice of an optimal time step size, leading to a reduced computational effort.

Linear stability of the presented integrators. Following [KEC23], the order of discretization and application of the DLRA approach makes a crucial difference for the projector-splitting integrator. While the "first discretize, then low-rank" approach is shown to be L^2-unstable due to the discrete S-step backwards in time, the "first low-rank, then discretize" approach, under a certain CFL condition, can lead to an L^2-stable discretization. Indeed, as proven in [KEC23], the BUG integrator is shown to be L^2-stable independently of the order of discretization and derivation of the DLRA equations. This result directly translates to the rank-adaptive augmented BUG integrator [EKK+25].

Energy estimates. Existing work on the stability of DLRA schemes often takes energy estimates as a complementary approach to classic stability considerations into account. Accordingly, in [KS23] an L^2-stability result for radiation therapy is derived. The concept of energy stability, which has been introduced in Section 3.2.4, is treated in [EHK24, FKP25] in the context of DLRA schemes for linear RTEs. This method is not limited to linear equations and a stability result for non-linear thermal radiative transfer is presented in [PK25]. The contributions of this thesis on low-rank discretizations for linear thermal radiative transfer published in [BEKK24a, BEKK25b] and on the linear BGK equation [BEKK24b] also make use of the concept of showing stability estimates in a suitable norm, which may be related to the energy of the underlying system.

4.4.2 Conservation of physical invariants

Since DLRA is a numerical reduction technique it cannot be expected to preserve all relevant information related to the physical system over time and important information ensuring the conservation of physical invariants may get lost. To overcome this problem, techniques for the preservation of conservation properties have been introduced. We distinguish between *local* and *global* conservation laws. While global conservation ensures the preservation of macroscopic quantities such as total mass, total momentum or total energy as described in (2.7), the concept of local conservation guarantees the validity of a local conservation laws as given in (2.6). Global conservation is easier to achieve and is obtained from local conservation by integration over the spatial domain [EL19].

Globally conservative DLRA schemes. A result on the global conservation of mass for the RTE can be found in [PMF20]. In this research article a rescaling of the solution is

performed in each step of the DLRA scheme to ensure the result of the algorithm to match the expected mass that is computed from the underlying equation. This approach is only available for the zeroth order moment. For the preservation of higher order moments the evolution equations have to be adjusted. In [EL19] a DLRA scheme conserving global mass and momentum is proposed for the Vlasov equation. The results derived in [PM21] guarantee global mass and global momentum conservation for the RTE under a suitable modification of the evolution equations.

Locally conservative DLRA schemes. The local preservation of conservation properties is significantly more demanding than the global one as stronger constraints on the dynamics of the system have to be fulfilled. In [EJ21] a weighted L^2-space with a corresponding modification in the L-step of the projector-splitting integrator is introduced, ensuring local conservation of mass, momentum and energy on a continuous level for the Vlasov equation. The proposed integrator is not stable with respect to small singular values but the ideas for its construction have influenced further research. For instance, in [EOS23, GQ24, EKS23] locally conservative DLRA algorithms for the Vlasov equation as well as for the RTE, which incorporate a conservative truncation step, are presented. In this thesis, we employ the rank-adaptive augmented BUG integrator, which is flexible to basis augmentations. In this setting, different from the considerations in [EOS23] and as explained in [EKS23], we do not need to adjust the L-step equation but solely include the basis functions related to the preserved quantities (for instance $\mathbf{v} \to 1$ for mass conservation) in the velocity basis and implement a conservative truncation step. However, this procedure requires an explicit forward Euler step in at least the S-step of the scheme.

Mass conservative truncation strategy. In the following two chapters, we focus on mass conservative DLRA schemes, for which the additional basis augmentations

$$\widehat{\mathbf{X}}^{n+1} = \left[\mathbf{u}_0^{n+1}, \widehat{\mathbf{X}}^{n+1}\right] \in \mathbb{R}^{N_x \times (2r+1)} \qquad \text{and} \qquad \widehat{\mathbf{V}}^{n+1} = \left[\mathbf{e}_1, \widehat{\mathbf{V}}^{n+1}\right] \in \mathbb{R}^{N_v \times (2r+1)}$$

are applied. The vector $\mathbf{u}_0^{n+1}$ denotes the updated zeroth order moment and $\mathbf{e}_1$ the first unit vector in $\mathbb{R}^{N_v}$. They are both stored in the first column of the updated $\widehat{\mathbf{X}}^{n+1}$ and $\widehat{\mathbf{V}}^{n+1}$, respectively. Also the coefficient matrix has to be updated to $\widehat{\mathbf{S}}^{n+1} \in \mathbb{R}^{(2r+1)\times(2r+1)}$ accordingly. In detail, the concrete form of the corresponding updated $\widehat{\mathbf{S}}^{n+1}$ will be explained in Chapter 5 and Chapter 6, respectively. An extension ensuring the preservation of further physical invariants is straightforward and can be achieved as described in [EOS23, EKS23]. The mass conservative truncation strategy proceeds as follows:

(i) We set $\widehat{\mathbf{K}}^{n+1} = \widehat{\mathbf{X}}^{n+1}\widehat{\mathbf{S}}^{n+1}$ and split it into two parts $\widehat{\mathbf{K}}^{n+1} = \left[\widehat{\mathbf{K}}^{n+1,\text{cons}}, \widehat{\mathbf{K}}^{n+1,\text{rem}}\right]$, where $\widehat{\mathbf{K}}^{n+1,\text{cons}}$ corresponds to the first and $\widehat{\mathbf{K}}^{n+1,\text{rem}}$ to the remaining columns of $\widehat{\mathbf{K}}^{n+1}$. Analogously, we split $\widehat{\mathbf{V}}^{n+1}$ into $\widehat{\mathbf{V}}^{n+1} = [\widehat{\mathbf{V}}^{n+1,\text{cons}}, \widehat{\mathbf{V}}^{n+1,\text{rem}}]$, where $\widehat{\mathbf{V}}^{n+1,\text{cons}}$ corresponds to the first and $\widehat{\mathbf{V}}^{n+1,\text{rem}}$ to the remaining columns of $\widehat{\mathbf{V}}^{n+1}$.

(ii) We compute $\widehat{\mathbf{X}}^{n+1,\text{cons}} = \dfrac{\widehat{\mathbf{K}}^{n+1,\text{cons}}}{\left\|\widehat{\mathbf{K}}^{n+1,\text{cons}}\right\|}$ and $\widehat{\mathbf{S}}^{n+1,\text{cons}} = \left\|\widehat{\mathbf{K}}^{n+1,\text{cons}}\right\|$.

(iii) We perform a QR-decomposition to obtain $\widehat{\mathbf{X}}^{n+1,\mathrm{rem}}\widehat{\mathbf{S}}^{n+1,\mathrm{rem}} = \mathrm{qr}\left(\widehat{\mathbf{K}}^{n+1,\mathrm{rem}}\right)$.

(iv) We compute $\widehat{\mathbf{P}}\boldsymbol{\Sigma}\widehat{\mathbf{Q}}^{\top} = \mathrm{svd}\left(\widehat{\mathbf{S}}^{n+1,\mathrm{rem}}\right)$ from an SVD, where $\widehat{\mathbf{P}}, \widehat{\mathbf{Q}} \in \mathbb{R}^{2r\times 2r}$ are orthogonal matrices and $\boldsymbol{\Sigma} \in \mathbb{R}^{2r\times 2r}$ is the diagonal matrix containing the singular values $\sigma_1, ..., \sigma_{2r}$. The new rank $\widetilde{r} \leq 2r$ is determined such that

$$\left(\sum_{j=\widetilde{r}+1}^{2r} \sigma_j^2\right)^{1/2} \leq \vartheta,$$

where ϑ denotes a prescribed tolerance parameter. Then we set $\mathbf{S}^{n+1,\mathrm{rem}} \in \mathbb{R}^{\widetilde{r}\times\widetilde{r}}$ to be the matrix containing the $\widetilde{r}$ largest singular values of $\widehat{\mathbf{S}}^{n+1,\mathrm{rem}}$ and the matrices $\widehat{\mathbf{P}}^{\mathrm{rem}}, \widehat{\mathbf{Q}}^{\mathrm{rem}} \in \mathbb{R}^{2r\times\widetilde{r}}$ to contain the first $\widetilde{r}$ columns of $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{Q}}$, respectively. Finally, we compute $\mathbf{X}^{n+1,\mathrm{rem}} = \widehat{\mathbf{X}}^{n+1,\mathrm{rem}}\widehat{\mathbf{P}}^{\mathrm{rem}} \in \mathbb{R}^{N_x\times\widetilde{r}}$ and $\mathbf{V}^{n+1,\mathrm{rem}} = \widehat{\mathbf{V}}^{n+1,\mathrm{rem}}\widehat{\mathbf{Q}}^{\mathrm{rem}} \in \mathbb{R}^{N_v\times\widetilde{r}}$.

(v) We set $\widetilde{\mathbf{X}}^{n+1} = [\widehat{\mathbf{X}}^{n+1,\mathrm{cons}}, \mathbf{X}^{n+1,\mathrm{rem}}]$ and $\widetilde{\mathbf{V}}^{n+1} = [\widehat{\mathbf{V}}^{n+1,\mathrm{cons}}, \mathbf{V}^{n+1,\mathrm{rem}}]$ and perform a QR-decomposition to obtain $\mathbf{X}^{n+1}\mathbf{R}^1 = \mathrm{qr}\left(\widetilde{\mathbf{X}}^{n+1}\right)$ and $\mathbf{V}^{n+1}\mathbf{R}^2 = \mathrm{qr}\left(\widetilde{\mathbf{V}}^{n+1}\right)$, respectively.

(vi) We compute

$$\mathbf{S}^{n+1} = \mathbf{R}^1 \begin{bmatrix} \widehat{\mathbf{S}}^{n+1,\mathrm{cons}} & 0 \\ 0 & \mathbf{S}^{n+1,\mathrm{rem}} \end{bmatrix} \mathbf{R}^{2,\top}.$$

Then the new rank r_{n+1} is given by $r_{n+1} = \widetilde{r} + 1$.

Altogether, this leads to the updated solution $\mathbf{f}_r^{n+1} = \mathbf{X}^{n+1}\mathbf{S}^{n+1}\mathbf{V}^{n+1,\top}$ after one time step at time $t_{n+1} = t_n + \Delta t$.

Part I

Stability analysis for DLRA schemes

5

A DLRA scheme for the Su-Olson problem

Thermal radiative transfer problems are a class of kinetic transport equations modeling the motion of particles that move through and interact with a background medium, for instance by scattering or absorption. By this interaction the background medium can heat up and itself emit new particles, enforcing the exchange of energy between particles and the background material.
In this chapter, we focus on the thermal radiative transfer equations (RTEs) with *Su-Olson closure*, leading to a linearized coupled internal energy model, for which the corresponding background information is given in Section 5.1. In Section 5.2 the method of DLRA is applied to this system and the continuous DLRA evolution equations obtained with the rank-adaptive augmented BUG integrator are derived. Section 5.3 discretizes the resulting equations in angle and space and provides an energy stability result for the semi-discrete time-continuous system. The main method is presented in Section 5.4, where a provably energy stable space-time discretization is proposed. Local mass conservation, proven in Section 5.5, is achieved by additional basis augmentations and the implementation of a conservative truncation strategy. Numerical experiments explained in Section 5.6 underline the theoretical properties before Section 5.7 provides a short summary and conclusion. The results of this chapter closely follow the presentation in [BEKK24a].

5.1 Thermal radiative transfer equations

The process of thermal radiative transfer is modeled by two coupled equations, the *thermal RTEs*. With absorbing background material they are given in 1D slab geometry by

$$\begin{aligned} \frac{1}{c}\partial_t f(t,x,\mu) + \mu\partial_x f(t,x,\mu) &= \sigma\left(B(t,x) - f(t,x,\mu)\right), \\ \partial_t e(t,x) &= \sigma\left\langle f(t,x,\mu) - B(t,x)\right\rangle_\mu, \end{aligned}$$

where the distribution function $f(t,x,\mu)$ describes the particle density and $e(t,x)$ the internal energy of the material. The variable $t \in \mathbb{R}^+$ denotes time, $x \in \Omega_x \subseteq \mathbb{R}$ represents

the spatial and $\mu \in \Omega_\mu = [-1, 1]$ the angular (or directional) variable. When restricting the 1D velocity variable to the interval $[-1, 1]$, using μ instead of v corresponds to common notation. The opacity σ encodes the rate at which particles are absorbed by the medium and we use brackets $\langle\cdot\rangle_x, \langle\cdot\rangle_\mu$ to indicate an integration over the spatial and the angular domain, respectively. Moreover, the speed of light is denoted by c and the black body radiation at the material temperature T is denoted by $B(T)$. It can be described by the Stefan-Boltzmann law

$$B(T) = acT^4,$$

where $a = \frac{4\sigma_{\mathrm{SB}}}{c}$ is the radiation density constant and σ_{SB} the Stefan-Boltzmann constant. Further information on the thermal RTEs and their relevance in physics can be found in [Pom73, BG70, HMDS20]. The above set of equations is not closed and different closures exist to determine a relation between the temperature T and the internal energy e [OAH00]. We follow the ideas of Pomraning [Pom79] and Su and Olson [SO97] and set $e(T) = \alpha B(T)$. From this point on, we call $\alpha B(T)$ the *internal energy* of the material. Further, we perform a rescaling $\tau = \frac{t}{c}$ and by an abuse of notation write t instead of τ in the remainder. This leads to the system

$$\partial_t f(t,x,\mu) + \mu \partial_x f(t,x,\mu) = \sigma\left(B(t,x) - f(t,x,\mu)\right), \tag{5.1a}$$

$$\partial_t B(t,x) = \sigma\left\langle f(t,x,\mu) - B(t,x)\right\rangle_\mu, \tag{5.1b}$$

where without loss of generality we assume $\alpha = 1$. This system is a closed linearized internal energy model, which is analytically solvable and serves as a common benchmark for numerical considerations [MELD08, MHB08a, MHB08b]. In the remaining thesis, we call equations (5.1) the *Su-Olson problem*. Note that for the moment we omit initial and boundary conditions. In subsequent considerations, our studies include the conservation properties of the derived numerical scheme. For the Su-Olson problem, the mass and the momentum of the system are defined as follows.

Definition 5.1 (Macroscopic quantities)**.** The *mass* and the *momentum* of the Su-Olson problem are defined as

$$\rho(t,x) := \int f(t,x,\mu)\,\mathrm{d}\mu + B(t,x) \qquad \text{and} \qquad \overline{u}(t,x) := \int \mu f(t,x,\mu)\,\mathrm{d}\mu.$$

In particular, the Su-Olson problem satisfies the local conservation law

$$\partial_t \rho(t,x) + \partial_x \overline{u}(t,x) = 0. \tag{5.2}$$

Numerical solution of the thermal RTEs. Constructing numerical schemes for the solution of the Su-Olson problem (5.1) is challenging. First, the potentially stiff opacity term on the right-hand side of both equations presented in (5.1) must be treated by an implicit time integration scheme. Second, for 3D spatial domains the computational costs and memory requirements for finely resolved spatial and angular discretizations become prohibitive. A widely used strategy to address this issue is to choose coarse numerical

discretizations and mitigate numerical artifacts [Lat68, Mat99, MWLP03] which arise due to the insufficient resolution, see e.g. [CFKK19, FKCH20, AS01, Lat71, Ten16]. Despite the success of these approaches in a large number of applications, the requirement of picking user-determined and problem dependent tuning parameters can render them impracticable.

Thermal RTEs and DLRA. Another approach to deal with the high dimensionality of the problem is the application of DLRA methods, which are able to yield accurate solutions while not requiring an expensive offline training phase. Earlier work on radiative transfer with DLRA methods has focused on asymptotic-preserving schemes [EHW21, EHK24, FKP25], mass conservation [PM21], stable discretizations [KEC23], imposing boundary conditions [KS23], and implicit temporal discretizations [PM23]. A discontinuous Galerkin discretization of the DLRA evolution equations for thermal radiative transfer has been proposed in [CFK22]. In this chapter, we focus on energy stability and mass conservation results for the thermal RTEs with Su-Olson closure.

5.2 Continuous DLRA equations for Su-Olson

We begin with a formulation of the continuous DLRA equations for the Su-Olson problem presented in (5.1). The distribution function f is approximated as

$$f(t,x,\mu) \approx \sum_{m,\eta=1}^{r} X_m(t,x)\, S_{m\eta}(t)\, V_\eta(t,\mu),$$

where $\{X_m(t,x) : m = 1,...,r\}$ are the spatial orthonormal basis functions and $\{V_\eta(t,\mu) : \eta = 1,...,r\}$ are the angular orthonormal basis functions. To simplify notation, we identify f with its low-rank approximation f_r and, throughout the following considerations, denote both the full rank and the low-rank solution by f. All theoretical considerations are performed in one spatial and one angular variable. However, an extension to higher dimensions is straightforward.

The rank-adaptive augmented BUG integrator introduced in Section 4.2.2 is employed in its continuous formulation and the corresponding evolution equations for system (5.1) are derived. In the first step, the DLRA evolution equations for the particle density (5.1a) are given as follows:

***K*-step:** We write $K_\eta(t,x) = \sum_{m=1}^{r} X_m(t,x)\, S_{m\eta}(t)$. This leads to the representation $f(t,x,\mu) = \sum_{\eta=1}^{r} K_\eta(t,x)\, V_\eta^n(\mu)$ for the low-rank approximation of the solution, where $\{V_\eta^n(\mu)\}$ denotes the set of angular orthonormal basis functions, which is kept fixed in this step. Inserting this representation of f into (5.1a) and projecting onto $V_p^n(\mu)$ yields the PDE

$$\partial_t K_p(t,x) = -\sum_{\eta=1}^{r} \partial_x K_\eta(t,x) \left\langle V_p^n, \mu V_\eta^n \right\rangle_\mu + \sigma\left(B(t,x)\langle V_p^n\rangle_\mu - K_p(t,x)\right). \tag{5.3a}$$

Together with the initial condition $K_\eta(t_n, x) = \sum_{m=1}^{r} X_m^n(x) S_{m\eta}^n$, the spatial basis functions $X_m^n(x)$ with $m = 1, ..., r$ are updated to $\widehat{X}_m^{n+1}(x)$ with $m = 1, ..., 2r$ by applying Gram Schmidt to $[K_\eta(t_{n+1}, x), X_m^n(x)] = \sum_{i=1}^{2r} \widehat{X}_m^{n+1}(x) R_{m\eta}^1$. Note that $R_{m\eta}^1$ is discarded after this step. We compute and store $\widehat{M}_{mq} = \left\langle \widehat{X}_m^{n+1}, X_q^n \right\rangle_x$.

***L*-step:** We write $L_m(t, \mu) = \sum_{\eta=1}^{r} S_{m\eta}(t) V_\eta(t, \mu)$. This leads to the representation $f(t, x, \mu) = \sum_{m=1}^{r} X_m^n(x) L_m(t, \mu)$ for the low-rank approximation of the solution, $\{X_m^n(x)\}$ denotes the set of spatial orthonormal basis functions, which is kept fixed in this step. Inserting this representation of f into (5.1a) and projecting onto $X_p^n(x)$ yields the PDE

$$\partial_t L_p(t, \mu) = -\mu \sum_{m=1}^{r} \left\langle X_p^n, \frac{\mathrm{d}}{\mathrm{d}x} X_m^n \right\rangle_x L_m(t, \mu) + \sigma \left(\left\langle X_p^n, B(t, x) \right\rangle_x - L_p(t, \mu) \right). \tag{5.3b}$$

Together with the initial condition $L_m(t_n, \mu) = \sum_{\eta=1}^{r} S_{m\eta}^n V_\eta^n(\mu)$, the angular basis functions $V_\eta^n(\mu)$ with $\eta = 1, ..., r$ are updated to $\widehat{V}_\eta^{n+1}(\mu)$ with $\eta = 1, ..., 2r$ by applying Gram Schmidt to $\left[L_m(t_{n+1}, \mu), V_\eta^n(\mu)\right] = \sum_{j=1}^{2r} \widehat{V}_\eta^{n+1}(\mu) R_{m\eta}^2$. Note that $R_{m\eta}^2$ is discarded after this step. We compute and store $\widehat{N}_{\eta p} = \left\langle \widehat{V}_\eta^{n+1}, V_p^n \right\rangle_\mu$.

Lastly the augmented Galerkin step of the rank-adaptive augmented BUG integrator is constructed according to:

***S*-step:** We fix the updated spatial basis functions $\widehat{X}_m^{n+1}$ with $m = 1, ..., 2r$ and the updated angular basis functions $\widehat{V}_\eta^{n+1}$ with $\eta = 1, ..., 2r$ and introduce the notation $\widehat{S}_{m\eta}(t) = \sum_{q,p=1}^{r} \widehat{M}_{mq} S_{qp}(t) \widehat{N}_{\eta p}$. For an update of the entries S_{qp}^n of the coefficient matrix with $q, p = 1, ..., r$ we insert the representation $f(t, x, \mu) = \sum_{m,\eta=1}^{2r} \widehat{X}_m^{n+1}(x) \widehat{S}_{m\eta}(t) \widehat{V}_\eta^{n+1}(\mu)$ into (5.1a) and test against $\widehat{X}_q^{n+1}(x)$ and $\widehat{V}_p^{n+1}(\mu)$. This yields the ODE

$$\begin{aligned} \dot{\widehat{S}}_{qp}(t) = &- \sum_{m,\eta=1}^{2r} \left\langle \widehat{X}_q^{n+1}, \frac{\mathrm{d}}{\mathrm{d}x} \widehat{X}_m^{n+1} \right\rangle_x \widehat{S}_{m\eta}(t) \left\langle \widehat{V}_p^{n+1}, \mu \widehat{V}_\eta^{n+1} \right\rangle_\mu \\ &+ \sigma \left(\left\langle \widehat{X}_q^{n+1}, B(t, x) \right\rangle_x \left\langle \widehat{V}_p^{n+1} \right\rangle_\mu - \widehat{S}_{qp}(t) \right). \end{aligned} \tag{5.3c}$$

Together with the initial condition $\widehat{S}_{m\eta}(t_n) = \sum_{q,p=1}^{r} \widehat{M}_{mq} S_{qp}^n \widehat{N}_{\eta p}$ we obtain the updated augmented quantities $\widehat{S}_{qp}^{n+1}$ with $q, p = 1, ..., 2r$.

For the evolution equation of the internal energy we insert all augmented low-rank factors into (5.1b) and obtain the PDE

$$\partial_t B(t, x) = \sigma \left(\sum_{m,\eta=1}^{2r} \widehat{X}_m^{n+1}(x) \widehat{S}_{m\eta}(t) \left\langle \widehat{V}_\eta^{n+1} \right\rangle_\mu - 2B(t, x) \right). \tag{5.3d}$$

Before repeating this process and evolving the subequations further in time, we truncate the augmented quantities to a new rank $r_{n+1} \leq 2r$ by using a suitable truncation strategy. Note, when employing the rank-adaptive augmented BUG integrator we are not limited to augmenting with the old basis in the K- and L-step.

5.3 Discretization in angle and space

Having derived the K-, L- and S-step of the rank-adaptive augmented BUG integrator for the Su-Olson problem, Sections 5.3.1 and 5.3.2 are devoted to the angular and spatial discretization of the evolution equations. This leads to a semi-discrete time-continuous system, for which energy stability is proven in Section 5.3.3.

5.3.1 Angular discretization

For the angular discretization a modal representation with normalized rescaled Legendre polynomials $P_\ell(\mu)$ as introduced in Section 3.3.2 is employed. The rescaled Legendre polynomials constitute a complete set of orthogonal functions on the interval $[-1,1]$ and satisfy $\langle P_k(\mu), P_\ell(\mu)\rangle_\mu = \delta_{k\ell}$. We approximate

$$V_\eta(t,\mu) \approx \sum_{\ell=0}^{N_\mu-1} V_{\ell\eta}(t) P_\ell(\mu), \quad L_m(t,\mu) \approx \sum_{\ell=0}^{N_\mu-1} L_{\ell m}(t) P_\ell(\mu),$$

and insert these representations into the evolution equations (5.3). We multiply (5.3b) with $P_k(\mu)$ and integrate over μ. In addition, we exploit the fact that with $\mathbf{A} \in \mathbb{R}^{N_\mu \times N_\mu}$ as defined in (3.26) we can rewrite $\left\langle V_p^n(\mu), \mu V_\eta^n(\mu)\right\rangle_\mu = \sum_{k,\ell=0}^{N_\mu-1} V_{\ell\eta}^n A_{k\ell} V_{kp}^n$. Then the evolution equations with angular discretization are given by

$$\partial_t K_p(t,x) = -\sum_{\eta=1}^{r} \partial_x K_\eta(t,x) \sum_{k,\ell=0}^{N_\mu-1} V_{\ell\eta}^n A_{k\ell} V_{kp}^n + \sigma\left(\sqrt{2} B(t,x) V_{0p}^n - K_p(t,x)\right), \tag{5.4a}$$

$$\begin{aligned}\dot{L}_{kp}(t) = &-\sum_{m=1}^{r} \left\langle X_p^n, \frac{\mathrm{d}}{\mathrm{d}x} X_m^n \right\rangle_x \sum_{\ell=0}^{N_\mu-1} L_{\ell m}(t) A_{k\ell} \\ &+ \sigma\left(\left\langle X_p^n, B(t,x)\right\rangle_x \delta_{k0} - L_{kp}(t)\right),\end{aligned} \tag{5.4b}$$

$$\begin{aligned}\dot{\widehat{S}}_{qp}(t) = &-\sum_{m,\eta=1}^{2r} \left\langle \widehat{X}_q^{n+1}, \frac{\mathrm{d}}{\mathrm{d}x} \widehat{X}_m^{n+1} \right\rangle_x \widehat{S}_{m\eta}(t) \sum_{k,\ell=0}^{N_\mu-1} \widehat{V}_{\ell\eta}^{n+1} A_{k\ell} \widehat{V}_{kp}^{n+1} \\ &+ \sigma\left(\sqrt{2}\left\langle \widehat{X}_q^{n+1}, B(t,x)\right\rangle_x \widehat{V}_{0p}^{n+1} - \widehat{S}_{qp}(t)\right).\end{aligned} \tag{5.4c}$$

For the angular discretization of (5.3d) we obtain the equation

$$\partial_t B(t,x) = \sigma\left(\sqrt{2} \sum_{m,\eta=1}^{2r} \widehat{X}_m^{n+1}(x) \widehat{S}_{m\eta}(t) \widehat{V}_{0\eta}^{n+1} - 2B(t,x)\right). \tag{5.4d}$$

5.3.2 Spatial discretization

To derive a spatial discretization, we construct a spatial grid with N_x grid cells and equidistant spacing $\Delta x = \frac{1}{N_x}$. Spatially dependent quantities are approximated at the

grid points x_j for $j = 1, ..., N_x$ and denoted by

$$X_{jp}(t) \approx X_p(t, x_j), \qquad K_{jp}(t) \approx K_p(t, x_j), \qquad B_j(t) \approx B(t, x_j).$$

Assuming periodic boundary conditions, first-order spatial derivatives ∂_x are approximated using the centered FD method. For stability reasons, a diffusion term involving second-order derivatives ∂_{xx} is added. This term is also approximated by the centered FD method. We employ the tridiagonal stencil matrices $\mathbf{D}^x \in \mathbb{R}^{N_x \times N_x}$ given in (3.8) and $\mathbf{D}^{xx} \in \mathbb{R}^{N_x \times N_x}$ defined in (3.11). Recall that the symmetric matrix $\mathbf{A}$ is diagonalizable in the form $\mathbf{A} = \mathbf{Q}\mathbf{M}\mathbf{Q}^\top$ with $\mathbf{Q}$ being orthogonal and $\mathbf{M} = \mathrm{diag}(\sigma_0, ..., \sigma_{N_\mu - 1})$ and that we have defined $|\mathbf{A}| = \mathbf{Q}|\mathbf{M}|\mathbf{Q}^\top$. The following matrix ODEs are obtained:

$$\begin{aligned}
\dot{K}_{jp}(t) = &- \sum_{i=1}^{N_x} D^x_{ji} \sum_{\eta=1}^{r} K_{i\eta}(t) \sum_{k,\ell=0}^{N_\mu - 1} V^n_{\ell\eta} A_{k\ell} V^n_{kp} \\
&+ \frac{\Delta x}{2} \sum_{i=1}^{N_x} D^{xx}_{ji} \sum_{\eta=1}^{r} K_{i\eta}(t) \sum_{k,\ell=0}^{N_\mu - 1} V^n_{\ell\eta} |A|_{k\ell} V^n_{kp} \\
&+ \sigma \left(\sqrt{2} B_j(t) V^n_{0p} - K_{jp}(t) \right), \\
\dot{L}_{kp}(t) = &- \sum_{\ell=0}^{N_\mu - 1} A_{k\ell} \sum_{m=1}^{r} L_{\ell m}(t) \sum_{i,j=1}^{N_x} X^n_{im} D^x_{ij} X^n_{jp} \\
&+ \frac{\Delta x}{2} \sum_{\ell=0}^{N_\mu - 1} |A|_{k\ell} \sum_{m=1}^{r} L_{\ell m}(t) \sum_{i,j=1}^{N_x} X^n_{im} D^{xx}_{ij} X^n_{jp} \\
&+ \sigma \left(\delta_{k0} \sum_{j=1}^{N_x} B_j(t) X^n_{jp} - L_{kp}(t) \right), \\
\dot{\widehat{S}}_{qp}(t) = &- \sum_{i,j=1}^{N_x} \widehat{X}^{n+1}_{jq} D^x_{ji} \sum_{m,\eta=1}^{2r} \widehat{X}^{n+1}_{im} \widehat{S}_{m\eta}(t) \sum_{k,\ell=0}^{N_\mu - 1} \widehat{V}^{n+1}_{\ell\eta} A_{k\ell} \widehat{V}^{n+1}_{kp} \\
&+ \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} \widehat{X}^{n+1}_{jq} D^{xx}_{ji} \sum_{m,\eta=1}^{2r} \widehat{X}^{n+1}_{im} \widehat{S}_{m\eta}(t) \sum_{k,\ell=0}^{N_\mu - 1} \widehat{V}^{n+1}_{\ell\eta} |A|_{k\ell} \widehat{V}^{n+1}_{kp} \\
&+ \sigma \left(\sqrt{2} \sum_{j=1}^{N_x} \widehat{X}^{n+1}_{jq} B_j(t) \widehat{V}^{n+1}_{0p} - \widehat{S}_{qp}(t) \right).
\end{aligned} \tag{5.5a, 5.5b, 5.5c}$$

Lastly, for the internal energy B we receive the spatially discretized equation

$$\begin{aligned}
\partial_t B_j(t) &= \sigma \left(\sqrt{2} \sum_{m,\eta=1}^{2r} \widehat{X}^{n+1}_{jm} \widehat{S}_{m\eta}(t) \widehat{V}^{n+1}_{0\eta} - 2B_j(t) \right) \\
&= \sigma \left(\sqrt{2} u^{n+1}_{j0}(t) - 2B_j(t) \right),
\end{aligned} \tag{5.5d}$$

where we use the notation $\sum_{m,\eta=1}^{2r} \widehat{X}^{n+1}_{jm} \widehat{S}_{m\eta}(t) \widehat{V}^{n+1}_{k\eta} =: u^{n+1}_{jk}(t)$.

5.3.3 Energy stability of the semi-discrete system

The aim of this section is showing energy stability of the semi-discrete time-continuous system (5.5). First, a formal definition of the *total energy* of the system is presented.

Definition 5.2 (Total energy)**.** Let us denote $\mathbf{u}^{n+1}(t) = \left(u_{jk}^{n+1}(t)\right) \in \mathbb{R}^{N_x \times N_\mu}$ for the particle density and $\mathbf{B}(t) = (B_j(t)) \in \mathbb{R}^{N_x}$ for the internal energy. The quantity

$$E(t) := \frac{1}{2} \left\| \mathbf{u}^{n+1}(t) \right\|_F^2 + \frac{1}{2} \left\| \mathbf{B}(t) \right\|_E^2,$$

where $\|\cdot\|_F$ denotes the Frobenius and $\|\cdot\|_E$ the Euclidean norm, is called the *total energy* of the system (5.5).

Then, dissipation of the total energy can be shown for system (5.5).

Theorem 5.3 (Energy stability of the semi-discrete system)**.** *The semi-discrete time-continuous system (5.5) is energy stable, i.e. it holds* $\dot{E}(t) \leq 0$.

Proof. Let us start from the S-step presented in (5.5c), which is given by

$$\begin{aligned}
\dot{\widehat{S}}_{qp}(t) = &- \sum_{i,j=1}^{N_x} \widehat{X}_{jq}^{n+1} D_{ji}^{x} \sum_{m,\eta=1}^{2r} \widehat{X}_{im}^{n+1} \widehat{S}_{m\eta}(t) \sum_{k,\ell=0}^{N_\mu - 1} \widehat{V}_{\ell\eta}^{n+1} A_{k\ell} \widehat{V}_{kp}^{n+1} \\
&+ \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} \widehat{X}_{jq}^{n+1} D_{ji}^{xx} \sum_{m,\eta=1}^{2r} \widehat{X}_{im}^{n+1} \widehat{S}_{m\eta}(t) \sum_{k,\ell=0}^{N_\mu - 1} \widehat{V}_{\ell\eta}^{n+1} |A|_{k\ell} \widehat{V}_{kp}^{n+1} \\
&+ \sigma \left(\sqrt{2} \sum_{j=1}^{N_x} \widehat{X}_{jq}^{n+1} B_j(t) \widehat{V}_{0p}^{n+1} - \widehat{S}_{qp}(t) \right).
\end{aligned}$$

We multiply with $\widehat{X}_{\alpha q}^{n+1} \widehat{V}_{\beta p}^{n+1}$, where $\alpha = 1, ..., N_x$ and $\beta = 0, ..., N_\mu - 1$, and sum over q and p. Further, the projection operators $P_{\alpha j}^{X^{n+1}} = \sum_{q=1}^{2r} \widehat{X}_{\alpha q}^{n+1} \widehat{X}_{jq}^{n+1}$ and $P_{k\beta}^{V^{n+1}} = \sum_{p=1}^{2r} \widehat{V}_{kp}^{n+1} \widehat{V}_{\beta p}^{n+1}$ are introduced. We obtain the representation

$$\begin{aligned}
\dot{u}_{\alpha\beta}^{n+1}(t) = &- \sum_{i,j=1}^{N_x} P_{\alpha j}^{X^{n+1}} D_{ji}^{x} \sum_{k,\ell=0}^{N_\mu - 1} u_{i\ell}^{n+1}(t) A_{k\ell} P_{k\beta}^{V^{n+1}} \\
&+ \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} P_{\alpha j}^{X^{n+1}} D_{ji}^{xx} \sum_{k,\ell=0}^{N_\mu - 1} u_{i\ell}^{n+1}(t) |A|_{k\ell} P_{k\beta}^{V^{n+1}} \\
&+ \sigma \left(\sqrt{2} \sum_{j=1}^{N_x} P_{\alpha j}^{X^{n+1}} B_j(t) \delta_{k0} P_{k\beta}^{V^{n+1}} - u_{\alpha\beta}^{n+1}(t) \right).
\end{aligned}$$

In the next step, we multiply with $u_{\alpha\beta}^{n+1}(t)$ and sum over α and β. Note that it holds

$$\sum_{\alpha=1}^{N_x} P_{\alpha j}^{X^{n+1}} u_{\alpha\beta}^{n+1}(t) = u_{j\beta}^{n+1}(t) \qquad \text{and} \qquad \sum_{\beta=0}^{N_\mu - 1} P_{k\beta}^{V^{n+1}} u_{j\beta}^{n+1}(t) = u_{jk}^{n+1}(t).$$

This leads to

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{u}^{n+1}(t)\right\|_F^2 = &-\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1}(t)\, D_{ji}^x u_{i\ell}^{n+1}(t)\, A_{k\ell} \\ &+\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1}(t)\, D_{ji}^{xx} u_{i\ell}^{n+1}(t)\, |A|_{k\ell} \\ &+\sigma\left(\sqrt{2}\sum_{j=1}^{N_x} u_{jk}^{n+1}(t)\, B_j(t)\,\delta_{k0} - \left\|\mathbf{u}^{n+1}(t)\right\|_F^2\right).\end{aligned}$$

Recall that the matrix $\mathbf{A}$ can be decomposed as $\mathbf{A} = \mathbf{Q}\mathbf{M}\mathbf{Q}^\top$ with $\mathbf{Q}$ being a orthogonal matrix and $\mathbf{M} = \mathrm{diag}(\sigma_0, ..., \sigma_{N_\mu-1})$. Inserting this representation gives

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{u}^{n+1}(t)\right\|_F^2 = &-\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1}(t)\, D_{ji}^x u_{i\ell}^{n+1}(t) \sum_{m=0}^{N_\mu-1} Q_{\ell m}\sigma_m Q_{km} \\ &+\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1}(t)\, D_{ji}^{xx} u_{i\ell}^{n+1}(t) \sum_{m=0}^{N_\mu-1} Q_{\ell m}|\sigma_m| Q_{km} \\ &+\sigma\left(\sqrt{2}\sum_{j=1}^{N_x} u_{jk}^{n+1}(t)\, B_j(t)\,\delta_{k0} - \left\|\mathbf{u}^{n+1}(t)\right\|_F^2\right) \\ = &-\sum_{m=0}^{N_\mu-1}\sigma_m\sum_{i,j=1}^{N_x}\widetilde{u}_{jm}^{n+1}(t)\, D_{ji}^x \widetilde{u}_{im}^{n+1}(t) \\ &+\frac{\Delta x}{2}\sum_{m=0}^{N_\mu-1}|\sigma_m|\sum_{i,j=1}^{N_x}\widetilde{u}_{jm}^{n+1}(t)\, D_{ji}^{xx}\widetilde{u}_{im}^{n+1}(t) \\ &+\sigma\left(\sqrt{2}\sum_{j=1}^{N_x} u_{jk}^{n+1}(t)\, B_j(t)\,\delta_{k0} - \left\|\mathbf{u}^{n+1}(t)\right\|_F^2\right),\end{aligned}$$

where $\widetilde{u}_{jm}^{n+1}(t) := \sum_{k=0}^{N_\mu-1} u_{jk}^{n+1}(t)\, Q_{km}$. Using the properties of the stencil matrices shown in Lemma 3.1, we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{u}^{n+1}(t)\right\|_F^2 = &-\frac{\Delta x}{2}\sum_{j=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(\sum_{i=1}^{N_x}\sum_{k=0}^{N_\mu-1} D_{ji}^{+} u_{ik}^{n+1}(t)\,|A|_{k\ell}^{1/2}\right)^2 \\ &+\sigma\left(\sqrt{2}\sum_{j=1}^{N_x} u_{jk}^{n+1}(t)\, B_j(t)\,\delta_{k0} - \left\|\mathbf{u}^{n+1}(t)\right\|_F^2\right).\end{aligned} \tag{5.6}$$

In the following step, we consider equation (5.5d). Multiplication with $B_j(t)$ and sum-

mation over j yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{B}\left(t\right)\right\|_E^2 = \sigma\left(\sqrt{2}\sum_{j=1}^{N_x} u_{j0}^{n+1}\left(t\right)B_j\left(t\right) - 2\left\|\mathbf{B}\left(t\right)\right\|_E^2\right). \tag{5.7}$$

Adding the evolution equations (5.6) and (5.7) and using the concept of the total energy provided in Definition 5.2, leads to

$$\begin{aligned}
\dot{E}\left(t\right) = &-\frac{\Delta x}{2}\sum_{j=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(\sum_{i=1}^{N_x}\sum_{k=0}^{N_\mu-1} D_{ji}^{+}u_{ik}^{n+1}\left(t\right)\left|A\right|_{k\ell}^{1/2}\right)^2 \\
&+\sigma\left(\sqrt{2}\sum_{j=1}^{N_x} u_{j0}^{n+1}\left(t\right)B_j\left(t\right) - \left\|\mathbf{u}^{n+1}\left(t\right)\right\|_F^2\right) \\
&+\sigma\left(\sqrt{2}\sum_{j=1}^{N_x} u_{j0}^{n+1}\left(t\right)B_j\left(t\right) - 2\left\|\mathbf{B}\left(t\right)\right\|_E^2\right) \\
= &-\frac{\Delta x}{2}\sum_{j=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(\sum_{i=1}^{N_x}\sum_{k=0}^{N_\mu-1} D_{ji}^{+}u_{ik}^{n+1}\left(t\right)\left|A\right|_{k\ell}^{1/2}\right)^2 \\
&-\sigma\left(\sum_{j=1}^{N_x}\left(u_{j0}^{n+1}\left(t\right) - \sqrt{2}B_j\left(t\right)\right)^2 + \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\left(t\right)\right)^2\left(1-\delta_{k0}\right)\right) \leq 0,
\end{aligned}$$

where in the last step we have rewritten $\left\|\mathbf{u}^{n+1}\left(t\right)\right\|_F^2 = \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\left(t\right)\right)^2$ and $\left\|\mathbf{B}\left(t\right)\right\|_E^2 = \sum_{j=1}^{N_x}\left(B_j\left(t\right)\right)^2$. The expression obtained is non-positive, which means that E is dissipated in time. Hence, the system is energy stable. □

5.4 Discretization in time

The aim of this section is the construction of a conservative DLRA scheme that is energy stable under a sharp time step restriction. First, the definition of the total energy is extended to the fully discrete framework.

Definition 5.4 (Fully discrete total energy)**.** Let $\mathbf{u}^n = \left(u_{jk}^n\right) \in \mathbb{R}^{N_x\times N_\mu}$ with entries $u_{jk}^n = \sum_{m,\eta=1}^{r} X_{jm}^n S_{m\eta}^n V_{k\eta}^n$ and $\mathbf{B}^n = \left(B_j^n\right) \in \mathbb{R}^{N_x}$. The quantity

$$E^n := \frac{1}{2}\left\|\mathbf{u}^n\right\|_F^2 + \frac{1}{2}\left\|\mathbf{B}^n\right\|_E^2$$

is called the *fully discrete total energy at time* t_n.

Constructing temporally discretized schemes that preserve the energy dissipation shown in Theorem 5.3 while not suffering from the potentially stiff opacity term is not trivial.

In fact, as shown in Section 5.4.1, a naive IMEX time discretization may increase the total energy. This unphysical behavior is overcome by carefully constructing an energy stable space-time discretization in Section 5.4.2, for which rigorous mathematical proofs are given.

5.4.1 Naive temporal discretization

The analysis starts from system (5.5) which still depends continuously on time. For the temporal discretization, a naive IMEX Euler scheme performing a splitting of internal energy and radiation transport is applied. This means that we use an explicit Euler step for the transport part of the evolution equations, treat the internal energy B explicitly and apply an implicit Euler step for the radiation absorption term. The evolution from time t_n to time $t_{n+1} = t_n + \Delta t$ is described as follows:

$$\begin{aligned} K_{jp}^{n+1} &= K_{jp}^n - \Delta t \sum_{i=1}^{N_x} D_{ji}^x \sum_{\eta=1}^{r} K_{i\eta}^n \sum_{k,\ell=0}^{N_\mu-1} V_{\ell\eta}^n A_{k\ell} V_{kp}^n \\ &\quad + \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} D_{ji}^{xx} \sum_{\eta=1}^{r} K_{i\eta}^n \sum_{k,\ell=0}^{N_\mu-1} V_{\ell\eta}^n \left|A\right|_{k\ell} V_{kp}^n \\ &\quad + \sigma \Delta t \left(\sqrt{2} B_j^n V_{0p}^n - K_{jp}^{n+1} \right), \end{aligned} \tag{5.8a}$$

$$\begin{aligned} L_{kp}^{n+1} &= L_{kp}^n - \Delta t \sum_{\ell=0}^{N_\mu-1} A_{k\ell} \sum_{m=1}^{r} L_{\ell m}^n \sum_{i,j=1}^{N_x} X_{im}^n D_{ji}^x X_{jp}^n \\ &\quad + \Delta t \frac{\Delta x}{2} \sum_{\ell=0}^{N_\mu-1} \left|A\right|_{k\ell} \sum_{m=1}^{r} L_{\ell m}^n \sum_{i,j=1}^{N_x} X_{im}^n D_{ji}^{xx} X_{jp}^n \\ &\quad + \sigma \Delta t \left(\delta_{k0} \sum_{j=1}^{N_x} B_j^n X_{jp}^n - L_{kp}^{n+1} \right). \end{aligned} \tag{5.8b}$$

The augmented and time-updated spatial basis $\widehat{X}_{jp}^{n+1}$ and velocity basis $\widehat{V}_{kp}^{n+1}$ are obtained from a QR-decomposition of the augmented quantities $\widehat{X}_{jp}^{n+1} = \mathrm{qr}\left(\left[K_{jp}^{n+1}, X_{jp}^n\right]\right)$ and $\widehat{V}_{kp}^{n+1} = \mathrm{qr}\left(\left[L_{kp}^{n+1}, V_{kp}^n\right]\right)$, according to the rank-adaptive augmented BUG integrator. Lastly, a Galerkin step for the augmented bases is performed according to

$$\begin{aligned} \widehat{S}_{qp}^{n+1} &= \widetilde{S}_{qp}^n - \Delta t \sum_{i,j=1}^{N_x} \widehat{X}_{jq}^{n+1} D_{ji}^x \sum_{m,\eta=1}^{2r} \widehat{X}_{im}^{n+1} \widetilde{S}_{m\eta}^n \sum_{k,\ell=0}^{N_\mu-1} \widehat{V}_{\ell\eta}^{n+1} A_{k\ell} \widehat{V}_{kp}^{n+1} \\ &\quad + \Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} \widehat{X}_{jq}^{n+1} D_{ji}^{xx} \sum_{m,\eta=1}^{2r} \widehat{X}_{im}^{n+1} \widetilde{S}_{m\eta}^n \sum_{k,\ell=0}^{N_\mu-1} \widehat{V}_{\ell\eta}^{n+1} \left|A\right|_{k\ell} \widehat{V}_{kp}^{n+1} \\ &\quad + \sigma \Delta t \left(\sqrt{2} \sum_{j=1}^{N_x} \widehat{X}_{jq}^{n+1} B_j^n \widehat{V}_{0p}^{n+1} - \widehat{S}_{qp}^{n+1} \right), \end{aligned} \tag{5.8c}$$

where $\widetilde{S}^n_{m\eta} = \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\sum_{q,p=1}^{r}\widehat{X}^{n+1}_{jm}X^n_{jq}S^n_{qp}V^n_{kp}\widehat{V}^{n+1}_{k\eta}$. The internal energy B is updated through

$$\begin{aligned} B^{n+1}_j &= B^n_j + \sigma\Delta t\left(\sqrt{2}\sum_{m,\eta=1}^{2r}\widehat{X}^{n+1}_{jm}\widehat{S}^{n+1}_{m\eta}\widehat{V}^{n+1}_{0\eta} - 2B^{n+1}_j\right) \\ &= B^n_j + \sigma\Delta t\left(\sqrt{2}u^{n+1}_{j0} - 2B^{n+1}_j\right). \end{aligned} \tag{5.8d}$$

However, in Theorem 5.5 we prove that this numerical method exhibits the undesirable property of potentially increasing the total energy during a single time step. This behavior is inconsistent with the governing physical principles.

Theorem 5.5. *There exist initial value pairs* $(\mathbf{u}^n, \mathbf{B}^n)$ *and time step sizes* Δt *such that the naive scheme* (5.8) *results in* $\left(\mathbf{u}^{n+1}, \mathbf{B}^{n+1}\right)$ *for which the fully discrete total energy increases, i.e. for which* $E^{n+1} > E^n$.

Proof. We multiply the S-step equation given in (5.8c) with $\widehat{X}^{n+1}_{\alpha q}\widehat{V}^{n+1}_{\beta p}$ and sum over q and p. Together with the projection operators $P^{X^{n+1}}_{\alpha j} = \sum_{q=1}^{2r}\widehat{X}^{n+1}_{\alpha q}\widehat{X}^{n+1}_{jq}$ and $P^{V^{n+1}}_{k\beta} = \sum_{p=1}^{2r}\widehat{V}^{n+1}_{kp}\widehat{V}^{n+1}_{\beta p}$ and the definition of $\widetilde{S}^n_{m\eta}$, we obtain

$$\begin{aligned} u^{n+1}_{\alpha\beta} = u^n_{\alpha\beta} &- \Delta t\sum_{i,j=1}^{N_x}P^{X^{n+1}}_{\alpha j}D^x_{ji}\sum_{k,\ell=0}^{N_\mu-1}u^n_{i\ell}A_{k\ell}P^{V^{n+1}}_{k\beta} \\ &+ \Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}P^{X^{n+1}}_{\alpha j}D^{xx}_{ji}\sum_{k,\ell=0}^{N_\mu-1}u^n_{i\ell}\left|A\right|_{k\ell}P^{V^{n+1}}_{k\beta} \\ &+ \sigma\Delta t\left(\sqrt{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\widehat{P}^{X^{n+1}}_{\alpha j}B^n_j\delta_{k0}\widehat{P}^{Vn+1}_{k\beta} - u^{n+1}_{\alpha\beta}\right). \end{aligned} \tag{5.9}$$

Let us choose a solution $\mathbf{u}$ and an internal energy $\mathbf{B}$ which at all times are constant in space. Then all terms in (5.9) containing the stencil matrices $\mathbf{D}^x$ and $\mathbf{D}^{xx}$ drop out. In addition, we conclude that all projections in the last term of (5.9) are exact since B^n_j is constant in space and δ_{k0} lies in the span of the basis. Hence, it follows that

$$u^{n+1}_{\alpha\beta} = u^n_{\alpha\beta} + \sigma\Delta t\left(\sqrt{2}B^n_\alpha\delta_{\beta0} - u^{n+1}_{\alpha\beta}\right). \tag{5.10}$$

Let us now set $B^{n+1}_\alpha = B^{n+1}$ and $u^{n+1}_{\alpha\beta} = u^{n+1}\delta_{\beta0}$. The scalar values B^{n+1} and u^{n+1} are chosen such that $B^{n+1} = \frac{1}{\sqrt{2}}u^{n+1} + \gamma$, where

$$0 < \gamma < \frac{2\sqrt{2}\sigma\Delta t}{2 + 3\sigma\Delta t + 4\sigma^2\left(\Delta t\right)^2 + 4\sigma^3\left(\Delta t\right)^3}u^{n+1}.$$

It can be verified that the chosen values for B^{n+1}_α and $u^{n+1}_{\alpha\beta}$ are retrieved after a single

step of the scheme (5.8) when using the initial conditions

$$B_\alpha^n = B^{n+1} + 2\sigma\Delta t\gamma = \frac{1}{\sqrt{2}}u^{n+1} + \gamma\left(1 + 2\sigma\Delta t\right), \tag{5.11a}$$

$$u_{\alpha\beta}^n = \left(u^{n+1} + \sigma\Delta t(u^{n+1} - \sqrt{2}B_\alpha^n)\right)\delta_{\beta 0} = \left(u^{n+1} - \sqrt{2}\sigma\Delta t\gamma\left(1 + 2\sigma\Delta t\right)\right)\delta_{\beta 0}. \tag{5.11b}$$

Inserting the initial values (5.11) into (5.10), we directly obtain $u_{\alpha\beta}^{n+1} = u^{n+1}\delta_{\beta 0}$. Similarly, by inserting (5.11) into (5.8d) we obtain $B_\alpha^{n+1} = B^{n+1}$. Then we square both of the initial terms (5.11). This leads to

$$\begin{aligned}
(B_\alpha^n)^2 &= \left(B^{n+1}\right)^2 + 4\sigma\Delta t\gamma B^{n+1} + 4\sigma^2\left(\Delta t\right)^2\gamma^2 \\
&= \left(B^{n+1}\right)^2 + 4\sigma\Delta t\gamma\left(\frac{1}{\sqrt{2}}u^{n+1} + \gamma\right) + 4\sigma^2\left(\Delta t\right)^2\gamma^2, \\
\left(u_{\alpha\beta}^n\right)^2 &= \left(\left(u^{n+1}\right)^2 - 2\sqrt{2}\sigma\Delta t\gamma u^{n+1}\left(1 + 2\sigma\Delta t\right) + 2\sigma^2\left(\Delta t\right)^2\gamma^2\left(1 + 2\sigma\Delta t\right)^2\right)\delta_{\beta 0}.
\end{aligned}$$

Summing the first equation over α, the second equation over α and β, adding the two terms, and multiplying with $\frac{1}{2}$, together with Definition 5.4 yields

$$E^{n+1} = E^n + \frac{1}{2}\sum_{\alpha=1}^{N_x}\left(2\sigma\Delta t\gamma\left(2\sqrt{2}\sigma\Delta t u^{n+1} - \gamma\left(2 + 2\sigma\Delta t + \sigma\Delta t\left(1 + 2\sigma\Delta t\right)^2\right)\right)\right).$$

Note that $E^{n+1} > E^n$ if

$$2\sqrt{2}\sigma\Delta t u^{n+1} - \gamma\left(2 + 2\sigma\Delta t + \sigma\Delta t\left(1 + 2\sigma\Delta t\right)^2\right) > 0.$$

Rearranging the inequality gives

$$\gamma < \frac{2\sqrt{2}\sigma\Delta t}{2 + 3\sigma\Delta t + 4\sigma^2\left(\Delta t\right)^2 + 4\sigma^3\left(\Delta t\right)^3}u^{n+1}.$$

This is exactly the domain γ is chosen from. Hence, we have $E^{n+1} > E^n$ and the unphysical behavior of the scheme (5.8) is proven. □

5.4.2 Energy stable space-time discretization

The naive scheme presented in (5.8) can increase the total energy in one time step. The main goal of this section is to construct a novel energy stable time integration scheme for which the corresponding analysis leads to a classic hyperbolic CFL condition which enables operating up to a time step size of $\Delta t = C_{\mathrm{CFL}} \cdot \Delta x$.

Energy stable DLRA scheme for Su-Olson. For constructing this energy stable scheme, the original equations are split in two parts, followed by a basis augmentation and a correction step.

In detail, we first solve

$$K^*_{jp} = K^n_{jp} - \Delta t \sum_{i=1}^{N_x} D^x_{ji} \sum_{\eta=1}^{r} K^n_{i\eta} \sum_{k,\ell=0}^{N_\mu-1} V^n_{\ell\eta} A_{k\ell} V^n_{kp} \tag{5.12a}$$
$$+ \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} D^{xx}_{ji} \sum_{\eta=1}^{r} K^n_{i\eta} \sum_{k,\ell=0}^{N_\mu-1} V^n_{\ell\eta} |A|_{k\ell} V^n_{kp},$$
$$L^*_{kp} = L^n_{kp} - \Delta t \sum_{\ell=0}^{N_\mu-1} A_{k\ell} \sum_{m=1}^{r} L^n_{\ell m} \sum_{i,j=1}^{N_x} X^n_{im} D^x_{ji} X^n_{jp} \tag{5.12b}$$
$$+ \Delta t \frac{\Delta x}{2} \sum_{\ell=0}^{N_\mu-1} |A|_{k\ell} \sum_{m=1}^{r} L^n_{\ell m} \sum_{i,j=1}^{N_x} X^n_{im} D^{xx}_{ji} X^n_{jp}.$$

The updated bases $\widehat{\mathbf{X}}^*$ of rank $2r$ and $\widehat{\mathbf{V}}^*$ of rank $2r$ are obtained from a QR-decomposition of the augmented quantities $\widehat{\mathbf{X}}^* = \mathrm{qr}([\mathbf{K}^*, \mathbf{X}^n])$ and $\widehat{\mathbf{V}}^* = \mathrm{qr}([\mathbf{L}^*, \mathbf{V}^n])$. Using the notation $\widetilde{S}^n_{m\eta} = \sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu-1} \sum_{q,p=1}^{r} \widehat{X}^*_{jm} X^n_{jq} S^n_{qp} V^n_{kp} \widehat{V}^*_{k\eta}$, we solve the S-step equation

$$\widehat{S}^*_{qp} = \widetilde{S}^n_{qp} - \Delta t \sum_{i,j=1}^{N_x} \widehat{X}^*_{jq} D^x_{ji} \sum_{m,\eta=1}^{2r} \widehat{X}^*_{im} \widetilde{S}^n_{m\eta} \sum_{k,\ell=0}^{N_\mu-1} \widehat{V}^*_{\ell\eta} A_{k\ell} \widehat{V}^*_{kp} \tag{5.12c}$$
$$+ \Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} \widehat{X}^*_{jq} D^{xx}_{ji} \sum_{m,\eta=1}^{2r} \widehat{X}^*_{im} \widetilde{S}^n_{m\eta} \sum_{k,\ell=0}^{N_\mu-1} \widehat{V}^*_{\ell\eta} |A|_{k\ell} \widehat{V}^*_{kp}.$$

Second, we solve the coupled equations for the internal energy $\mathbf{B} \in \mathbb{R}^{N_x}$ and the zeroth order moment $\widetilde{\mathbf{u}}^{n+1}_0 = (\widetilde{u}^{n+1}_{j0})_j \in \mathbb{R}^{N_x}$ according to

$$\widetilde{u}^{n+1}_{j0} = \sum_{m,\eta=1}^{r} X^n_{jm} S^n_{m\eta} V^n_{0\eta} - \Delta t \sum_{i=1}^{N_x} D^x_{ji} \sum_{m,\eta=1}^{2r} \widehat{X}^*_{im} \widetilde{S}^n_{m\eta} \sum_{\ell=0}^{N_\mu-1} \widehat{V}^*_{\ell\eta} A_{0\ell} \tag{5.12d}$$
$$+ \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} D^{xx}_{ji} \sum_{m,\eta=1}^{2r} \widehat{X}^*_{im} \widetilde{S}^n_{m\eta} \sum_{\ell=0}^{N_\mu-1} \widehat{V}^*_{\ell\eta} |A|_{0\ell} + \sigma \Delta t \left(\sqrt{2} B^{n+1}_j - \widetilde{u}^{n+1}_{j0} \right),$$
$$B^{n+1}_j = B^n_j + \sigma \Delta t \left(\sqrt{2} \widetilde{u}^{n+1}_{j0} - 2 B^{n+1}_j \right). \tag{5.12e}$$

Following [KEC23], we perform the opacity update only on $\mathbf{L} = \widehat{\mathbf{V}}^* \widehat{\mathbf{S}}^*$, i.e. we compute

$$L^{*,\mathrm{abs}}_{kp} = \frac{1}{1 + \sigma \Delta t} L_{kp} \qquad \text{for } \; p \neq 0. \tag{5.12f}$$

We perform a QR-decomposition $\widehat{\mathbf{V}}^{*,\mathrm{abs}} \widehat{\mathbf{S}}^{*,\mathrm{abs},\top} = \mathrm{qr}\left(\mathbf{L}^{*,\mathrm{abs}}\right)$ to retrieve the factorized basis $\widehat{\mathbf{V}}^{*,\mathrm{abs}}$ and the coefficients contained in the matrix $\widehat{\mathbf{S}}^{*,\mathrm{abs}}$. In the next step, we augment the basis matrices according to

$$\widehat{\mathbf{X}}^{n+1} = \mathrm{qr}\left(\left[\widetilde{\mathbf{u}}^{n+1}_0, \widehat{\mathbf{X}}^* \right] \right) \qquad \text{and} \qquad \widehat{\mathbf{V}}^{n+1} = \mathrm{qr}\left(\left[\mathbf{e}_1, \widehat{\mathbf{V}}^{*,\mathrm{abs}} \right] \right), \tag{5.12g}$$

where $\mathbf{e}_1$ denotes the first unit vector in $\mathbb{R}^{N_\mu}$. Third, the coefficient matrix is updated to $\widehat{\mathbf{S}}^{n+1} \in \mathbb{R}^{(2r+1)\times(2r+1)}$ through

$$\widehat{\mathbf{S}}^{n+1} = \widehat{\mathbf{X}}^{n+1,\top}\widehat{\mathbf{X}}^{*}\widehat{\mathbf{S}}^{*,\mathrm{abs}}\widehat{\mathbf{V}}^{*,\mathrm{abs},\top}\left(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^\top\right)\widehat{\mathbf{V}}^{n+1} + \widehat{\mathbf{X}}^{n+1,\top}\widetilde{\mathbf{u}}_0^{n+1}\mathbf{e}_1^\top\widehat{\mathbf{V}}^{n+1}. \tag{5.12h}$$

The updated solution $\mathbf{u} \in \mathbb{R}^{N_x\times N_\mu}$ is obtained as $\mathbf{u}^{n+1} = \widehat{\mathbf{X}}^{n+1}\widehat{\mathbf{S}}^{n+1}\widehat{\mathbf{V}}^{n+1}$. Lastly, we truncate the augmented quantities $\widehat{\mathbf{X}}^{n+1}, \widehat{\mathbf{S}}^{n+1}$ and $\widehat{\mathbf{V}}^{n+1}$ from rank $2r+1$ to a new rank r_{n+1} by using a suitable truncation strategy such as proposed in Section 4.4.2. This finally leads to the low-rank factors $\mathbf{X}^{n+1}, \mathbf{S}^{n+1}$ and $\mathbf{V}^{n+1}$. To provide an overview of the scheme, its main steps are visualized in Algorithm 1.

Proof of energy stability of the proposed low-rank scheme. For showing energy stability of the DLRA scheme given in (5.12), we first provide the following auxiliary results.

Lemma 5.6 (Young's inequality, [CR16])**.** *Let $1 < p < q < \infty$ be such that $\frac{1}{p} + \frac{1}{q} = 1$ and let $a, b \in \mathbb{R}^+$. Then it holds*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

with equality if and only if $a^p = b^q$.

Proof. See for instance [CR16]. □

A practically useful result within the framework of Fourier analysis is obtained following the ideas presented in [KEC23].

Lemma 5.7. *Let us define the matrix $\mathbf{E} \in \mathbb{C}^{N_x\times N_x}$ with entries*

$$E_{j\alpha} = \sqrt{\frac{\Delta x}{|\Omega_x|}}\exp\left(2\pi i\alpha x_j\right), \qquad \text{for } \; j,\alpha = 1,...,N_x,$$

where $|\Omega_x|$ denotes the length of the domain Ω_x. Then, $\mathbf{E}$ is a unitary matrix, i.e. $\mathbf{E}\mathbf{E}^H = \mathbf{E}^H\mathbf{E} = \mathbf{I}$, where the superscript H denotes the complex transpose and $\mathbf{I} \in \mathbb{R}^{N_x\times N_x}$ represents the identity matrix. In addition, it diagonalizes the stencil matrices

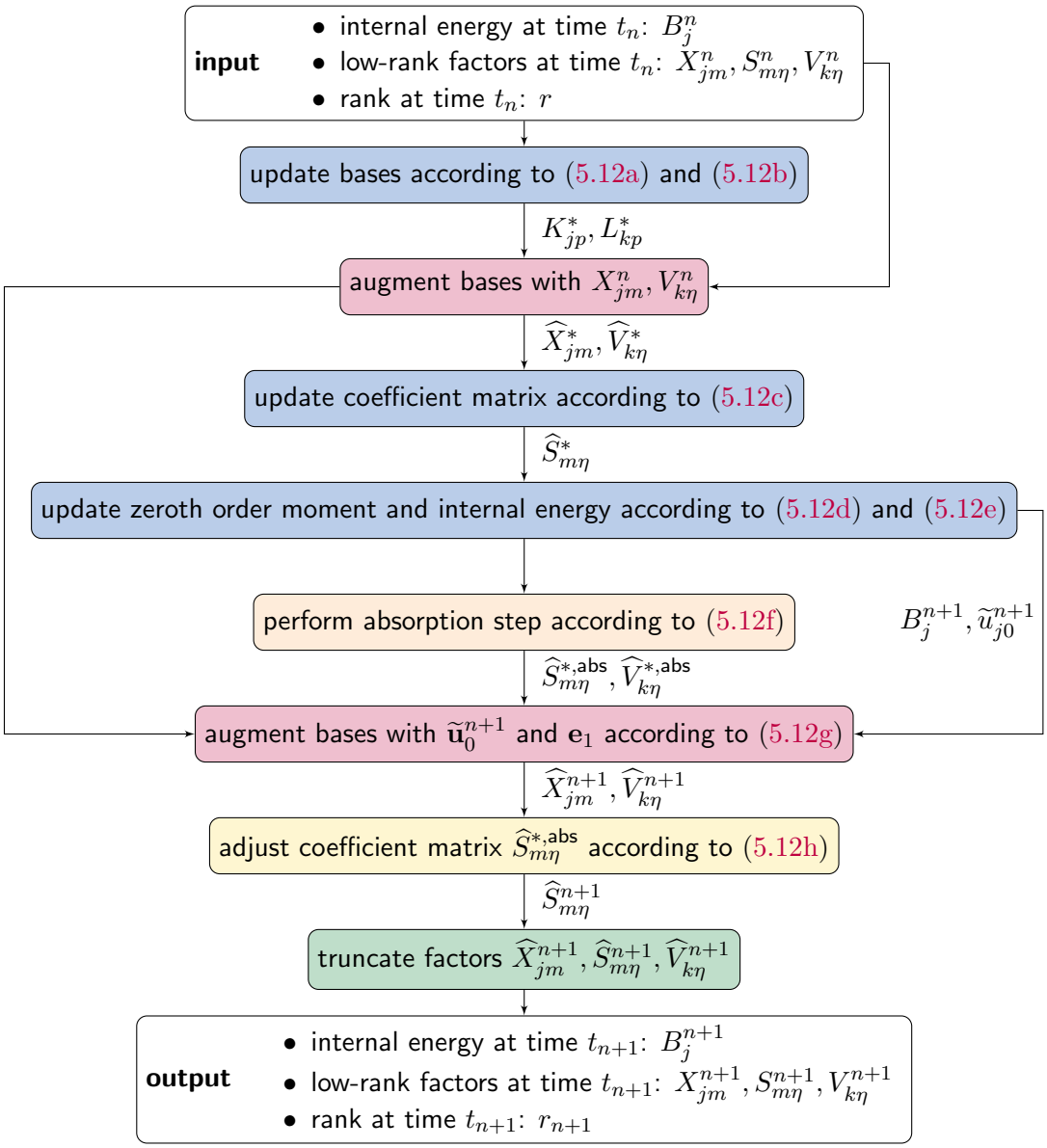
$$\mathbf{D}^\gamma\mathbf{E} = \mathbf{E}\mathbf{\Lambda}^\gamma \qquad \text{with } \; \gamma \in \{x, xx, +\},$$

and $\mathbf{\Lambda}^\gamma \in \mathbb{C}^{N_x\times N_x}$ are the diagonal matrices with entries

$$\begin{aligned}
\lambda^x_{\alpha\alpha} &= \frac{1}{2\Delta x}\left(e^{2\pi i\alpha\Delta x} - e^{-2\pi i\alpha\Delta x}\right) = \frac{i}{\Delta x}\sin\left(\nu_\alpha\right),\\
\lambda^{xx}_{\alpha\alpha} &= \frac{1}{(\Delta x)^2}\left(e^{2\pi i\alpha\Delta x} - 2 + e^{-2\pi i\alpha\Delta x}\right) = \frac{2}{(\Delta x)^2}\left(\cos(\nu_\alpha) - 1\right),\\
\lambda^{+}_{\alpha\alpha} &= \frac{1}{\Delta x}\left(e^{2\pi i\alpha\Delta x} - 1\right) = \frac{1}{\Delta x}\left(\cos(\nu_\alpha) + i\sin(\nu_\alpha) - 1\right),
\end{aligned}$$

where $\nu_\alpha := 2\pi\alpha\Delta x$.

Algorithm 1 Flowchart of the energy stable and mass conservative DLRA scheme (5.12).



Proof. The assertions follow directly by inserting the definitions of the matrix $\mathbf{E}$, the spatial stencil matrices $\mathbf{D}^\gamma$ and the diagonal matrices $\mathbf{\Lambda}^\gamma$ for $\gamma \in \{x, xx, +\}$, assuming periodic boundary conditions. □

Also the following lemma is indispensable for the proof of energy stability.

Lemma 5.8. *Under the time step restriction $\Delta t \leq \Delta x$ it holds*

$$\Delta t \left\| \mathbf{D}^x \mathbf{u}^{n+1} \mathbf{A} - \frac{\Delta x}{2} \mathbf{D}^{xx} \mathbf{u}^{n+1} |\mathbf{A}| \right\|_F^2 - \Delta x \left\| \mathbf{D}^+ \mathbf{u}^{n+1} |\mathbf{A}|^{1/2} \right\|_F^2 \leq 0. \tag{5.13}$$

Proof. We employ a Fourier analysis similar to [KEC23] and use Lemma 5.7 introducing the matrices $\mathbf{E}$ and $\mathbf{\Lambda}^\gamma$. Moreover, we recall that the matrix $\mathbf{A}$ can be decomposed as $\mathbf{A} = \mathbf{Q}\mathbf{M}\mathbf{Q}^\top$ with $\mathbf{Q}$ being orthogonal and $\mathbf{M} = \operatorname{diag}(\sigma_0, ..., \sigma_{N_\mu - 1})$. Let us denote $\widehat{\mathbf{u}}^{n+1} = \left(\widehat{u}_{\alpha m}^{n+1}\right) \in \mathbb{C}^{N_x \times N_\mu}$ with entries $\widehat{u}_{\alpha m}^{n+1} = \sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu - 1} E_{\alpha j} u_{jk}^{n+1} Q_{km}$. By applying Parseval's identity as stated in Proposition 3.14, we obtain

$$\begin{aligned}
&\Delta t \left\| \mathbf{D}^x \mathbf{u}^{n+1} \mathbf{A} - \frac{\Delta x}{2} \mathbf{D}^{xx} \mathbf{u}^{n+1} |\mathbf{A}| \right\|_F^2 - \Delta x \left\| \mathbf{D}^+ \mathbf{u}^{n+1} |\mathbf{A}|^{1/2} \right\|_F^2 \\
&= \Delta t \left\| \mathbf{E}\mathbf{\Lambda}^x \widehat{\mathbf{u}}^{n+1} \mathbf{M}\mathbf{Q}^\top - \frac{\Delta x}{2} \mathbf{E}\mathbf{\Lambda}^{xx} \widehat{\mathbf{u}}^{n+1} |\mathbf{M}| \, \mathbf{Q}^\top \right\|_F^2 - \Delta x \left\| \mathbf{E}\mathbf{\Lambda}^+ \widehat{\mathbf{u}}^{n+1} |\mathbf{M}|^{1/2} \, \mathbf{Q}^\top \right\|_F^2 \\
&= \Delta t \left\| \mathbf{\Lambda}^x \widehat{\mathbf{u}}^{n+1} \mathbf{M} - \frac{\Delta x}{2} \mathbf{\Lambda}^{xx} \widehat{\mathbf{u}}^{n+1} |\mathbf{M}| \right\|_F^2 - \Delta x \left\| \mathbf{\Lambda}^+ \widehat{\mathbf{u}}^{n+1} |\mathbf{M}|^{1/2} \right\|_F^2 \\
&= 2 \sum_{\alpha=1}^{N_x} \sum_{m=0}^{N_\mu - 1} \left(\Delta t \frac{|\sigma_m|^2}{(\Delta x)^2} |1 - \cos(\nu_\alpha)| - \frac{|\sigma_m|}{\Delta x} |1 - \cos(\nu_\alpha)| \right) \left| \widehat{u}_{\alpha m}^{n+1} \right|^2 .
\end{aligned}$$

A sufficient condition to ensure negativity is that for each index m it must hold

$$\Delta t \frac{|\sigma_m|^2}{(\Delta x)^2} |1 - \cos(\nu_\alpha)| \leq \frac{|\sigma_m|}{\Delta x} |1 - \cos(\nu_\alpha)| .$$

Hence, for $\Delta t \leq \frac{\Delta x}{|\sigma_m|}$, equation (5.13) holds. Since $|\sigma_m| \leq 1$, we have proven the lemma. □

We can now show energy stability of the proposed scheme.

Theorem 5.9 (Energy stability of the proposed DLRA scheme)**.** *Under the time step restriction $\Delta t \leq \Delta x$ the fully discrete DLRA scheme presented in (5.12) is energy stable, i.e. it holds*

$$\frac{1}{2} \left\| \mathbf{B}^{n+1} \right\|_E^2 + \frac{1}{2} \left\| \mathbf{X}^{n+1} \mathbf{S}^{n+1} \mathbf{V}^{n+1,\top} \right\|_F^2 \leq \frac{1}{2} \left\| \mathbf{B}^n \right\|_E^2 + \frac{1}{2} \left\| \mathbf{X}^n \mathbf{S}^n \mathbf{V}^{n,\top} \right\|_F^2 . \tag{5.14}$$

Proof. First, we multiply equation (5.12e) with B_j^{n+1} and obtain

$$\left(B_j^{n+1} \right)^2 = B_j^n B_j^{n+1} + \sigma \Delta t \left(\sqrt{2} \widetilde{u}_{j0}^{n+1} B_j^{n+1} - 2 \left(B_j^{n+1} \right)^2 \right) .$$

Let us note that

$$B_j^n B_j^{n+1} = \frac{1}{2}\left(B_j^{n+1}\right)^2 + \frac{1}{2}\left(B_j^n\right)^2 - \frac{1}{2}\left(B_j^{n+1} - B_j^n\right)^2. \tag{5.15}$$

Inserting this relation and summing over j leads to

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1}\right)^2 = \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^n\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1} - B_j^n\right)^2 \\ + \sigma\Delta t\sum_{j=1}^{N_x}\left(\sqrt{2}\widetilde{u}_{j0}^{n+1}B_j^{n+1} - 2\left(B_j^{n+1}\right)^2\right).\end{aligned} \tag{5.16}$$

To obtain a similar expression for $\left(u_{jk}^{n+1}\right)^2$, we multiply (5.12c) with $\widehat{X}_{\alpha q}^*\widehat{V}_{\beta p}^*$ and sum over q and p. For simplicity of notation, let us introduce $u_{\alpha\beta}^* := \sum_{q,p=1}^{2r}\widehat{X}_{\alpha q}^*\widehat{S}_{qp}^*\widehat{V}_{\beta p}^*$ and $u_{\alpha\beta}^n := \sum_{q,p=1}^{2r}\widehat{X}_{\alpha q}^*\widetilde{S}_{qp}^n\widehat{V}_{\beta p}^*$ as well as the projection operators $P_{\alpha j}^{X^*} = \sum_{q=1}^{2r}\widehat{X}_{\alpha q}^*\widehat{X}_{jq}^*$ and $P_{k\beta}^{V^*} = \sum_{p=1}^{2r}\widehat{V}_{kp}^*\widehat{V}_{\beta p}^*$. Then we obtain

$$\begin{aligned}u_{\alpha\beta}^* = u_{\alpha\beta}^n - \Delta t\sum_{i,j=1}^{N_x}P_{\alpha j}^{X^*}D_{ji}^x\sum_{k,\ell=0}^{N_\mu-1}u_{i\ell}^n A_{k\ell}P_{k\beta}^{V^*} \\ + \Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}P_{\alpha j}^{X^*}D_{ji}^{xx}\sum_{k,\ell=0}^{N_\mu-1}u_{i\ell}^n\left|A\right|_{k\ell}P_{k\beta}^{V^*}.\end{aligned} \tag{5.17}$$

Note that with $u_{\alpha\beta}^{n+1} = \sum_{q,p=1}^{2r}\widehat{X}_{\alpha q}^{n+1}\widehat{S}_{qp}^{n+1}\widehat{V}_{\beta p}^{n+1}$ and by construction it holds

$$u_{\alpha\beta}^{n+1} = \frac{u_{\alpha\beta}^\star(1-\delta_{\beta 0})}{1+\sigma\Delta t} + \widetilde{u}_{\alpha 0}^{n+1}\delta_{\beta 0}.$$

Hence, inserting the schemes for $u_{\alpha\beta}^*$ and $\widetilde{u}_{\alpha 0}^{n+1}$, i.e. equations (5.17) and (5.12d), leads to

$$\begin{aligned}(1+\sigma\Delta t)\,u_{\alpha\beta}^{n+1} = \Bigg(u_{\alpha\beta}^n - \Delta t\sum_{i,j=1}^{N_x}P_{\alpha j}^{X^*}D_{ji}^x\sum_{k,\ell=0}^{N_\mu-1}u_{i\ell}^n A_{k\ell}P_{k\beta}^{V^*} \\ + \Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}P_{\alpha j}^{X^*}D_{ji}^{xx}\sum_{k,\ell=0}^{N_\mu-1}u_{i\ell}^n\left|A\right|_{k\ell}P_{k\beta}^{V^*}\Bigg)(1-\delta_{\beta 0}) \\ + \Bigg(u_{\alpha 0}^n - \Delta t\sum_{i=1}^{N_x}D_{\alpha i}^x\sum_{\ell=0}^{N_\mu-1}u_{i\ell}^n A_{0\ell} \\ + \Delta t\frac{\Delta x}{2}\sum_{i=1}^{N_x}D_{\alpha i}^{xx}\sum_{\ell=0}^{N_\mu-1}u_{i\ell}^n\left|A\right|_{0\ell} + \sqrt{2}\sigma\Delta t B_j^{n+1}\Bigg)\delta_{\beta 0}.\end{aligned}$$

In the next step, we multiply with $u_{\alpha\beta}^{n+1}$ and sum over α and β. Note that it holds

$$\sum_{\alpha=1}^{N_x} P_{\alpha j}^{X^*} u_{\alpha\beta}^{n+1} = u_{j\beta}(t) \qquad \text{and} \qquad \sum_{\beta=0}^{N_\mu-1} P_{k\beta}^{V^*} u_{j\beta}^{n+1} = u_{jk}^{n+1}.$$

To be consistent in notation, we change the summation indices in the corresponding terms from α to j and from β to k. Let us note that

$$\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1} u_{jk}^n u_{jk}^{n+1} = \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1} \left(\frac{1}{2}\left(u_{jk}^{n+1}\right)^2 + \frac{1}{2}\left(u_{jk}^n\right)^2 - \frac{1}{2}\left(u_{jk}^{n+1} - u_{jk}^n\right)^2\right). \tag{5.18}$$

This results in

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\right)^2 = \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^n\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}-u_{jk}^n\right)^2 \\ - \Delta t\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x u_{i\ell}^n A_{k\ell} + \Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^n |A|_{k\ell} \\ + \sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1} u_{jk}^{n+1}\left(\sqrt{2}B_j^{n+1}\delta_{k0} - u_{jk}^{n+1}\right).\end{aligned}$$

Let us now add the zero term $\Delta t\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x u_{i\ell}^{n+1} A_{k\ell}$ and add and subtract the term $\Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^{n+1} |A|_{k\ell}$. This yields

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\right)^2 = \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^n\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}-u_{jk}^n\right)^2 \\ - \Delta t\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x\left(u_{i\ell}^n - u_{i\ell}^{n+1}\right) A_{k\ell} \qquad \text{(I)} \\ + \Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx}\left(u_{i\ell}^n - u_{i\ell}^{n+1}\right)|A|_{k\ell} \qquad \text{(II)} \\ + \Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^{n+1}|A|_{k\ell} \qquad \text{(III)} \\ + \sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1} u_{jk}^{n+1}\left(\sqrt{2}B_j^{n+1}\delta_{k0} - u_{jk}^{n+1}\right).\end{aligned}$$

We proceed by analyzing the terms (I), (II), and (III) separately. Let us start with (I) and (II) and apply Young's inequality given in Lemma 5.6. For the sum (I) + (II) this

results in

$$
\begin{aligned}
&- \Delta t \sum_{i,j=1}^{N_x} \sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{x} \left(u_{i\ell}^{n} - u_{i\ell}^{n+1}\right) A_{k\ell} + \Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} \sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} \left(u_{i\ell}^{n} - u_{i\ell}^{n+1}\right) |A|_{k\ell} \\
&= - \Delta t \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} \left(u_{i\ell}^{n} - u_{i\ell}^{n+1}\right) \left(\sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu-1} \left(D_{ji}^{x} u_{jk}^{n+1} A_{k\ell} - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell} \right) \right) \\
&\leq \frac{1}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} \left(u_{i\ell}^{n} - u_{i\ell}^{n+1}\right)^2 \\
&\quad + \frac{(\Delta t)^2}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} \left(\sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu-1} \left(D_{ji}^{x} u_{jk}^{n+1} A_{k\ell} - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell} \right) \right)^2 .
\end{aligned}
$$

For (III) we exploit the properties of the spatial stencil matrices given in Lemma 3.1. This leads to the equality

$$
\Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} \sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^{n+1} |A|_{k\ell} = -\Delta t \frac{\Delta x}{2} \sum_{j=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} \left(\sum_{i=1}^{N_x} \sum_{k=0}^{N_\mu-1} D_{ji}^{+} u_{ik}^{n+1} |A|_{k\ell}^{1/2} \right)^2 .
$$

Hence, inserting these relations, yields

$$
\begin{aligned}
\frac{1}{2} \sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu-1} \left(u_{jk}^{n+1}\right)^2 \leq \frac{1}{2} \sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu-1} \left(u_{jk}^{n}\right)^2 & \\
+ \frac{(\Delta t)^2}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} \left(\sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu-1} \left(D_{ji}^{x} u_{jk}^{n+1} A_{k\ell} \right.\right. & \\
\left.\left. - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell} \right) \right)^2 & \qquad (5.19) \\
- \Delta t \frac{\Delta x}{2} \sum_{j=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} \left(\sum_{i=1}^{N_x} \sum_{k=0}^{N_\mu-1} D_{ji}^{+} u_{ik}^{n+1} |A|_{k\ell}^{1/2} \right)^2 & \\
+ \sigma \Delta t \sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu-1} u_{jk}^{n+1} \left(\sqrt{2} B_j^{n+1} \delta_{k0} - u_{jk}^{n+1} \right) . &
\end{aligned}
$$

As for the continuous case, we add equations (5.19) and (5.16) to obtain a time update

for the total energy introduced in Definition 5.4. This establishes the inequality

$$
\begin{aligned}
E^{n+1} \leq E^n &+ \frac{(\Delta t)^2}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \left(\sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu - 1} \left(D_{ji}^x u_{jk}^{n+1} A_{k\ell} - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell} \right) \right)^2 \\
&- \Delta t \frac{\Delta x}{2} \sum_{j=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \left(\sum_{i=1}^{N_x} \sum_{k=0}^{N_\mu - 1} D_{ji}^+ u_{ik}^{n+1} |A|_{k\ell}^{1/2} \right)^2 \\
&+ \sigma \Delta t \sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu - 1} \left(\sqrt{2} u_{j0}^{n+1} B_j^{n+1} - \left(u_{jk}^{n+1} \right)^2 \right) \\
&- \frac{1}{2} \sum_{j=1}^{N_x} \left(B_j^{n+1} - B_j^n \right)^2 + \sigma \Delta t \sum_{j=1}^{N_x} \left(\sqrt{2} u_{j0}^{n+1} B_j^{n+1} - 2 \left(B_j^{n+1} \right)^2 \right).
\end{aligned}
$$

We estimate the opacity term and derive

$$
\begin{aligned}
E^{n+1} \leq E^n &+ \frac{(\Delta t)^2}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \left(\sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu - 1} \left(D_{ji}^x u_{jk}^{n+1} A_{k\ell} - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell} \right) \right)^2 \\
&- \Delta t \frac{\Delta x}{2} \sum_{j=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \left(\sum_{i=1}^{N_x} \sum_{k=0}^{N_\mu - 1} D_{ji}^+ u_{ik}^{n+1} |A|_{k\ell}^{1/2} \right)^2 \\
&- \sigma \Delta t \sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu - 1} \left(\sqrt{2} B_j^{n+1} - u_{jk}^{n+1} \right)^2 - \frac{1}{2} \sum_{j=1}^{N_x} \left(B_j^{n+1} - B_j^n \right)^2.
\end{aligned}
$$

With Lemma 5.8 we obtain

$$
\begin{aligned}
&\Delta t \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \left(\sum_{j=1}^{N_x} \sum_{k=0}^{N_\mu - 1} \left(D_{ji}^x u_{jk}^{n+1} A_{k\ell} - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell} \right) \right)^2 \\
&- \Delta x \sum_{j=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \left(\sum_{i=1}^{N_x} \sum_{k=0}^{N_\mu - 1} D_{ji}^+ u_{ik}^{n+1} |A|_{k\ell}^{1/2} \right)^2 \leq 0
\end{aligned}
$$

for $\Delta t \leq \Delta x$. Since the truncation step is designed to not alter the zeroth order moment, this means that $E^{n+1} \leq E^n$ and we can conclude that the full scheme is energy stable under the time step restriction $\Delta t \leq \Delta x$. □

5.5 Mass conservation

A drawback of the DLRA method using the integrators introduced in Section 4.2 is that physical invariants are not preserved. This problem can be overcome when implementing the rank-adaptive augmented BUG integrator introduced in [CKL22] together with

suitable basis augmentation steps and a conservative truncation strategy as described in Section 4.4.2. We first translate the macroscopic quantities given in Definition 5.1 to the fully discrete setting.

Definition 5.10 (Fully discrete macroscopic quantities)**.** The *mass* and the *momentum* of the fully discretized Su-Olson problem at time t_n are defined as

$$\rho_j^n := \sqrt{2}u_{j0}^n + B_j^n \qquad \text{and} \qquad \overline{u}_j^n := \sqrt{2}\sum_{\ell=0}^{N_\mu-1} u_{j\ell}^n A_{0\ell}.$$

Then we can show that besides being energy stable our DLRA scheme ensures local conservation of mass.

Theorem 5.11 (Mass conservation of the proposed DLRA scheme)**.** *The DLRA scheme (5.12) together with the conservative truncation strategy described in Section 4.4.2 is locally mass conservative, i.e. it fulfills the local conservation law*

$$\begin{aligned}&\frac{1}{\Delta t}\left(\sqrt{2}\Phi_j^{n+1} + B_j^{n+1} - \left(\sqrt{2}\Phi_j^n + B_j^n\right)\right) \\ &= -\sqrt{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^x u_{i\ell}^n A_{0\ell} + \sqrt{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{xx} u_{i\ell}^n |A|_{0\ell},\end{aligned} \tag{5.20}$$

where $\Phi_j^n := \sum_{m,\eta=1}^r X_{jm}^n S_{m\eta}^n V_{0\eta}^n$ *and* $\Phi_j^{n+1} := \sum_{m,\eta=1}^{r_{n+1}} X_{jm}^{n+1} S_{m\eta}^{n+1} V_{0\eta}^{n+1}$*. As done before, we denote* $u_{jk}^n = \sum_{m,\eta=1}^r X_{jm}^n S_{m\eta}^n V_{k\eta}^n$*. This is a discretization of the continuous local conservation law given in (5.2).*

Proof. The conservative truncation strategy is designed to not alter the zeroth order moment, i.e. it holds $\sum_{m,\eta=1}^{2r} \widehat{X}_{jm}^{n+1}\widehat{S}_{m\eta}^{n+1}\widehat{V}_{0\eta}^{n+1} = u_{j0}^{n+1}$. In addition, the basis augmentations performed in (5.12g) and the adjustment step stated in (5.12h) ensure that $\sum_{m,\eta=1}^{2r} \widehat{X}_{jm}^{n+1}\widehat{S}_{m\eta}^{n+1}\widehat{V}_{0\eta}^{n+1} = \sum_{m,\eta=1}^{r_{n+1}} X_{jm}^{n+1} S_{m\eta}^{n+1} V_{0\eta}^{n+1}$. Combining both, this leads to the equality

$$\Phi_j^{n+1} = \sum_{m,\eta=1}^{r_{n+1}} X_{jm}^{n+1} S_{m\eta}^{n+1} V_{0\eta}^{n+1} = \sum_{m,\eta=1}^{2r} \widehat{X}_{jm}^{n+1}\widehat{S}_{m\eta}^{n+1}\widehat{V}_{0\eta}^{n+1} = u_{j0}^{n+1}.$$

We insert this relation into the coupled equations (5.12d) and (5.12e) and multiply (5.12d) with $\sqrt{2}$. This yields

$$\begin{aligned}\sqrt{2}\Phi_j^{n+1} &= \sqrt{2}\Phi_j^n - \sqrt{2}\Delta t\sum_{i=1}^{N_x} D_{ji}^x \sum_{m,\eta=1}^{2r} \widehat{X}_{im}^* \widetilde{S}_{m\eta}^n \sum_{\ell=0}^{N_\mu-1} \widehat{V}_{\ell\eta}^* A_{0\ell} \\ &\quad + \sqrt{2}\Delta t\frac{\Delta x}{2}\sum_{i=1}^{N_x} D_{ji}^{xx} \sum_{m,\eta=1}^{2r} \widehat{X}_{im}^* \widetilde{S}_{m\eta}^n \sum_{\ell=0}^{N_\mu-1} \widehat{V}_{\ell\eta}^* |A|_{0\ell} + \sigma\Delta t\left(2B_j^{n+1} - \sqrt{2}\Phi_j^{n+1}\right),\end{aligned} \tag{5.21a}$$

$$B_j^{n+1} = B_j^n + \sigma\Delta t\left(\sqrt{2}\Phi_j^{n+1} - 2B_j^{n+1}\right). \tag{5.21b}$$

Due to the basis augmentations with $\mathbf{X}^n$ and $\mathbf{V}^n$ introduced by the rank-adaptive augmented BUG integrator it can be concluded that

$$\sum_{m,\eta=1}^{2r} \widehat{X}^*_{im} \widetilde{S}^n_{m\eta} \widehat{V}^*_{\ell\eta} = \sum_{m,\eta=1}^{r} X^n_{im} S^n_{m\eta} V^n_{\ell\eta} = u^n_{i\ell}.$$

We insert this relation into (5.21a), add equations (5.21a) and (5.21b), and rearrange the obtained expression. This leads to the local conservation law (5.20), ensuring local conservation of mass. □

Hence, equipped with a conservative truncation step, the energy stable DLRA algorithm presented in (5.12) locally conserves mass.

5.6 Numerical results

In this section, we provide numerical results to validate the energy stable and mass conservative DLRA scheme proposed in (5.12). Sections 5.6.1 and 5.6.2 are devoted to commonly considered 1D test examples in radiative transfer before in Section 5.6.3 an experiment in two spatial dimensions is presented.

5.6.1 1D plane source

We consider the thermal RTEs as described in (5.1) on the spatial domain $\Omega_x = [-10, 10]$ and the angular domain $\Omega_\mu = [-1, 1]$. As initial distribution we choose the cutoff Gaussian

$$u\left(t=0,x\right) = \max\left(10^{-4}, \frac{1}{\sqrt{2\pi\sigma_{\mathrm{IC}}^2}} \exp\left(-\frac{(x-1)^2}{2\sigma_{\mathrm{IC}}^2}\right)\right)$$

with constant deviation $\sigma_{\mathrm{IC}} = 0.03$. Particles are initially centered around $x = 1$ and move into all directions $\mu \in [-1, 1]$. The initial value for the internal energy is set to $B^0 = 1$ and the opacity to the constant value $\sigma = 1$. For the low-rank computations an initial rank of $r = 20$ is prescribed. Note that this setting is an extension of the so-called *plane source* problem, which is a common test case for the RTE [GBD+01, Gan08]. In the context of DLRA it has been studied for instance in [CKL22, KEC23, PMF20, PM21]. We compare the solution of the full coupled-implicit system

$$\begin{aligned} u^{n+1}_{jk} &= u^n_{jk} - \Delta t \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} D^x_{ji} u^n_{i\ell} A_{k\ell} + \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} D^{xx}_{ji} u^n_{i\ell} |A|_{k\ell} \\ &\quad + \sigma \Delta t \left(\sqrt{2} B^{n+1}_j \delta_{k0} - u^{n+1}_{jk}\right), \end{aligned} \tag{5.22a}$$

$$B^{n+1}_j = B^n_j + \sigma \Delta t \left(\sqrt{2} u^{n+1}_{j0} - 2 B^{n+1}_j\right), \tag{5.22b}$$

to the solution obtained from the energy stable and mass conservative DLRA scheme given in (5.12). We refer to (5.22) as the *full system*. The total mass at time t_n is defined as $m^n := \Delta x \sum_{j=1}^{N_x} \left(\sqrt{2} u_{j0}^n + B_j^n\right)$. As computational parameters we use $N_x = 1000$ cells in the spatial domain and $N_\mu = 500$ moments to represent the angular variable. The time step size is chosen to be $\Delta t = C_{\mathrm{CFL}} \cdot \Delta x$ with a CFL number of $C_{\mathrm{CFL}} = 0.99$.

In Figure 5.1 we present the computational results for the solution $f(x, \mu)$, the scalar flux $\Phi = \frac{1}{\sqrt{2}} \langle f \rangle_\mu$ and the dimensionless temperature $T = \sqrt[4]{B}$ at the end time $t_{\mathrm{end}} = 8$. Further, the evolution of the rank r in time and the evolution of the relative mass error $\frac{|m^0 - m^n|}{|m^0|}$ in time are shown. It is observable that the DLRA scheme captures well the behavior of the full system. For a chosen tolerance parameter of $\vartheta = 10^{-1} \|\mathbf{\Sigma}\|_F$ the rank increases up to $r = 23$ before it significantly decreases again. The relative mass error is of order $\mathcal{O}\left(10^{-13}\right)$. Hence, our proposed scheme is mass conservative up to machine precision. These results confirm our theoretical considerations.

5.6.2 1D external source

For the next test problem, a source term $Q(x)$ is added to the previously investigated equations, leading to

$$\begin{aligned} \partial_t f(t, x, \mu) + \mu \partial_x f(t, x, \mu) &= \sigma \left(B(t, x) - f(t, x, \mu)\right) + Q(x), \\ \partial_t B(t, x) &= \sigma \left\langle f(t, x, \mu) - B(t, x) \right\rangle_\mu. \end{aligned}$$

This source term generates radiation particles moving through and interacting with the background material. The interaction is driven by the opacity σ. In turn, particles heat up the material, leading to a traveling temperature front, also called a *Marshak wave* [Mar58]. Again this traveling heat wave can lead to the emission of new particles from the background material, generating a particle wave. In our example we use the source function $Q(x) = \chi_{[-0.5,0.5]}(x)/a$ with $a = \frac{4\sigma_{\mathrm{SB}}}{c}$ being the radiation and $\chi_{[-0.5,0.5]}(x)$ denoting the indicator function on $[-0.5, 0.5]$. The initial value for the internal energy is set to $B^0 = 50$. All other initial settings and computational parameters remain unchanged from the previous test example given in Section 5.6.1.

In Figure 5.2 we display the numerical results for the solution $f(x, \mu)$, the scalar flux $\Phi = \frac{1}{\sqrt{2}} \langle f \rangle_\mu$ and the temperature $T = \sqrt[4]{B}$ at a given time point $t_{\mathrm{end}} = 3.16$. We add the same source term to the full coupled-implicit system (5.22) as well as to the presented energy stable and mass conservative DLRA scheme given in (5.12) and compare the solution. Further, the evolution of the rank in time is presented for a chosen tolerance parameter of $\vartheta = 10^{-2} \|\mathbf{\Sigma}\|_F$. Again we observe that the proposed DLRA scheme approximates well the behavior of the full system. In addition, a very low rank is sufficient to obtain accurate results. Note that due to the additional source term there is no mass conservation in this example.

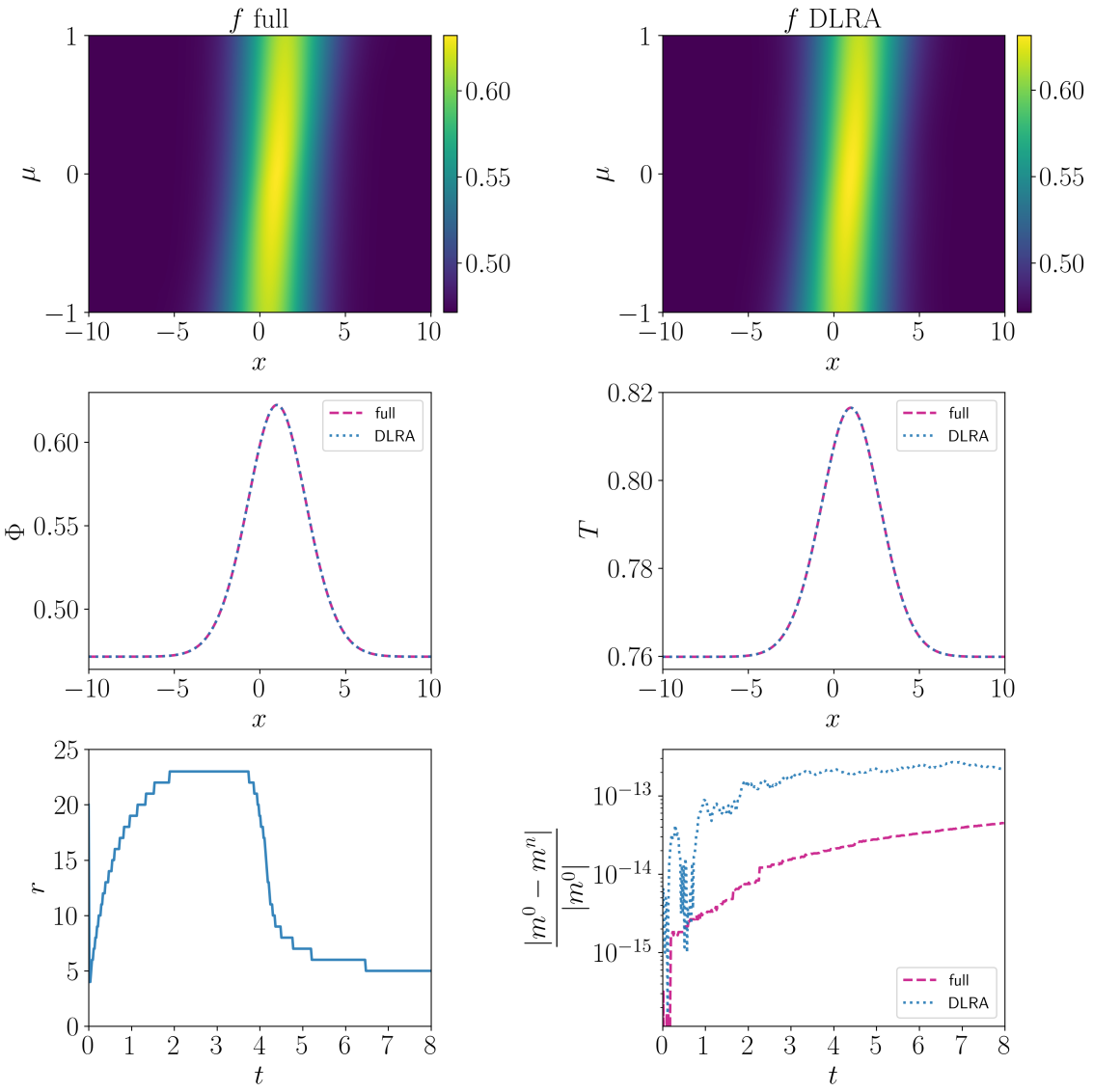


Figure 5.1: Top row: Numerical results for the solution $f(x, \mu)$ of the plane source problem at time $t_{\text{end}} = 8$ computed with the full coupled-implicit solver (left) and the DLRA scheme (right). **Middle row:** Scalar flux Φ (left) and temperature T (right) for both the full solver and the DLRA scheme. **Bottom row:** Evolution of the rank in time for the DLRA method (left) and evolution of the relative mass error in time compared for both methods (right).

5.6.3 2D beam

To approve computational benefits of the presented DLRA algorithm, we extend it to a 2D spatial and a 2D angular setting. The corresponding set of equations becomes

$$\begin{aligned} \partial_t f(t, \mathbf{x}, \mathbf{\Omega}) + \mathbf{\Omega} \cdot \nabla_{\mathbf{x}} f(t, \mathbf{x}, \mathbf{\Omega}) &= \sigma \left(B(t, \mathbf{x}) - f(t, \mathbf{x}, \mathbf{\Omega}) \right), \\ \partial_t B(t, \mathbf{x}) &= \sigma \left\langle f(t, \mathbf{x}, \mathbf{\Omega}) - B(t, \mathbf{x}) \right\rangle_{\mathbf{\Omega}}. \end{aligned}$$

For the numerical experiments let $\mathbf{x} = (x, y) \in [-1, 1] \times [-1, 1]$ and $\mathbf{\Omega} = (\Omega_x, \Omega_y, \Omega_z) \in \mathcal{S}^2$ be represented in 3D Cartesian coordinates as explained in Section 3.3.2. The initial

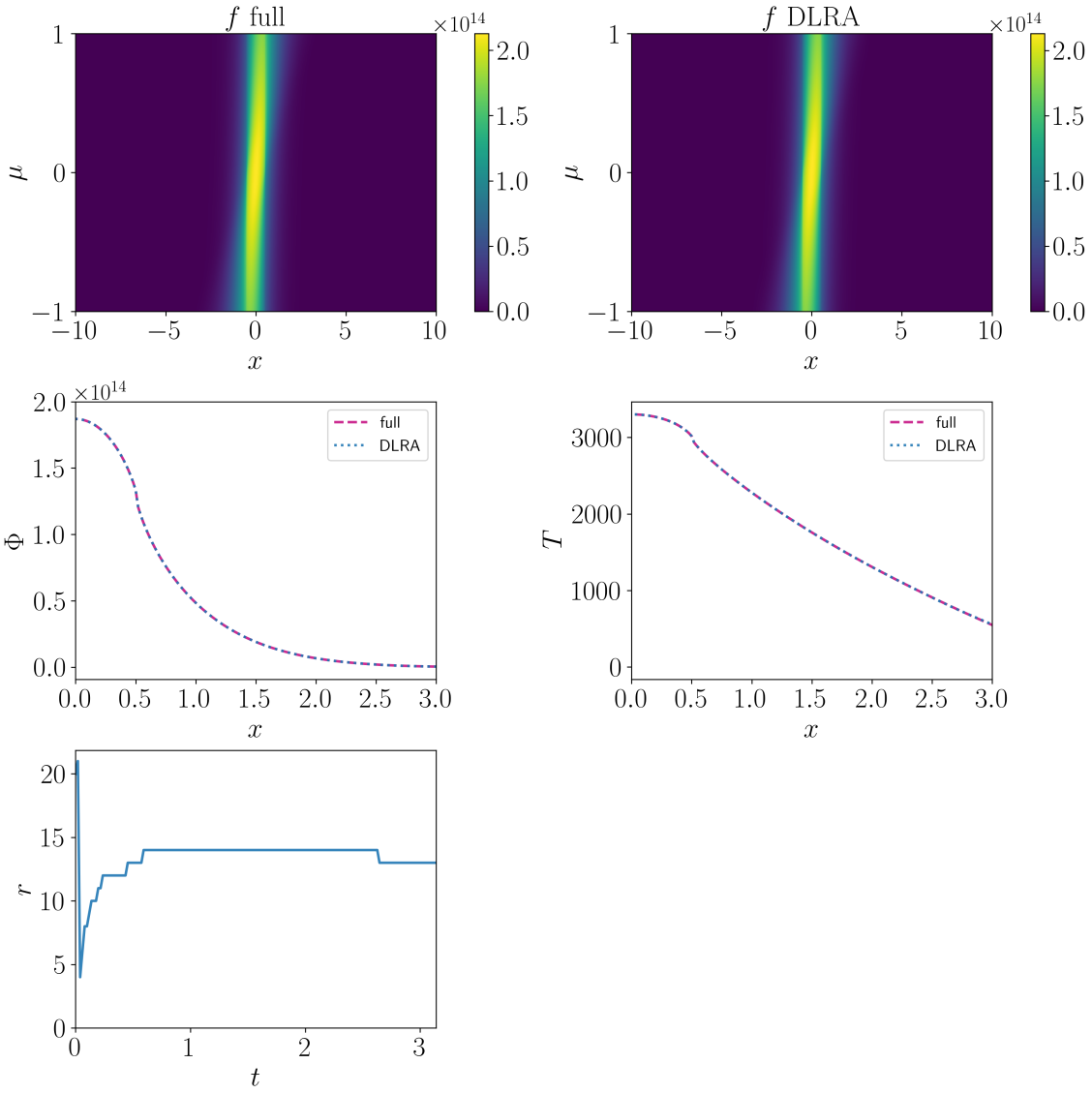


Figure 5.2: Top row: Numerical results for the solution $f(x, \mu)$ of the external source problem at time $t_{\text{end}} = 3.16$ computed with the full coupled-implicit solver (left) and the DLRA scheme (right). **Middle row:** Scalar flux Φ (left) and temperature T (right) for both the full solver and the DLRA scheme. **Bottom row:** Evolution of the rank in time for the DLRA method.

condition of the 2D beam is given by

$$f(t=0, \mathbf{x}, \mathbf{\Omega}) = 10^6 \cdot \frac{1}{2\pi\sigma_{\mathbf{x}}^2} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma_{\mathbf{x}}^2}\right) \cdot \frac{1}{2\pi\sigma_{\mathbf{\Omega}}^2} \exp\left(-\frac{(\Omega_x - \Omega^*)^2 + (\Omega_z - \Omega^*)^2}{2\sigma_{\mathbf{\Omega}}^2}\right),$$

where $\Omega^* = \frac{1}{\sqrt{2}}$ and $\sigma_{\mathbf{x}} = \sigma_{\mathbf{\Omega}} = 0.1$. The initial value for the internal energy is set to $B^0 = 1$ and the opacity to the constant value $\sigma = 0.5$. The low-rank computations are performed with an initial rank of $r = 100$. The total mass at any time t_n is defined as $m^n := \Delta x \Delta y \sum_{j=1}^{N_x \cdot N_y} \left(u_{j0}^n + B_j^n\right)$. We perform our computations on a spatial grid with $N_x = 500$ cells in x and $N_y = 500$ cells in y. For the 2D angular discretization, we use

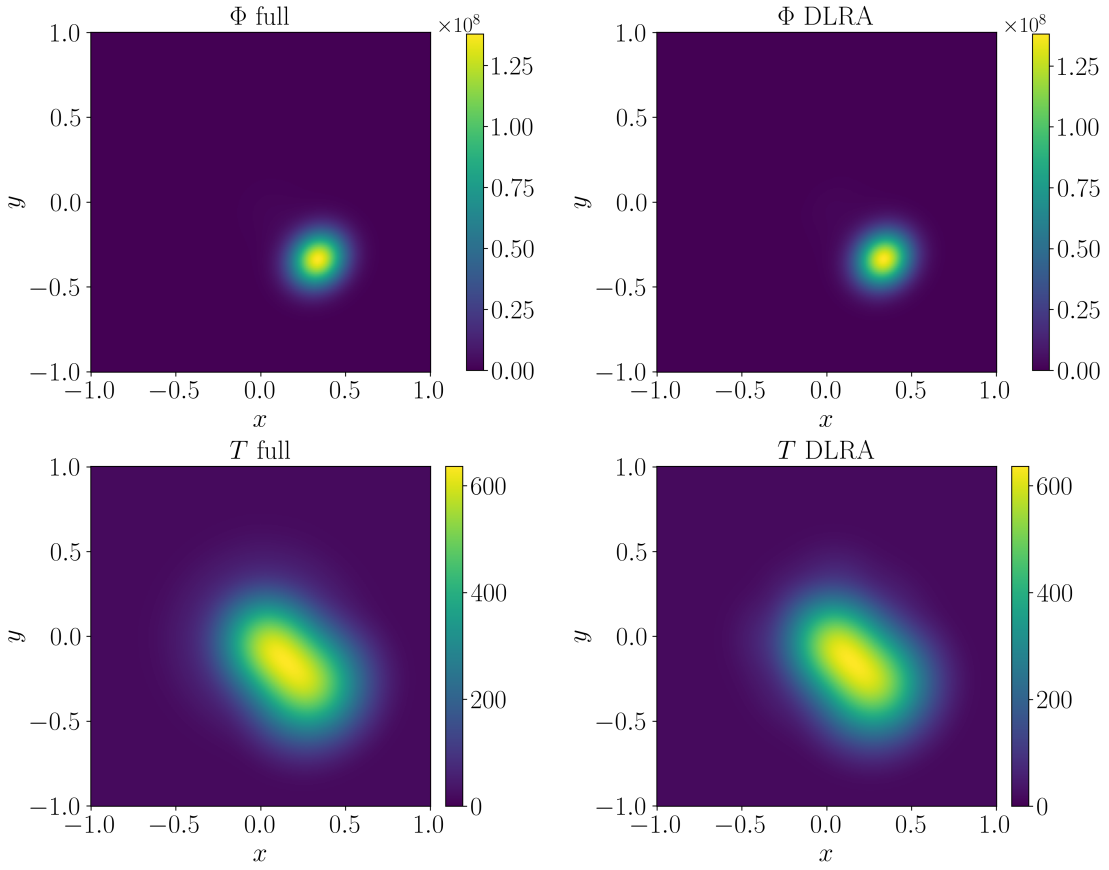


Figure 5.3: Numerical results for the scalar flux Φ and the temperature T for the 2D beam problem computed with the full coupled-implicit solver (left) and the DLRA scheme (right) at time $t_{\text{end}} = 0.5$.

the spherical harmonics method introduced in Section 3.3.2. We consider a polynomial degree of $N_{\boldsymbol{\Omega}} = 29$, corresponding to 900 expansion coefficients in angle. In general, the polynomial degree shall be chosen large enough to ensure a correct behavior of the scheme but still small enough to stay in a reasonable computational regime. The time step size is chosen to be $\Delta t = C_{\text{CFL}} \cdot \Delta x$ with a CFL number of $C_{\text{CFL}} = 0.7$. We compare the solution of the 2D full system corresponding to (5.22) to the solution obtained from the 2D DLRA scheme corresponding to (5.12). The extension to two dimensions is straightforward.
In Figure 5.3 we show numerical results for the scalar flux $\Phi = \int_{S^2} f\,(t, \mathbf{x}, \boldsymbol{\Omega})\,\mathrm{d}\boldsymbol{\Omega}$ and the temperature $T = 4\pi\sqrt{2}\sqrt[4]{B}$ at the end time $t_{\text{end}} = 0.5$. We again observe the accuracy of the proposed DLRA scheme. For the evolution of the rank r in time and the evolution of the relative mass error $\frac{|m^0 - m^n|}{|m^0|}$ in time we consider a time interval up to $t_{\text{end}} = 1.5$. In Figure 5.4 one can observe that for a chosen tolerance parameter of $\vartheta = 5 \cdot 10^{-4} \|\boldsymbol{\Sigma}\|_F$ the rank increases but does not approach its maximal allowed value of $r_{\max} = 100$. Further, the relative mass error stagnates at order $\mathcal{O}\left(10^{-11}\right)$ and the DLRA method shows its mass conservation property. For this setup, the computational benefit of the DLRA method is significant. The scheme is implemented in Julia v1.7 and performed on a MacBook Pro with M1 chip, resulting in a decrease of run time by a factor of approximately 8 from 20023 seconds to 2509 seconds.

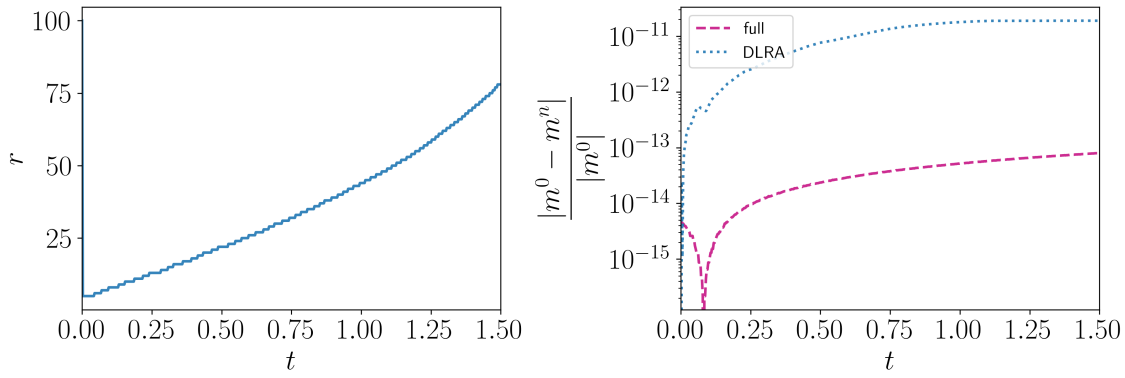


Figure 5.4: Evolution of the rank in time for the 2D beam problem for the DLRA method (left) and evolution of the relative mass error in time compared for both methods (right) until time $t_{\text{end}} = 1.5$.

5.7 Summary and conclusion

We have introduced an energy stable and mass conservative DLRA scheme for the Su-Olson problem. The main research contributions are:

(i) *An energy stable numerical scheme with rigorous mathematical proofs:* We have shown that a naive IMEX scheme fails to guarantee energy stability. To overcome this unphysical behavior, a DLRA scheme which advances radiation and internal energy in a coupled-implicit way has been proposed. In addition, a classic hyperbolic CFL condition has been derived, enabling to operate up to an optimal time step size of $\Delta t = C_{\text{CFL}} \cdot \Delta x$.

(ii) *A mass conservative and rank-adaptive augmented integrator:* We have employed the basis augmentation step described in [CKL22] as well as an adaption of the conservative truncation strategy presented in [EOS23, EKS23] to guarantee local mass conservation and rank adaptivity.

(iii) *Numerical test examples confirming the theoretical properties:* We have compared the numerical results obtained from the DLRA scheme with the solution of the full system for different test examples both in 1D and 2D, underlining the derived properties while showing significantly reduced computational costs and memory requirements, especially in the 2D setting.

Altogether, we have proposed a novel coupled-implicit energy stable DLRA scheme from which conclusions on an appropriate discretization strategy regarding stability can be drawn. For future work, we propose to implement the parallel integrator described in [CKL24] for further enhancing the efficiency of the DLRA method.

6

A multiplicative DLRA scheme for the Su-Olson problem

For the construction of efficient DLRA schemes the structure of the underlying problem has to be taken into account. For instance, it has been shown in [EHY21] that for deriving an efficient DLRA scheme for the non-linear isothermal Boltzmann-BGK equation it is advantageous to consider a multiplicative splitting of the distribution function. This allows for a separation of a generally not low-rank Maxwellian from a remaining low-rank function, to which the DLRA scheme is subsequently applied. To transfer knowledge about the construction of efficient DLRA schemes from the Su-Olson problem considered in Chapter 5 to more general kinetic equations such as given in [EHY21], we reconsider the Su-Olson problem in this chapter and decide on a multiplicative splitting of the distribution function. One difficulty arising in this context is the treatment of the spatial derivatives. For the temporal discretization again the potentially stiff opacity term has to be taken into account, leading to a coupled-implicit scheme, which is complicated to solve. In addition, the multiplicative splitting poses further challenges for the proof of energy stability and for the construction of a DLRA scheme to which we account for instance by pursuing a "first discretize, then low-rank" approach.

The structure of this chapter is as follows. In Section 6.1 we explain the considered multiplicative structure and derive two possible systems for the thermal radiative transfer equations (RTEs) with multiplicative splitting, which in the continuous setting are equivalent. In Section 6.2 a discretization of both systems in angle, space and time is given. Section 6.3 is devoted to the subject of energy stability. We show that the advection form of the multiplicative Su-Olson problem is generally not stable in the sense of von Neumann, whereas for the conservative form an energy estimate can be derived under a classic hyperbolic CFL condition. In Section 6.4 an energy stable DLRA scheme is presented. In addition, mass conservation is shown in Section 6.5 when additional basis augmentations and a suitable truncation strategy are used. The numerical results in Section 6.6 confirm the derived properties before in Section 6.7 a brief summary and conclusion are given. The results of this chapter closely follow the presentation in [BEKK25b].

6.1 Thermal radiative transfer equations with multiplicative splitting

We start from the Su-Olson problem given in equations (5.1) and decide on a multiplicative splitting of the distribution function f of the form

$$f(t,x,\mu) = B(t,x)\, g(t,x,\mu). \tag{6.1}$$

Similar to [EHY21], we apply a DLRA approach to the function g. For this system, we derive a mathematically rigorous proof of energy stability and a hyperbolic CFL condition. As resembling in structure, this chapter can be understood as an intermediate step from the Su-Olson problem treated in Chapter 5 towards more complicated Boltzmann-BGK problems with multiplicative splitting as treated in [EHY21], where the time step size of the proposed algorithm is not theoretically determined by means of analytical considerations but experimentally chosen small enough to ensure good agreement in numerical experiments. We insert the multiplicative splitting (6.1) into the continuous Su-Olson problem (5.1) and obtain the set of equations

$$\begin{aligned}\partial_t g(t,x,\mu) = &-\mu\partial_x g(t,x,\mu) + \sigma(1-g(t,x,\mu)) - \frac{g(t,x,\mu)}{B(t,x)}\partial_t B(t,x) \\ &- \mu\frac{g(t,x,\mu)}{B(t,x)}\partial_x B(t,x),\end{aligned} \tag{6.2a}$$

$$\partial_t B(t,x) = \sigma B(t,x)\left(\langle g(t,x,\mu)\rangle_\mu - 2\right), \tag{6.2b}$$

which is called the *advection form* of the multiplicative system. Using the product rule, it splits up the spatial derivatives for B and g in (6.2a). This corresponds to the form in which the multiplicative splitting in [EHY21] is applied to the non-linear isothermal Boltzmann-BGK equation. Equation (6.2a) can be equivalently rewritten into a *conservative form*, leaving the spatial derivative of Bg together and leading to the system

$$\begin{aligned}\partial_t g(t,x,\mu) = &-\frac{\mu}{B(t,x)}\partial_x\left(B(t,x)\, g(t,x,\mu)\right) + \sigma(1-g(t,x,\mu)) \\ &- \frac{g(t,x,\mu)}{B(t,x)}\partial_t B(t,x),\end{aligned} \tag{6.3a}$$

$$\partial_t B(t,x) = \sigma B(t,x)\left(\langle g(t,x,\mu)\rangle_\mu - 2\right). \tag{6.3b}$$

Note that for both systems we omit initial and boundary conditions for now. In subsequent considerations, our studies include the conservation properties of the derived numerical scheme. The mass and the momentum of the multiplicative system are defined as follows.

Definition 6.1 (Macroscopic quantities)**.** The *mass* of the multiplicative Su-Olson problem is defined as

$$\rho(t,x) := \int f(t,x,\mu)\,\mathrm{d}\mu + B(t,x) = B(t,x)\int g(t,x,\mu)\,\mathrm{d}\mu + B(t,x).$$

The *momentum* is given by

$$\bar{u}(t,x) := \int \mu f(t,x,\mu)\,\mathrm{d}\mu = B(t,x)\int \mu g(t,x,\mu)\,\mathrm{d}\mu.$$

In particular, the multiplicative Su-Olson problem satisfies the local conservation law

$$\partial_t \rho(t,x) + \partial_x \bar{u}(t,x) = 0. \tag{6.4}$$

In the following sections, we discretize both sets of equations (6.2) and (6.3) to compare them in terms of numerical stability. We derive an energy stable DLRA scheme and give a concrete hyperbolic CFL condition. Note that in contrast to Chapter 5 and to [EHY21], we first discretize the equations and then apply a DLRA approach here.

6.2 Discretization of the multiplicative system

In this section, we fully discretize the advection form (6.2) as well as the conservative form (6.3) of the multiplicative Su-Olson problem. We start with the angular and spatial discretizations in Sections 6.2.1 and 6.2.2, followed by the temporal discretization in Section 6.2.3.

6.2.1 Angular discretization

For the angular discretization a modal approach with normalized rescaled Legendre polynomials $P_\ell(\mu)$ as introduced in Section 3.3.2 is applied. The rescaled Legendre polynomials constitute a complete set of orthogonal functions on the interval $[-1,1]$ and satisfy $\langle P_k(\mu), P_\ell(\mu)\rangle_\mu = \delta_{k\ell}$. We approximate the distribution function g in terms of a finite expansion with N_μ expansion coefficients of the form

$$g(t,x,\mu) \approx g_{N_\mu}(t,x,\mu) = \sum_{\ell=0}^{N_\mu-1} v_\ell(t,x)\,P_\ell(\mu).$$

We insert this representation into the advection form (6.2), multiply (6.2a) with $P_k(\mu)$ and integrate over μ. Together with the matrix $\mathbf{A} \in \mathbb{R}^{N_\mu \times N_\mu}$ defined in (3.26) we obtain the angularly discretized equations

$$\begin{aligned}\partial_t v_k(t,x) = &-\sum_{\ell=0}^{N_\mu-1} \partial_x v_\ell(t,x)\,A_{k\ell} + \sigma\left(\sqrt{2}\delta_{k0} - v_k(t,x)\right) - \frac{v_k(t,x)}{B(t,x)}\partial_t B(t,x) \\ &- \sum_{\ell=1}^{N_\mu-1} \frac{v_\ell(t,x)}{B(t,x)}\partial_x B(t,x)\,A_{k\ell},\end{aligned} \tag{6.5a}$$

$$\partial_t B(t,x) = \sigma B(t,x)\left(\sqrt{2}v_0(t,x) - 2\right). \tag{6.5b}$$

Analogously, we obtain for the conservative form (6.3) the following equations

$$\partial_t v_k(t,x) = -\frac{1}{B(t,x)} \sum_{\ell=0}^{N_\mu-1} \partial_x \left(B(t,x)\, v_\ell(t,x)\right) A_{k\ell} + \sigma\left(\sqrt{2}\delta_{k0} - v_k(t,x)\right) \tag{6.6a}$$
$$- \frac{v_k(t,x)}{B(t,x)} \partial_t B(t,x),$$
$$\partial_t B(t,x) = \sigma B(t,x)\left(\sqrt{2} v_0(t,x) - 2\right). \tag{6.6b}$$

6.2.2 Spatial discretization

For the spatial discretization we construct a spatial grid with N_x grid cells and equidistant spacing $\Delta x = \frac{1}{N_x}$. Spatially dependent quantities are approximated as

$$B_j(t) \approx B(t, x_j), \qquad v_{jk}(t) \approx v_k(t, x_j) \qquad \text{for } \; j = 1, ..., N_x.$$

Assuming periodic boundary conditions, first-order spatial derivatives ∂_x are approximated using the centered FD method. For stability reasons, a diffusion term involving second-order derivatives ∂_{xx} is added. This term is also approximated by the centered FD method. We employ the tridiagonal spatial stencil matrices $\mathbf{D}^x \in \mathbb{R}^{N_x \times N_x}$ given in (3.8) and $\mathbf{D}^{xx} \in \mathbb{R}^{N_x \times N_x}$ defined in (3.11). Recall that the symmetric matrix $\mathbf{A}$ is diagonalizable in the form $\mathbf{A} = \mathbf{Q}\mathbf{M}\mathbf{Q}^\top$ with $\mathbf{Q}$ being orthogonal and $\mathbf{M} = \mathrm{diag}(\sigma_0, ..., \sigma_{N_\mu-1})$ and that we have defined $|\mathbf{A}| = \mathbf{Q}|\mathbf{M}|\mathbf{Q}^\top$. We insert the proposed discretization into the advection form (6.5) and add a second-order stabilization term for $\partial_x v$. This leads to

$$\dot{v}_{jk}(t) = -\sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} D^x_{ji} v_{i\ell}(t)\, A_{k\ell} + \frac{\Delta x}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} D^{xx}_{ji} v_{i\ell}(t)\, |A|_{k\ell} \tag{6.7a}$$
$$+ \sigma\left(\sqrt{2}\delta_{k0} - v_{jk}(t)\right) - \frac{v_{jk}(t)}{B_j(t)} \dot{B}_j(t) - \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} \frac{v_{j\ell}(t)}{B_j(t)} D^x_{ji} B_i(t)\, A_{k\ell},$$
$$\dot{B}_j(t) = \sigma B_j(t)\left(\sqrt{2} v_{j0}(t) - 2\right). \tag{6.7b}$$

Inserting the discretization into the conservative form (6.6) and adding a second-order stabilization term to $\partial_x(Bv)$ gives

$$\dot{v}_{jk}(t) = -\frac{1}{B_j(t)} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} D^x_{ji} B_i(t)\, v_{i\ell}(t)\, A_{k\ell} + \frac{\Delta x}{2} \frac{1}{B_j(t)} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu-1} D^{xx}_{ji} B_i(t)\, v_{i\ell}(t)\, |A|_{k\ell}$$
$$+ \sigma\left(\sqrt{2}\delta_{k0} - v_{jk}(t)\right) - \frac{v_{jk}(t)}{B_j(t)} \dot{B}_j(t), \tag{6.8a}$$
$$\dot{B}_j(t) = \sigma B_j(t)\left(\sqrt{2} v_{j0}(t) - 2\right). \tag{6.8b}$$

Note that due to the different structure of the equations the stabilization term in (6.7a) is applied to $\partial_x v$, whereas in (6.8a) it is added for $\partial_x(Bv)$.

6.2.3 Temporal discretization

In Section 5.4 it is shown that constructing a fully discrete energy stable scheme for the Su-Olson problem is challenging. For the advection form we begin with equations (6.7) and apply an explicit Euler step to the transport terms. The potentially stiff absorption term is treated implicitly and the time derivative $\partial_t B$ is approximated by its difference quotient. We obtain the following fully discrete space-time discretization

$$\begin{aligned} v_{jk}^{n+1} &= v_{jk}^n - \Delta t \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} D_{ji}^x v_{i\ell}^n A_{k\ell} + \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} D_{ji}^{xx} v_{i\ell}^n |A|_{k\ell} \\ &\quad + \sigma \Delta t \left(\sqrt{2} \delta_{k0} - v_{jk}^{n+1} \right) - \frac{v_{jk}^{n+1}}{B_j^n} \left(B_j^{n+1} - B_j^n \right) - \Delta t \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \frac{v_{j\ell}^n}{B_j^n} D_{ji}^x B_i^n A_{k\ell}, \end{aligned} \tag{6.9a}$$

$$B_j^{n+1} = B_j^n + \sigma \Delta t B_j^{n+1} \left(\sqrt{2} v_{j0}^{n+1} - 2 \right), \tag{6.9b}$$

which describes one time step from time t_n to time $t_{n+1} = t_n + \Delta t$. For the conservative form (6.8), we again apply an explicit Euler step to the transport parts, treat the absorption term implicitly and approximate $\partial_t B$ by its difference quotient. In addition, we add the factor $\frac{B_j^{n+1}}{B_j^n}$ in the absorption term of (6.8a). This gives the fully discrete scheme

$$\begin{aligned} v_{jk}^{n+1} &= v_{jk}^n - \Delta t \frac{1}{B_j^n} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} D_{ji}^x B_i^n v_{i\ell}^n A_{k\ell} + \Delta t \frac{\Delta x}{2} \frac{1}{B_j^n} \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} D_{ji}^{xx} B_i^n v_{i\ell}^n |A|_{k\ell} \\ &\quad + \sigma \Delta t \frac{B_j^{n+1}}{B_j^n} \left(\sqrt{2} \delta_{k0} - v_{jk}^{n+1} \right) - \frac{v_{jk}^{n+1}}{B_j^n} \left(B_j^{n+1} - B_j^n \right), \end{aligned} \tag{6.10a}$$

$$B_j^{n+1} = B_j^n + \sigma \Delta t B_j^{n+1} \left(\sqrt{2} v_{j0}^{n+1} - 2 \right). \tag{6.10b}$$

Note that the evolution equations (6.9b) and (6.10b) for the internal energy B are the same in both schemes. The main difference of (6.9a) and (6.10a) consists in the distinct second-order stabilization terms and the additional factor $\frac{B_j^{n+1}}{B_j^n}$ in (6.10a), which will be explained later in the proof of energy stability.

6.3 Energy stability

The goal of this section consists in investigating energy stability of the derived schemes. Note that this section is closely related to the considerations in Section 5.4. We first introduce the following notations.

Definition 6.2. In the following considerations, we denote $u_{jk}^n := B_j^n v_{jk}^n$. Note that $\mathbf{u}^n = \left(u_{jk}^n \right) \in \mathbb{R}^{N_x \times N_\mu}$ corresponds to the angularly and spatially discretized $f(t, x, \mu)$ at time t_n and $\mathbf{v}^n = \left(v_{jk}^n \right) \in \mathbb{R}^{N_x \times N_\mu}$ corresponds to the angularly and spatially discretized $g(t, x, \mu)$ at time t_n in representation (6.1).

Then the definition of the total energy of a fully discrete system can be given.

Definition 6.3 (Fully discrete total energy)**.** Let $\mathbf{u}^n \in \mathbb{R}^{N_x \times N_\mu}$ be the fully discrete angular solution to the full Su-Olson problem at time t_n and $\mathbf{B}^n = \left(B_j^n\right) \in \mathbb{R}^{N_x}$ the internal energy at time t_n. The *fully discrete total energy at time* t_n is defined as

$$E^n := \frac{1}{2} \|\mathbf{u}^n\|_F^2 + \frac{1}{2} \|\mathbf{B}^n\|_E^2 ,$$

where $\|\cdot\|_F$ denotes the Frobenius and $\|\cdot\|_E$ the Euclidean norm.

In Section 6.3.1 it is shown that the advection form (6.9), in general, is not numerically stable. Section 6.3.2 presents a proof of energy stability for the conservative form (6.10).

6.3.1 Advection form

We begin with the advection form (6.9) of the Su-Olson problem, which is comparable to the considered DLRA discretization in [EHY21] for the isothermal Boltzmann-BGK equation in the sense that the term $\partial_x (Bv)$ is split up into the sum of $B\partial_x v$ and $v\partial_x B$. We can show that this scheme is, in general, not von Neumann stable.

Theorem 6.4. *There exist initial values* $\mathbf{v}^n \in \mathbb{R}^{N_x \times N_\mu}$ *and* $\mathbf{B}^n \in \mathbb{R}^{N_x}$ *such that the advection form* (6.9) *of the Su-Olson problem for* $\sigma = 0$ *is not von Neumann stable.*

Proof. Let us assume a solution v_{jk}^n that is constant in space and direction, e.g. $v_{jk}^n = 1$. For this solution all spatial derivatives are zero and the terms containing $\mathbf{D}^x\mathbf{v}^n$ and $\mathbf{D}^{xx}\mathbf{v}^n$ in (6.9a) drop out. We further assume that for the opacity it holds $\sigma = 0$, i.e. the Su-Olson problem reduces to a simple advection equation. From (6.9b) we thus derive the equality $B_j^{n+1} = B_j^n = B_j$, i.e. the internal energy is constant in time. We insert these results into (6.9a) and obtain

$$v_{jk}^{n+1} = 1 - \Delta t \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} \frac{1}{B_j} D_{ji}^x B_i A_{k\ell}.$$

Multiplication with B_j leads to

$$u_{jk}^{n+1} = u_{jk}^n - \Delta t \sum_{i=1}^{N_x} \sum_{\ell=0}^{N_\mu - 1} D_{ji}^x u_{i\ell}^n A_{k\ell}.$$

This is a discretization of $\partial_t u + \mu \partial_x u = 0$ with an explicit Euler step forward in time and a centered FD scheme in space. According to Remark 3.16 this discretization is not von Neumann stable. □

The consideration of this special case demonstrates that for the fully discrete advection form (6.9) of the multiplicative Su-Olson problem numerical stability in the sense of von Neumann as described in Section 3.2.3 cannot be guaranteed. In this sense, Theorem 6.4 serves as a motivation to seek a generally stable numerical discretization as done in the next section.

6.3.2 Conservative form

For the conservative form (6.10) we are able to derive a hyperbolic CFL condition and to show that under this time step restriction the total energy of the system dissipates.

Theorem 6.5 (Energy stability of the fully discrete system)**.** *Under the time step restriction $\Delta t \leq \Delta x$ the fully discrete system (6.10) is energy stable, i.e. it holds $E^{n+1} \leq E^n$.*

Proof. The proof of this theorem is similar to the proof of Theorem 5.9. We start with equation (6.10b) and multiply it with B_j^{n+1}. This gives

$$\left(B_j^{n+1}\right)^2 = B_j^n B_j^{n+1} + \sigma \Delta t \left(B_j^{n+1}\right)^2 \left(\sqrt{2} v_{j0}^{n+1} - 2\right).$$

We insert relation (5.15) and sum over j, leading to

$$\frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1}\right)^2 = \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^n\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1} - B_j^n\right)^2 + \sigma\Delta t \sum_{j=1}^{N_x}\left(B_j^{n+1}\right)^2\left(\sqrt{2}v_{j0}^{n+1} - 2\right). \tag{6.11}$$

To obtain a similar expression for $\left(u_{jk}^{n+1}\right)^2$, we multiply (6.10a) with $B_j^{n+1} B_j^n v_{jk}^{n+1}$, sum over j and k, and use the notation $u_{jk}^n = B_j^n v_{jk}^n$. We obtain

$$\begin{aligned}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1} B_j^{n+1} B_j^n \left(v_{jk}^{n+1}\right)^2 = &\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1} u_{jk}^n u_{jk}^{n+1} - \Delta t \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x u_{i\ell}^n A_{k\ell} \\ &+ \Delta t \frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^n |A|_{k\ell} \\ &+ \sigma\Delta t \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0} - v_{jk}^{n+1}\right) \\ &- \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1} B_j^{n+1}\left(v_{jk}^{n+1}\right)^2\left(B_j^{n+1} - B_j^n\right).\end{aligned} \tag{6.12}$$

Note that for this step the additional factor $\frac{B_j^{n+1}}{B_j^n}$ in (6.10a) is crucial. We insert relation (5.18) into (6.12), put the last term of (6.12) to the left-hand side and rearrange. Then,

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\right)^2 = &\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^n\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1} - u_{jk}^n\right)^2 \\ &- \Delta t \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x u_{i\ell}^n A_{k\ell} + \Delta t \frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^n |A|_{k\ell} \\ &+ \sigma\Delta t \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0} - v_{jk}^{n+1}\right).\end{aligned} \tag{6.13}$$

In the next step, we add artificial zero terms to the equation. Adding the zero term $\Delta t \sum_{i,j=1}^{N_x} \sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x u_{i\ell}^{n+1} A_{k\ell}$ and adding and subtracting the second-order term $\Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x} \sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^{n+1} |A|_{k\ell}$ leads to

$$\begin{aligned}
\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\right)^2 = \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n}\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}-u_{jk}^{n}\right)^2 \\
- \Delta t \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x \left(u_{i\ell}^{n}-u_{i\ell}^{n+1}\right) A_{k\ell} \qquad \text{(I)} \\
+ \Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} \left(u_{i\ell}^{n}-u_{i\ell}^{n+1}\right) |A|_{k\ell} \qquad \text{(II)} \\
+ \Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^{n+1} |A|_{k\ell} \qquad \text{(III)} \\
+ \sigma \Delta t \sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0}-v_{jk}^{n+1}\right).
\end{aligned}$$

We proceed by analyzing the terms (I), (II), and (III) separately. Let us start with (I) and (II) and apply Young's inequality given in Lemma 5.6. For the sum (I) + (II) this results in

$$\begin{aligned}
&- \Delta t \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^x \left(u_{i\ell}^{n}-u_{i\ell}^{n+1}\right) A_{k\ell} + \Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} \left(u_{i\ell}^{n}-u_{i\ell}^{n+1}\right) |A|_{k\ell} \\
&= - \Delta t \sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(u_{i\ell}^{n}-u_{i\ell}^{n+1}\right)\left(\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(D_{ji}^x u_{jk}^{n+1} A_{k\ell} - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell}\right)\right) \\
&\leq \frac{1}{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(u_{i\ell}^{n}-u_{i\ell}^{n+1}\right)^2 \\
&\quad + \frac{(\Delta t)^2}{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(D_{ji}^x u_{jk}^{n+1} A_{k\ell} - \frac{\Delta x}{2} D_{ji}^{xx} u_{jk}^{n+1} |A|_{k\ell}\right)\right)^2.
\end{aligned}$$

For (III) we exploit the properties of the stencil matrices given in Lemma 3.1. This leads to the equality

$$\Delta t \frac{\Delta x}{2} \sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1} D_{ji}^{xx} u_{i\ell}^{n+1} |A|_{k\ell} = -\Delta t \frac{\Delta x}{2} \sum_{j=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(\sum_{i=1}^{N_x}\sum_{k=0}^{N_\mu-1} D_{ji}^{+} u_{ik}^{n+1} |A|_{k\ell}^{1/2}\right)^2.$$

We insert both relations and obtain

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\right)^2 \leq \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n}\right)^2 \\ + \frac{(\Delta t)^2}{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(D_{ji}^{x}u_{jk}^{n+1}A_{k\ell}\right.\right. \\ \left.\left. - \frac{\Delta x}{2}D_{ji}^{xx}u_{jk}^{n+1}\left|A\right|_{k\ell}\right)\right)^2 \\ - \Delta t\frac{\Delta x}{2}\sum_{j=1}^{N_x}\sum_{\ell=0}^{N_\mu-1}\left(\sum_{i=1}^{N_x}\sum_{k=0}^{N_\mu-1}D_{ji}^{+}u_{ik}^{n+1}\left|A\right|_{k\ell}^{1/2}\right)^2 \\ + \sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0}-v_{jk}^{n+1}\right).\end{aligned}$$

With Lemma 5.8 we can conclude that under the time step restriction $\Delta t \leq \Delta x$ it holds

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\right)^2 \leq \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n}\right)^2 \\ + \sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0}-v_{jk}^{n+1}\right).\end{aligned} \tag{6.14}$$

To obtain an expression for the total energy of the system given in Definition 6.3, we add equations (6.14) and (6.11). This yields

$$\begin{aligned}E^{n+1} \leq E^n - \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1}-B_j^n\right)^2 + \sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0}-v_{jk}^{n+1}\right) \\ + \sigma\Delta t\sum_{j=1}^{N_x}\left(B_j^{n+1}\right)^2\left(\sqrt{2}v_{j0}^{n+1}-2\right).\end{aligned}$$

The term $-\frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1}-B_j^n\right)^2$ is non-positive. The remaining two terms on the right-hand side can be rewritten and bounded as follows:

$$\begin{aligned}&\sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0}-v_{jk}^{n+1}\right) + \sigma\Delta t\sum_{j=1}^{N_x}\left(B_j^{n+1}\right)^2\left(\sqrt{2}v_{j0}^{n+1}-2\right) \\ \leq\ &\sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2\left(-\left(v_{jk}^{n+1}\right)^2+2\sqrt{2}v_{jk}^{n+1}\delta_{k0}-2\delta_{k0}\right) \\ =\ &-\sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2\left(v_{jk}^{n+1}-\sqrt{2}\delta_{k0}\right)^2 \leq 0.\end{aligned}$$

Hence, we have shown that under the time step restriction $\Delta t \leq \Delta x$ it holds $E^{n+1} \leq E^n$, and the system is energy stable. □

6.4 Energy stable DLRA scheme for multiplicative Su-Olson

Having attained an energy stable discretization of the multiplicative Su-Olson problem, its practical implementation can still pose numerical challenges such as large memory demands and computational costs, especially in higher-dimensional settings. To overcome these problems, we apply the concept of DLRA to the energy stable conservative form (6.10) of the Su-Olson problem to evolve $\mathbf{v}^n = \left(v_{jk}^n\right)$ to $\mathbf{v}^{n+1} = \left(v_{jk}^{n+1}\right)$. First note that for the derivation of the DLRA scheme we rewrite the equations given in (6.10). In (6.10a), we put all terms containing v_{jk}^{n+1} to the left-hand side and divide by $1+\sigma\Delta t$. Further, we multiply (6.10b) with $\frac{1}{B_j^n}$. This establishes the system

$$\begin{aligned}\frac{B_j^{n+1}}{B_j^n} v_{jk}^{n+1} &= \frac{1}{1+\sigma\Delta t} v_{jk}^n - \frac{\Delta t}{1+\sigma\Delta t}\frac{1}{B_j^n}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^x B_i^n v_{i\ell}^n A_{k\ell} \\ &\quad + \frac{\Delta t}{1+\sigma\Delta t}\frac{\Delta x}{2}\frac{1}{B_j^n}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{xx} B_i^n v_{i\ell}^n |A|_{k\ell} + \frac{\sqrt{2}\sigma\Delta t}{1+\sigma\Delta t}\frac{B_j^{n+1}}{B_j^n}\delta_{k0}, \end{aligned} \tag{6.15a}$$

$$\frac{B_j^{n+1}}{B_j^n} = 1 + \sigma\Delta t \frac{B_j^{n+1}}{B_j^n}\left(\sqrt{2}v_{j0}^{n+1} - 2\right). \tag{6.15b}$$

In what follows, we derive an energy stable and mass conservative DLRA discretization for equations (6.15) which makes use of the rank-adaptive augmented BUG integrator described in [CKL22] together with additional basis augmentations and a conservative truncation strategy. In detail, the DLRA scheme works as follows.

In the first step of the scheme, an update of the quantity $v_{jk}^n = \sum_{m,\eta=1}^r X_{jm}^n S_{m\eta}^n V_{k\eta}^n$ to $\frac{B_j^{n+1}}{B_j^n} v_{jk}^* = \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{jm}^* \widehat{\widehat{S}}_{m\eta}^* \widehat{\widehat{V}}_{k\eta}^*$ is performed for $k \neq 0$. We introduce the notation $K_{j\eta}^n = \sum_{m=1}^r X_{jm}^n S_{m\eta}^n$ and solve the K-step equation

$$\begin{aligned} K_{jp}^* &= \frac{1}{1+\sigma\Delta t} K_{jp}^n - \frac{\Delta t}{1+\sigma\Delta t}\frac{1}{B_j^n}\sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{\eta=1}^{r} K_{i\eta}^n \sum_{k,\ell=0}^{N_\mu-1} V_{\ell\eta}^n A_{k\ell} V_{kp}^n \\ &\quad + \frac{\Delta t}{1+\sigma\Delta t}\frac{\Delta x}{2}\frac{1}{B_j^n}\sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{\eta=1}^{r} K_{i\eta}^n \sum_{k,\ell=0}^{N_\mu-1} V_{\ell\eta}^n |A|_{k\ell} V_{kp}^n. \end{aligned} \tag{6.16a}$$

The updated basis $\widehat{\mathbf{X}}^*$ of rank $2r$ is derived from a QR-decomposition of the augmented quantity $\widehat{\mathbf{X}}^* = \mathrm{qr}\left([\mathbf{K}^*, \mathbf{X}^n]\right)$. Moreover, we perform an additional basis augmentation

step according to

$$\widehat{\widehat{\mathbf{X}}}^* = \mathrm{qr}\left(\left[\widehat{\mathbf{X}}^*, \frac{1}{\mathbf{B}^n} \odot \mathbf{D}^x \left(\mathbf{B}^n \odot \mathbf{X}^n\right), \frac{1}{\mathbf{B}^n} \odot \mathbf{D}^{xx} \left(\mathbf{B}^n \odot \mathbf{X}^n\right)\right]\right), \tag{6.16b}$$

which ensures the exactness of the corresponding projection operators in the proof of energy stability of the DLRA scheme. The symbol $\odot$ denotes a pointwise multiplication and the vector $\frac{1}{\mathbf{B}^n} \in \mathbb{R}^{N_x}$ is defined to contain the elements $\frac{1}{B_j^n}$ for each $j = 1, ..., N_x$. In addition, we compute and store $\widehat{\widehat{\mathbf{M}}} = \widehat{\widehat{\mathbf{X}}}^{*,\top} \mathbf{X}^n$. Note that for this scheme we perform full rank updates, leading to an increase from rank $2r$ to $4r$. Quantities of rank $2r$ are denoted with one single hat and quantities of rank $4r$ with double hats.

The L-step can be computed in parallel with the K-step. We introduce the notation $L_{km}^n = \sum_{\eta=1}^r S_{m\eta}^n V_{k\eta}^n$ and solve

$$\begin{aligned} L_{kp}^* = &\frac{1}{1+\sigma\Delta t} L_{kp}^n - \frac{\Delta t}{1+\sigma\Delta t} \sum_{\ell=0}^{N_\mu - 1} A_{k\ell} \sum_{m=1}^{r} L_{\ell m}^n \sum_{i=1}^{N_x} X_{im}^n B_i^n \sum_{j=1}^{N_x} D_{ji}^x \frac{1}{B_j^n} X_{jp}^n \\ &+ \frac{\Delta t}{1+\sigma\Delta t} \frac{\Delta x}{2} \sum_{\ell=0}^{N_\mu - 1} |A|_{k\ell} \sum_{m=1}^{r} L_{\ell m}^n \sum_{i=1}^{N_x} X_{im}^n B_i^n \sum_{j=1}^{N_x} D_{ji}^{xx} \frac{1}{B_j^n} X_{jp}^n. \end{aligned} \tag{6.16c}$$

The updated basis $\widehat{\mathbf{V}}^*$ of rank $2r$ is derived from a QR-decomposition of the augmented quantity $\widehat{\mathbf{V}}^* = \mathrm{qr}\left([\mathbf{L}^*, \mathbf{V}^n]\right)$. Moreover, we perform an additional basis augmentation step according to

$$\widehat{\widehat{\mathbf{V}}}^* = \mathrm{qr}\left(\left[\widehat{\mathbf{V}}^*, \mathbf{A}^\top \mathbf{V}^n, |\mathbf{A}|^\top \mathbf{V}^n\right]\right), \tag{6.16d}$$

leading to a new augmented basis $\widehat{\widehat{\mathbf{V}}}^{n+1}$ of rank $4r$. This basis augmentation again ensures the exactness of the corresponding projection operators and will be made clear in the proof of energy stability of the DLRA scheme later. In addition, we compute and store $\widehat{\widehat{\mathbf{N}}} = \widehat{\widehat{\mathbf{V}}}^{*,\top} \mathbf{V}^n$.

For the S-step, the previously computed solutions obtained in the K- and L-step are used. We introduce the notation $\widetilde{S}_{m\eta}^n = \sum_{j,k=1}^r \widehat{\widehat{M}}_{mj} S_{jk}^n \widehat{\widehat{N}}_{\eta k}$ and solve the equation

$$\begin{aligned} \widehat{\widehat{S}}_{qp}^* = &\frac{1}{1+\sigma\Delta t} \widetilde{S}_{qp}^n \\ &- \frac{\Delta t}{1+\sigma\Delta t} \sum_{j=1}^{N_x} \widehat{\widehat{X}}_{jq}^* \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{im}^* \widetilde{S}_{m\eta}^n \sum_{k,\ell=0}^{N_\mu - 1} \widehat{\widehat{V}}_{\ell\eta}^* A_{k\ell} \widehat{\widehat{V}}_{kp}^* \\ &+ \frac{\Delta t}{1+\sigma\Delta t} \frac{\Delta x}{2} \sum_{j=1}^{N_x} \widehat{\widehat{X}}_{jq}^* \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{im}^* \widetilde{S}_{m\eta}^n \sum_{k,\ell=0}^{N_\mu - 1} \widehat{\widehat{V}}_{\ell\eta}^* |A|_{k\ell} \widehat{\widehat{V}}_{kp}^*. \end{aligned} \tag{6.16e}$$

In the next step, we consider the equations for $k = 0$. In this case, the expressions for

$\frac{B_j^{n+1}}{B_j^n}\widetilde{v}_{j0}^{n+1}$ and $\frac{B_j^{n+1}}{B_j^n}$ are coupled and we solve the system

$$\begin{aligned}\frac{B_j^{n+1}}{B_j^n}\widetilde{v}_{j0}^{n+1} &= \frac{1}{1+\sigma\Delta t}\sum_{m,\eta=1}^{r} X_{jm}^n S_{m\eta}^n V_{0\eta}^n \\ &\quad - \frac{\Delta t}{1+\sigma\Delta t}\frac{1}{B_j^n}\sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{im}^* \widetilde{S}_{m\eta}^n \sum_{\ell=0}^{N_\mu-1} \widehat{\widehat{V}}_{\ell\eta}^* A_{0\ell} \\ &\quad + \frac{\Delta t}{1+\sigma\Delta t}\frac{\Delta x}{2}\frac{1}{B_j^n}\sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{im}^* \widetilde{S}_{m\eta}^n \sum_{\ell=0}^{N_\mu-1} \widehat{\widehat{V}}_{\ell\eta}^* |A|_{0\ell} \\ &\quad + \frac{\sqrt{2}\sigma\Delta t}{1+\sigma\Delta t}\frac{B_j^{n+1}}{B_j^n},\end{aligned} \tag{6.16f}$$

$$\frac{B_j^{n+1}}{B_j^n} = 1+\sigma\Delta t\frac{B_j^{n+1}}{B_j^n}\left(\sqrt{2}\widetilde{v}_{j0}^{n+1}-2\right). \tag{6.16g}$$

Using equations (6.16f) and (6.16g), $\widetilde{\mathbf{v}}_0^{n+1} = \left(\widetilde{v}_{j0}^{n+1}\right)$ and $\mathbf{B}^{n+1} = \left(B_j^{n+1}\right)$ can be retrieved. The latter is used to multiply the DLRA expression of $\frac{B_j^{n+1}}{B_j^n}v_{jk}^* = \sum_{m,\eta=1}^{4r}\widehat{\widehat{X}}_{jm}^*\widehat{\widehat{S}}_{m\eta}^*\widehat{\widehat{V}}_{k\eta}^*$ with the factor $\frac{B_j^n}{B_j^{n+1}}$ in the form of a transformation step given by

$$K_{jp}^{*,\mathrm{trans}} = \frac{B_j^n}{B_j^{n+1}}K_{jp}^*. \tag{6.16h}$$

From a QR-decomposition we obtain $\widehat{\widehat{\mathbf{X}}}^{*,\mathrm{trans}}\widehat{\widehat{\mathbf{S}}}^{*,\mathrm{trans}} = \mathrm{qr}\left(\mathbf{K}^{*,\mathrm{trans}}\right)$. Then we perform an additional basis augmentation step according to

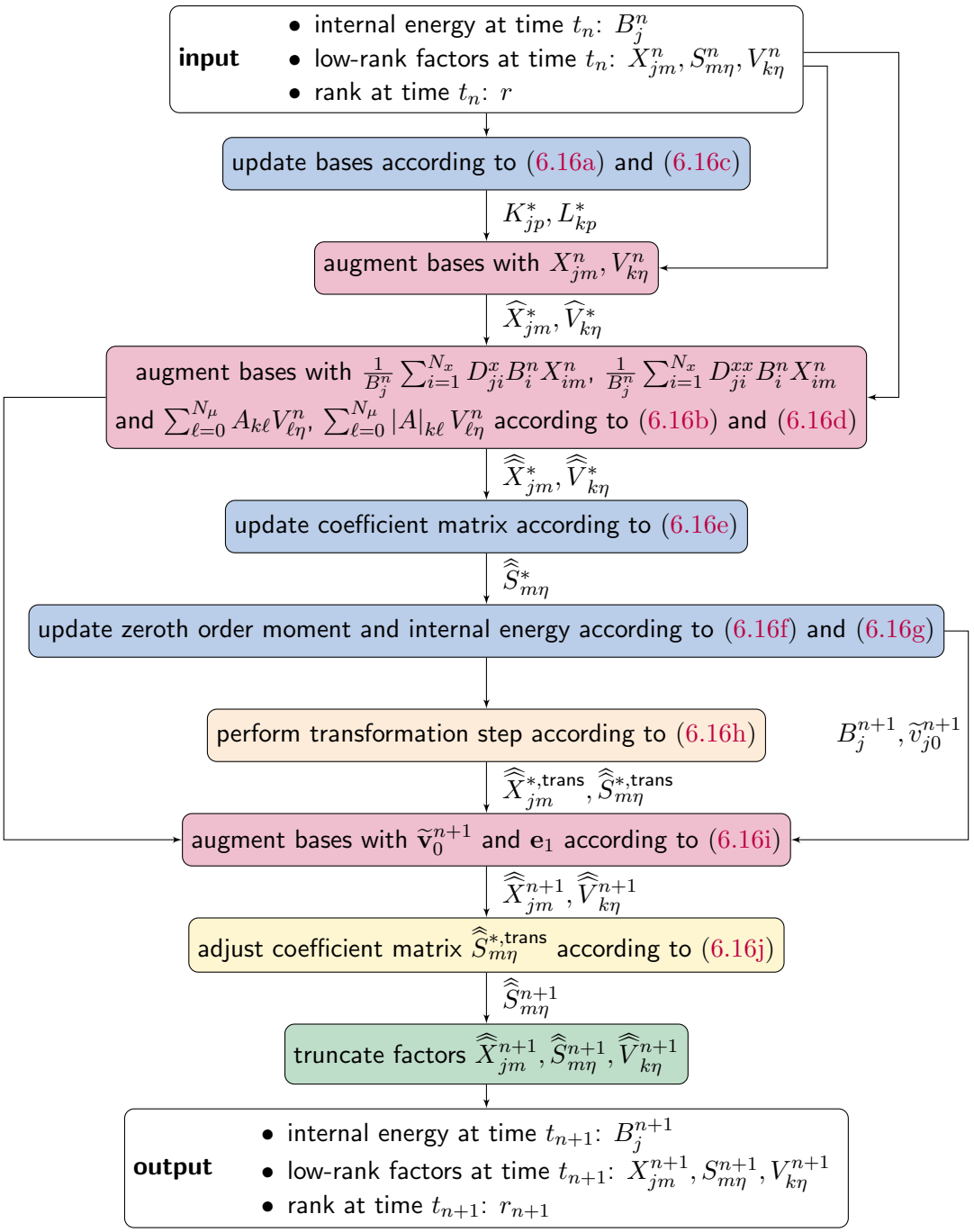
$$\widehat{\widehat{\mathbf{X}}}^{n+1} = \mathrm{qr}\left(\left[\widetilde{\mathbf{v}}_0^{n+1}, \widehat{\widehat{\mathbf{X}}}^{*,\mathrm{trans}}\right]\right) \qquad \text{and} \qquad \widehat{\widehat{\mathbf{V}}}^{n+1} = \mathrm{qr}\left(\left[\mathbf{e}_1, \widehat{\widehat{\mathbf{V}}}^*\right]\right), \tag{6.16i}$$

where we add $\widetilde{\mathbf{v}}_0^{n+1}$ to the updated spatial low-rank basis since the mass of the system is given by the zeroth order moment $\widetilde{\mathbf{v}}_0^{n+1}$. In the directional basis, we add $\mathbf{e}_1 \in \mathbb{R}^{N_\mu}$. Again, these basis augmentations ensure the conservation of mass of the DLRA scheme. Finally, we have to adjust the coefficient matrix from $\widehat{\widehat{\mathbf{S}}}^{*,\mathrm{trans}}$ to $\widehat{\widehat{\mathbf{S}}}^{n+1} \in \mathbb{R}^{(4r+1)\times(4r+1)}$ as

$$\widehat{\widehat{\mathbf{S}}}^{n+1} = \widehat{\widehat{\mathbf{X}}}^{n+1,\top}\widehat{\widehat{\mathbf{X}}}^{*,\mathrm{trans}}\widehat{\widehat{\mathbf{S}}}^{*,\mathrm{trans}}\widehat{\widehat{\mathbf{V}}}^*\left(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^\top\right)\widehat{\widehat{\mathbf{V}}}^{n+1} + \widehat{\widehat{\mathbf{X}}}^{n+1,\top}\widetilde{\mathbf{v}}_0^{n+1}\mathbf{e}_1^\top\widehat{\widehat{\mathbf{V}}}^{n+1}. \tag{6.16j}$$

We obtain the updated solution $\mathbf{v}^{n+1} = \widehat{\widehat{\mathbf{X}}}^{n+1}\widehat{\widehat{\mathbf{S}}}^{n+1}\widehat{\widehat{\mathbf{V}}}^{n+1} \in \mathbb{R}^{N_x\times N_\mu}$. In a last step, we truncate the augmented quantities $\widehat{\widehat{\mathbf{X}}}^{n+1}, \widehat{\widehat{\mathbf{S}}}^{n+1}$ and $\widehat{\widehat{\mathbf{V}}}^{n+1}$ from rank $4r+1$ to a new rank r_{n+1} by using a suitable truncation strategy such as proposed in Section 4.4.2. This eventually gives the low-rank factors $\mathbf{X}^{n+1}, \mathbf{S}^{n+1}$, and $\mathbf{V}^{n+1}$. To provide an overview of the DLRA scheme, its structure is visualized in Algorithm 2. Note that this scheme is significantly different from the one presented in Section 5.4.2 as the multiplicative structure leads to additional basis augmentations introduced in (6.16b) and (6.16d) as well as to a different solution of the coupled equations given in (6.16f) and (6.16g).

Algorithm 2 Flowchart of the energy stable and mass conservative multiplicative DLRA scheme (6.16).



Proof of energy stability of the proposed multiplicative low-rank scheme. It can be shown that the DLRA scheme proposed in (6.16) preserves the energy stability of the full system given in Section 5.4.2. The rewriting of equations (6.10) into (6.15) as well as the basis augmentations introduced in (6.16b) and (6.16d) differentiate this DLRA method from the existing scheme in Section 5.4.2 and are crucial for the proof.

Theorem 6.6 (Energy stability of the proposed multiplicative DLRA scheme)**.** *Under the time step restriction $\Delta t \leq \Delta x$ the fully discrete multiplicative DLRA scheme (6.16) is energy stable, i.e. it holds $E^{n+1} \leq E^n$.*

Proof. Similar to the proof of Theorem 6.5, an estimate for the fully discrete energy introduced in Definition 6.3 is sought. We begin with the internal energy $\mathbf{B}$ and multiply equation (6.16g) with $B_j^n B_j^{n+1}$. This leads to

$$\left(B_j^{n+1}\right)^2 = B_j^n B_j^{n+1} + \sigma \Delta t \left(B_j^{n+1}\right)^2 \left(\sqrt{2} v_{j0}^{n+1} - 2\right).$$

We insert relation (5.15) to rewrite the product $B_j^n B_j^{n+1}$ and sum over j, rendering the expression

$$\begin{aligned} \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1}\right)^2 = \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^n\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\left(B_j^{n+1} - B_j^n\right)^2 \\ + \sigma \Delta t \sum_{j=1}^{N_x}\left(B_j^{n+1}\right)^2 \left(\sqrt{2} v_{j0}^{n+1} - 2\right). \end{aligned} \tag{6.17}$$

This is the same equation as stated in (6.11) in the proof of Theorem 6.5. To obtain a similar expression for $\left(u_{jk}^{n+1}\right)^2$, we multiply equation (6.16e) with $\widehat{\widehat{X}}_{\alpha q}^* \widehat{\widehat{V}}_{\beta p}^*$ and sum over q and p. For simplicity of notation, we introduce $v_{\alpha\beta}^* := \sum_{q,p=1}^{4r} \widehat{\widehat{X}}_{\alpha q}^* \widehat{\widehat{S}}_{qp}^* \widehat{\widehat{V}}_{\beta p}^*$ and $v_{\alpha\beta}^n := \sum_{q,p=1}^{4r} \widehat{\widehat{X}}_{\alpha q}^* \widetilde{S}_{qp}^n \widehat{\widehat{V}}_{\beta p}^*$ as well as the projection operators $P_{\alpha j}^{X^*} = \sum_{q=1}^{4r} \widehat{\widehat{X}}_{\alpha q}^* \widehat{\widehat{X}}_{jq}^*$ and $P_{k\beta}^{V^*} = \sum_{p=1}^{4r} \widehat{\widehat{V}}_{kp}^* \widehat{\widehat{V}}_{\beta p}^*$. We obtain

$$\begin{aligned} v_{\alpha\beta}^* = \frac{1}{1+\sigma\Delta t} v_{\alpha\beta}^n - \frac{\Delta t}{1+\sigma\Delta t} \sum_{j=1}^{N_x} P_{\alpha j}^{X^*} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{k,\ell=0}^{N_\mu - 1} v_{i\ell}^n A_{k\ell} P_{k\beta}^{V^*} \\ + \frac{\Delta t}{1+\sigma\Delta t} \frac{\Delta x}{2} \sum_{j=1}^{N_x} P_{\alpha j}^{X^*} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{k,\ell=0}^{N_\mu - 1} v_{i\ell}^n |A|_{k\ell} P_{k\beta}^{V^*}. \end{aligned} \tag{6.18}$$

Further, we denote $v_{\alpha\beta}^{n+1} := \sum_{q,p=1}^{4r+1} \widehat{\widehat{X}}_{\alpha q}^{n+1} \widehat{\widehat{S}}_{qp}^{n+1} \widehat{\widehat{V}}_{\beta p}^{n+1}$. From equation (6.16j), we can derive the equation

$$\frac{B_\alpha^{n+1}}{B_\alpha^n} v_{\alpha\beta}^{n+1} = v_{\alpha\beta}^* \left(1 - \delta_{\beta 0}\right) + \frac{B_\alpha^{n+1}}{B_\alpha^n} \widetilde{v}_{\alpha 0}^{n+1} \delta_{\beta 0}.$$

Hence, inserting the schemes for $v_{\alpha\beta}^*$ and $\widetilde{v}_{\alpha 0}^{n+1}$, i.e. equations (6.18) and (6.16f), establishes

the expression

$$\begin{aligned}\frac{B_\alpha^{n+1}}{B_\alpha^n} v_{\alpha\beta}^{n+1}\,(1+\sigma\Delta t) = \Bigg(& v_{\alpha\beta}^n - \Delta t \sum_{j=1}^{N_x} P_{\alpha j}^{X^*} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{k,\ell=0}^{N_\mu-1} v_{i\ell}^n A_{k\ell} P_{k\beta}^{V^*} \\ & + \Delta t \frac{\Delta x}{2} \sum_{j=1}^{N_x} P_{\alpha j}^{X^*} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{k,\ell=0}^{N_\mu-1} v_{i\ell}^n |A|_{k\ell} P_{k\beta}^{V^*} \Bigg) (1-\delta_{\beta 0}) \\ + \Bigg(& v_{\alpha 0}^n - \Delta t \sum_{i=1}^{N_x} \frac{1}{B_\alpha^n} D_{\alpha i}^x B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n A_{0\ell} \\ & + \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} \frac{1}{B_\alpha^n} D_{\alpha i}^{xx} B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n |A|_{0\ell} + \sqrt{2}\sigma\Delta t \frac{B_\alpha^{n+1}}{B_\alpha^n} \Bigg) \delta_{\beta 0}.\end{aligned}$$

We proceed by employing the fact that we have augmented the spatial basis according to (6.16b) and (6.16d). This allows us to write any function $h_i^n \in \operatorname{span}(X_i^n)$ and $\tilde{h}_\ell^n \in \operatorname{span}(V_\ell^n)$ as

$$\sum_{j=1}^{N_x} P_{\alpha j}^{X^*} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^x B_i^n h_i^n = \frac{1}{B_\alpha^n} \sum_{i=1}^{N_x} D_{\alpha i}^x B_i^n h_i^n \quad \text{and} \quad \sum_{k,\ell=0}^{N_\mu-1} \tilde{h}_\ell^n A_{k\ell} P_{k\beta}^{V^*} = \sum_{\ell=0}^{N_\mu-1} \tilde{h}_\ell^n A_{\beta\ell},$$

$$\sum_{j=1}^{N_x} P_{\alpha j}^{X^*} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n h_i^n = \frac{1}{B_\alpha^n} \sum_{i=1}^{N_x} D_{\alpha i}^{xx} B_i^n h_i^n \quad \text{and} \quad \sum_{k,\ell=0}^{N_\mu-1} \tilde{h}_\ell^n |A|_{k\ell} P_{k\beta}^{V^*} = \sum_{\ell=0}^{N_\mu-1} \tilde{h}_\ell^n |A|_{\beta\ell}.$$

To be consistent in notation, we change the indices from α to j and from β to k at this point. The basis augmentations as well as the properties of the projection operators enable us to obtain a representation of the form

$$\frac{B_j^{n+1}}{B_j^n} v_{jk}^{n+1}\,(1+\sigma\Delta t) = F\,(1-\delta_{k0}) + F\delta_{k0} + \sqrt{2}\sigma\Delta t \frac{B_j^{n+1}}{B_j^n}\delta_{k0}$$

with

$$F = v_{jk}^n - \Delta t \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n A_{k\ell} + \Delta t \frac{\Delta x}{2} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n\,|A|_{k\ell}\,.$$

On the right-hand side the factor $F\delta_{k0}$ cancels out. This yields the equation

$$\begin{aligned}\frac{B_j^{n+1}}{B_j^n} v_{jk}^{n+1}\,(1+\sigma\Delta t) = & v_{jk}^n - \Delta t \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n A_{k\ell} \\ & + \Delta t \frac{\Delta x}{2} \frac{1}{B_j^n} \sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n\,|A|_{k\ell} + \sqrt{2}\sigma\Delta t \frac{B_j^{n+1}}{B_j^n}\delta_{k0}.\end{aligned}$$

In the next step we multiply this expression with $B_j^{n+1} B_j^n v_{jk}^{n+1}$, sum over j and k, rearrange the obtained equation, use the notation $u_{jk}^n = B_j^n v_{jk}^n$, and insert relation (5.18).

This leads to

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}\right)^2 = \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n}\right)^2 - \frac{1}{2}\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(u_{jk}^{n+1}-u_{jk}^{n}\right)^2 \\ - \Delta t\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1}D_{ji}^{x}u_{i\ell}^{n}A_{k\ell} + \Delta t\frac{\Delta x}{2}\sum_{i,j=1}^{N_x}\sum_{k,\ell=0}^{N_\mu-1} u_{jk}^{n+1}D_{ji}^{xx}v_{i\ell}^{n}\left|A\right|_{k\ell} \\ + \sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=0}^{N_\mu-1}\left(B_j^{n+1}\right)^2 v_{jk}^{n+1}\left(\sqrt{2}\delta_{k0}-v_{jk}^{n+1}\right),\end{aligned}$$

which is the same expression as in equation (6.13) in the proof of Theorem 6.5. We apply the same estimates as in the proof of Theorem 6.5 and add the resulting equation and equation (6.17). Analogously to the proof of Theorem 6.5 and due to the fact that the truncation step does not alter the zeroth order moment, we obtain energy stability of the multiplicative DLRA scheme under the time step restriction $\Delta t \leq \Delta x$. □

6.5 Mass conservation

The multiplicative DLRA scheme described in (6.16) can be shown to be locally mass conservative when using a suitable truncation strategy. For instance, the truncation strategy presented in Section 4.4.2 can be easily adjusted to the considered framework, which includes quantities of rank $4r$ instead of $2r$. We translate the macroscopic quantities introduced in Definition 6.1 to the fully discrete setting.

Definition 6.7 (Fully discrete macroscopic quantities)**.** The *mass* and the *momentum* of the fully discrete multiplicative Su-Olson problem at time t_n are defined as

$$\rho_j^n := \sqrt{2}B_j^n v_{j0}^n + B_j^n \qquad \text{and} \qquad \bar{u}_j^n := \sqrt{2}B_j^n\sum_{\ell=0}^{N_\mu-1} v_{j\ell}^n A_{0\ell}.$$

It can be shown that the DLRA algorithm proposed in (6.16) together with the conservative truncation strategy fulfills the following local conservation law.

Theorem 6.8 (Mass conservation of the proposed multiplicative DLRA scheme)**.** *The DLRA scheme given in* (6.16) *together with the conservative truncation strategy presented in Section 4.4.2 is locally mass conservative, i.e. it fulfills the local conservation law*

$$\begin{aligned}&\frac{1}{\Delta t}\left(\sqrt{2}B_j^{n+1}\Phi_j^{n+1} + B_j^{n+1} - \left(\sqrt{2}B_j^n\Phi_j^n + B_j^n\right)\right) \\ &= -\sqrt{2}\sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n A_{0\ell} + \sqrt{2}\sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{\ell=0}^{N_\mu-1} v_{i\ell}^n \left|A\right|_{0\ell},\end{aligned} \tag{6.19}$$

where $\Phi_j^n := \sum_{m,\eta=1}^{r} X_{jm}^n S_{m\eta}^n V_{0\eta}^n$ *and* $\Phi_j^{n+1} = \sum_{m,\eta=1}^{r_{n+1}} X_{jm}^{n+1} S_{m\eta}^{n+1} V_{0\eta}^{n+1}$. *As done before,*

we denote $v_{jk}^n = \sum_{m,\eta=1}^{r} X_{jm}^n S_{m\eta}^n V_{k\eta}^n$. *This is a discretization of the continuous local conservation law given in* (6.4).

Proof. The conservative truncation strategy is designed to leave the zeroth order moment unchanged, i.e. it holds $\sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{jm}^{n+1} \widehat{\widehat{S}}_{m\eta}^{n+1} \widehat{\widehat{V}}_{0\eta}^{n+1} = v_{j0}^{n+1}$. In addition, we know from the basis augmentation performed in (6.16i) and the adjustment step stated in (6.16j) that it holds $\sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{jm}^{n+1} \widehat{\widehat{S}}_{m\eta}^{n+1} \widehat{\widehat{V}}_{0\eta}^{n+1} = \sum_{m,\eta=1}^{r_{n+1}} X_{jm}^{n+1} S_{m\eta}^{n+1} V_{0\eta}^{n+1}$. Combining both equalities, we obtain

$$\Phi_j^{n+1} = \sum_{m,\eta=1}^{r_{n+1}} X_{jm}^{n+1} S_{m\eta}^{n+1} V_{0\eta}^{n+1} = \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{jm}^{n+1} \widehat{\widehat{S}}_{m\eta}^{n+1} \widehat{\widehat{V}}_{0\eta}^{n+1} = v_{j0}^{n+1}.$$

We insert this relation into the coupled equations (6.16f) and (6.16g). We multiply (6.16f) with $\sqrt{2}\,(1+\sigma\Delta t)$, rearrange it, and multiply both equations with B_j^n. This leads to

$$\begin{aligned}
\sqrt{2} B_j^{n+1} \Phi_j^{n+1} = \sqrt{2} B_j^n \Phi_j^n &- \sqrt{2}\Delta t \sum_{i=1}^{N_x} D_{ji}^x B_i^n \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{im}^* \widetilde{S}_{m\eta}^n \sum_{\ell=0}^{N_\mu - 1} \widehat{\widehat{V}}_{\ell\eta}^* A_{0\ell} \\
&+ \sqrt{2}\Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} D_{ji}^{xx} B_i^n \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{im}^* \widetilde{S}_{m\eta}^n \sum_{\ell=0}^{N_\mu - 1} \widehat{\widehat{V}}_{\ell\eta}^* |A|_{0\ell} \qquad (6.20a) \\
&+ \sigma \Delta t B_j^{n+1} \left(2 - \sqrt{2}\Phi_j^{n+1}\right), \\
B_j^{n+1} = B_j^n &+ \sigma \Delta t B_j^{n+1} \left(\sqrt{2}\Phi_j^{n+1} - 2\right). \qquad (6.20b)
\end{aligned}$$

Due to the basis augmentations with $\mathbf{X}^n$ and $\mathbf{V}^n$ introduced by the rank-adaptive augmented BUG integrator it can be concluded that

$$\sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}_{im}^* \widetilde{S}_{m\eta}^n \widehat{\widehat{V}}_{\ell\eta}^* = \sum_{m,\eta=1}^{r} X_{im}^n S_{m\eta}^n V_{\ell\eta}^n = v_{i\ell}^n.$$

We insert this relation into expression (6.20a), add equations (6.20a) and (6.20b), and rearrange the result. This leads to the local conservation law (6.19), ensuring the local conservation of mass. □

Hence, equipped with a conservative truncation step, the energy stable DLRA algorithm presented in (6.16) locally conserves mass.

6.6 Numerical results

In this section, we compare the solution of the DLRA scheme (6.16) to the solution of the full equations (6.15). We provide different test examples in 1D which validate our theoretical results. Section 6.6.1 reconsiders the 1D plane source problem, whereas Section 6.6.2 is devoted to the 1D Marshak wave problem with external source. Note that in this

chapter we focus on the theoretical difficulties arising for a multiplicative splitting of the distribution function and on proper theoretical results. For this reason we refrain from higher-dimensional examples but expect the DLRA scheme to equally provide accurate and efficient solutions similar to the 2D result given in Section 5.6.3.

6.6.1 1D plane source

We first examine the 1D plane source test case. This is a common test example for thermal radiative transfer and has already been treated in Section 5.6.1 for the non-multiplicative Su-Olson problem. We consider the spatial domain $\Omega_x = [-10, 10]$ and the angular domain $\Omega_\mu = [-1, 1]$. The initial distribution is chosen to be the cutoff Gaussian

$$v\left(t=0,x\right) = \frac{1}{B^0} \max\left(10^{-4}, \frac{1}{\sqrt{2\pi\sigma_{\mathrm{IC}}^2}} \exp\left(-\frac{\left(x-1\right)^2}{2\sigma_{\mathrm{IC}}^2}\right)\right),$$

with constant deviation $\sigma_{\mathrm{IC}} = 0.03$. The traveling particles are initially centered around $x = 1$ and move into all directions $\mu \in [-1, 1]$. The initial value for the internal energy is set to $B^0 = 1$ and for the opacity to $\sigma = 1$. For the low-rank computations an initial rank of $r = 10$ is prescribed. This value is chosen smaller than in Section 5.6.1 as we are concerned with quantities of rank $4r + 1$. The total mass m^n at time t_n is defined as $m^n := \Delta x \sum_{j=1}^{N_x} \left(\sqrt{2} B_j^n v_{j0}^n + B_j^n\right)$. As computational parameters we use $N_x = 1000$ cells in the spatial and $N_\mu = 500$ moments in the angular variable. The time step size is determined by $\Delta t = C_{\mathrm{CFL}} \cdot \Delta x$ with a CFL number of $C_{\mathrm{CFL}} = 0.99$.
In Figure 6.1 we compare the solution of the DLRA scheme with the solution of the full system. It is observable that the solution $f\left(x,\mu\right)$ as well as the scalar flux $\Phi = \frac{1}{\sqrt{2}} \left\langle f \right\rangle_\mu$ and the dimensionless temperature $T = \sqrt[4]{B}$ at the end time $t_{\mathrm{end}} = 8$ are captured well by the DLRA scheme. For a chosen tolerance parameter of $\vartheta = 10^{-1} \|\mathbf{\Sigma}\|_F$ the rank r increases up to $r = 23$ before it significantly decreases again. The relative mass error $\frac{\left|m^0 - m^n\right|}{\left|m^0\right|}$ is of order $\mathcal{O}\left(10^{-13}\right)$, i.e. the proposed DLRA scheme is mass conservative up to machine precision. These results confirm our theoretical considerations and match the results of the non-multiplicative Su-Olson problem described in Section 5.6.1.

6.6.2 1D external source

In a second example, an external source term $Q\left(x\right)$ is added to the conservative form of the Su-Olson system (6.3), leading to

$$\begin{aligned}
\partial_t g\left(t,x,\mu\right) = &-\frac{\mu}{B\left(t,x\right)} \partial_x \left(B\left(t,x\right) g\left(t,x,\mu\right)\right) + \sigma\left(1 - g\left(t,x,\mu\right)\right) \\
&- \frac{g\left(t,x,\mu\right)}{B\left(t,x\right)} \partial_t B\left(t,x\right) + \frac{Q\left(x\right)}{B\left(t,x\right)}, \\
\partial_t B\left(t,x\right) = &\sigma B\left(t,x\right) \left(\left\langle g\left(t,x,\mu\right)\right\rangle_\mu - 2\right).
\end{aligned}$$

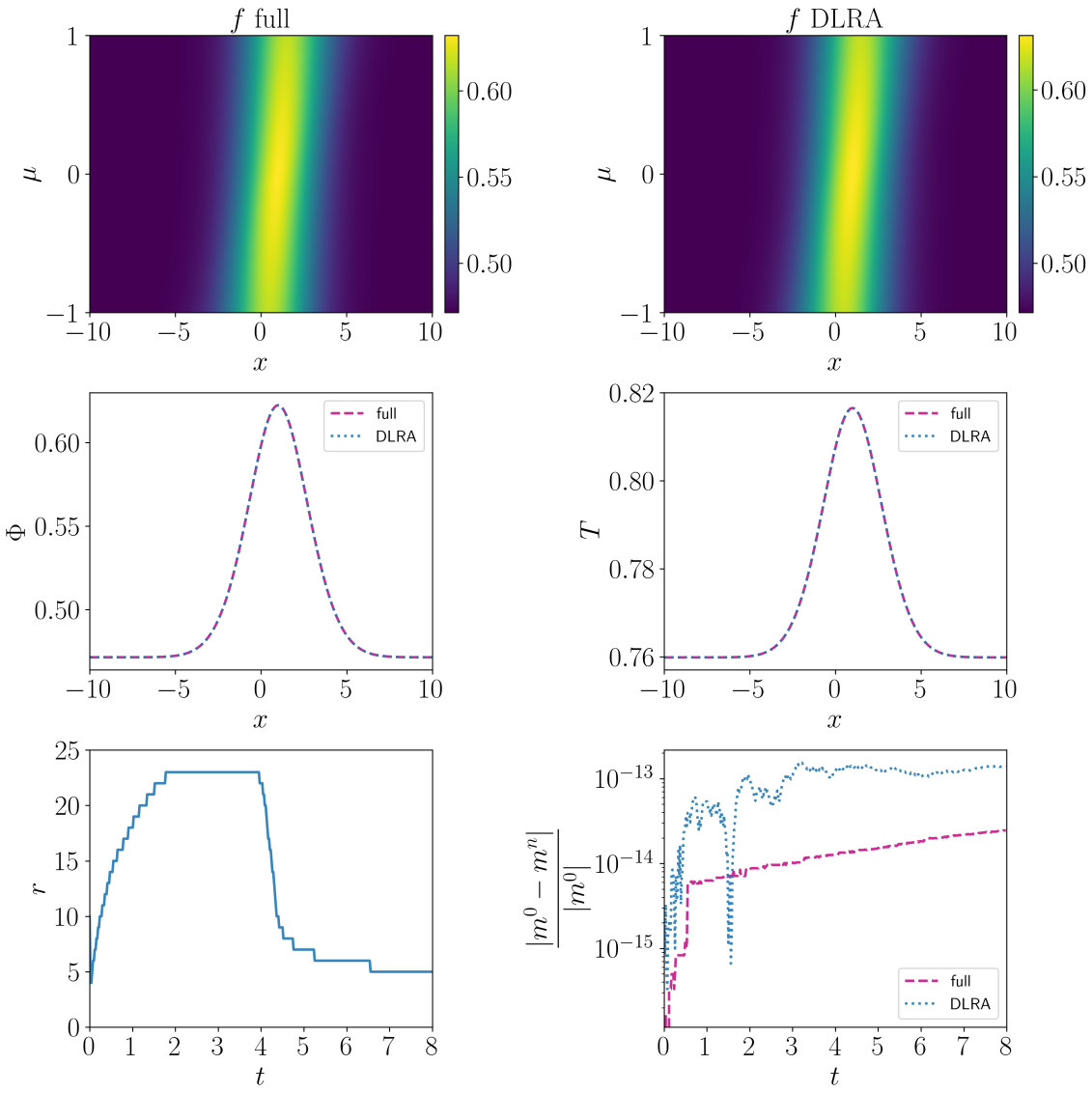


Figure 6.1: Top row: Numerical results for the solution $f(x, \mu)$ of the plane source problem at time $t_{\text{end}} = 8$ computed with the multiplicative full solver (left) and the multiplicative DLRA scheme (right). **Middle row:** Scalar flux Φ (left) and temperature T (right) for both the multiplicative full solver and the multiplicative DLRA scheme. **Bottom row:** Evolution of the rank in time for the multiplicative DLRA method (left) and evolution of the relative mass error in time compared for both methods (right).

This test example is known as the *Marshak wave problem* [Mar58] and has already been considered in Section 5.6.2 for the non-multiplicative Su-Olson problem. In our example, we use the source function $Q(x) = \chi_{[-0.5,0.5]}(x)/a$ with $a = \frac{4\sigma_{\text{SB}}}{c}$ being the radiation constant and $\chi_{[-0.5,0.5]}(x)$ denoting the indicator function on $[-0.5, 0.5]$. The initial value for the internal energy is set to $B^0 = 50$. All other initial settings and computational parameters remain unchanged from the previous test example given in Section 6.6.1.

In Figure 6.2 we compare the solution of the full equations (6.15) to the solution obtained from the DLRA scheme presented in (6.16). Both schemes are adjusted to take the

additional source term into account. The numerical results for the solution $f(x, \mu)$, for the scalar flux $\Phi = \frac{1}{\sqrt{2}} \langle f \rangle_\mu$ and for the dimensionless temperature $T = \sqrt[4]{B}$ at the end time $t_{\text{end}} = 3.16$, computed with both solvers, are shown. We again observe that the DLRA scheme captures the solution of the full system. For a chosen tolerance parameter of $\vartheta = 10^{-3}\|\mathbf{\Sigma}\|_F$ the rank r increases up to a value of $r = 23$. Due to the additional source term, there is no conservation of mass in this test example. These results confirm our theoretical considerations and match the results of the non-multiplicative Su-Olson problem described in Section 5.6.2. However, note that for an accurate solution of the DLRA scheme a smaller truncation tolerance parameter ϑ as well as a higher rank r are required, indicating that the multiplicative structure numerically poses additional challenges.

6.7 Summary and conclusion

We have presented a DLRA discretization for the multiplicative Su-Olson problem that is energy stable and mass conservative. The main research contributions are:

(i) *A multiplicative splitting of the distribution function:* Based on the insights gained in [EHY21] we have considered a multiplicative splitting of the distribution function for which the spatial discretization had to be carefully derived. Further, the multiplicative splitting has required additional modifications in the DLRA scheme in order to obtain an energy stable numerical discretization of the problem.

(ii) *An energy stable numerical scheme with rigorous mathematical proofs:* We have given rigorous mathematical proofs for the energy stability of the derived DLRA scheme, enabling to deduce a classic hyperbolic CFL condition. This allows to compute up to a maximal time step size of $\Delta t = C_{\text{CFL}} \cdot \Delta x$, enhancing the performance of the algorithm.

(iii) *A mass conservative and rank-adaptive augmented integrator:* We have implemented the rank-adaptive augmented BUG integrator presented in [CKL22]. Since this integrator allows for further basis modifications, we have included additional basis augmentation steps that ensure the exactness of the projection operators needed for the theoretical proof of energy stability as well as the local conservation of mass, which has been guaranteed in combination with a suitable truncation strategy as described in [EOS23, EKS23].

(iv) *Numerical test examples confirming the theoretical properties:* We have compared the numerical results obtained from the DLRA scheme with the solution of the full system for relevant test examples from the literature, validating the derived properties as well as the accuracy of the proposed DLRA method.

However, the extension of the considered stability analysis from a linear to a non-linear problem, for example the isothermal Boltzmann-BGK equation treated in [EHY21], poses

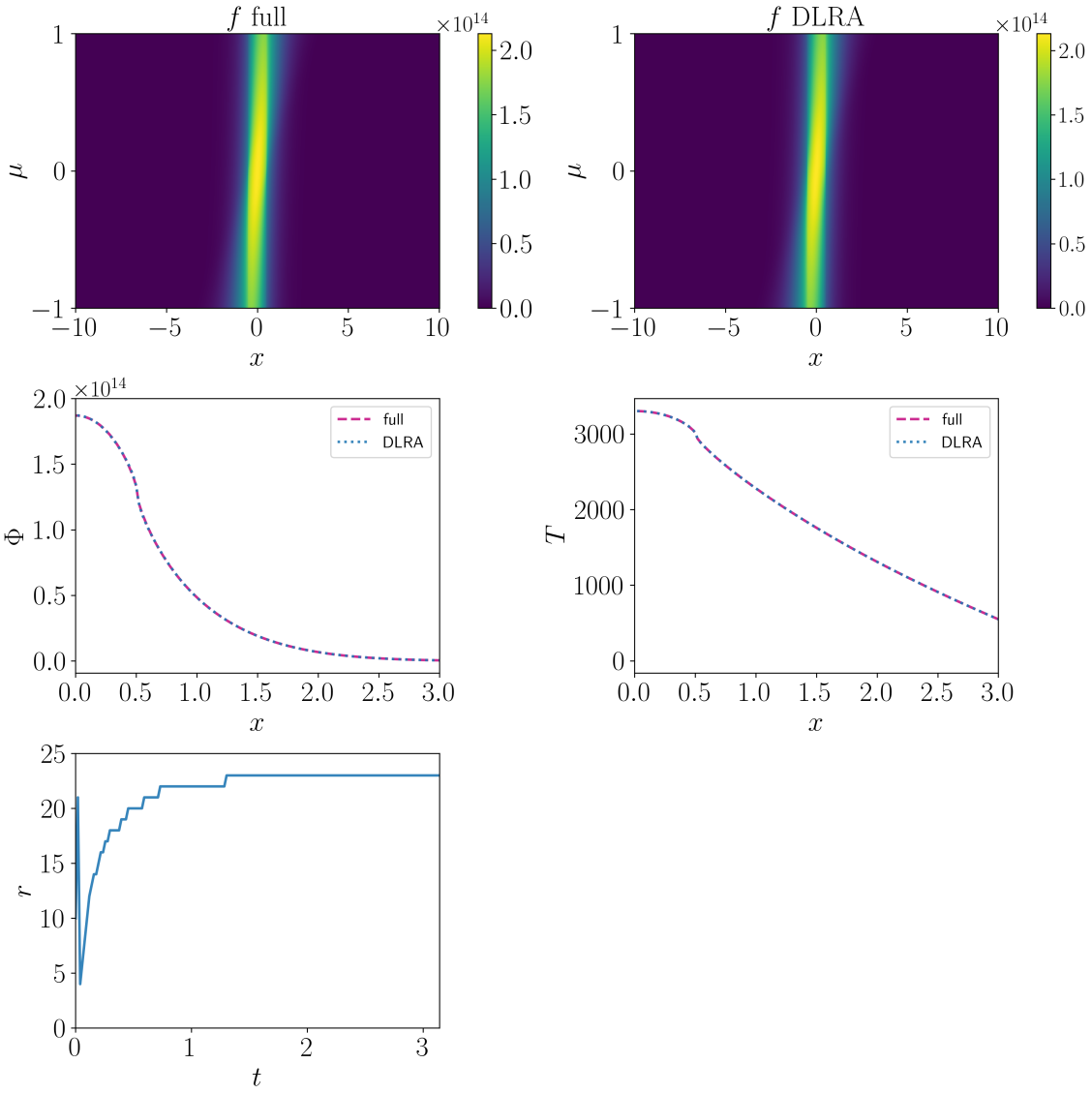


Figure 6.2: Top row: Numerical results for the solution $f(x, \mu)$ of the external source problem at time $t_{\text{end}} = 3.16$ computed with the multiplicative full solver (left) and the multiplicative DLRA scheme (right). **Middle row:** Scalar flux Φ (left) and temperature T (right) for both the multiplicative full system and the multiplicative DLRA scheme. **Bottom row:** Evolution of the rank in time for the multiplicative DLRA method.

additional challenges as the general theoretical setting is significantly more difficult. Nevertheless, the analysis performed for the multiplicative Su-Olson problem provides valuable insights into the choice of a suitable spatial discretization and stabilization when considering a multiplicative splitting of the distribution function. This splitting approach can be extremely useful for the construction of DLRA schemes for more complicated problems.

7

A multiplicative DLRA scheme for the linear Boltzmann-BGK equation

The Boltzmann equation is a fundamental model in kinetic theory describing a gas that is not in thermodynamic equilibrium. In its full formulation with quadratic Boltzmann collision operator as given in (2.10), numerically solving the Boltzmann equation is highly demanding. Instead, the Boltzmann-Bhatnagar-Gross-Krook (BGK) equation (2.11) can be considered. It simplifies the collision term while maintaining the key properties of the original equation. Still, especially in higher-dimensional settings occurring in practical applications, its solution can lead to prohibitive numerical costs. To overcome this last problem, the method of DLRA is applied to the Boltzmann-BGK equation in this chapter. Inspired by [EHY21, KS16], a multiplicative splitting of the distribution function of the form $f = Mg$ is considered, splitting a generally not low-rank Maxwellian M from a remaining distribution function g. In [EHY21], it has been shown that for the Boltzmann-BGK equation the remaining function g is of low rank even if the distribution function f is not (which is not true for the classic additive micro-macro decomposition). Hence, in order to obtain an efficient scheme, the DLRA approach is applied to the low-rank distribution function g. Difficulties may arise in the discretization. With the knowledge gained in Chapter 6 an advection and a conservative form of the evolution equation for g are derived and a "first discretize, then low-rank" approach is pursued. Further, the potentially stiff collision term is treated with an implicit temporal discretization. However, different from Chapter 6, the Boltzmann-BGK equation requires another notion of stability, giving rise to additional complexities in the proof of numerical stability. In addition, for the construction of the DLRA scheme new ideas for the basis augmentations as well as an adjusted truncation strategy are necessary.

The structure of this chapter is as follows. In Section 7.1 two possible systems for the linear Boltzmann-BGK equation with multiplicative splitting are derived. Both systems are equivalent in the continuous setting. In Section 7.2 a discretization in velocity, space and time is performed, leading to two different fully discretized schemes. It is shown in Section 7.3 that the advection form of the multiplicative linear Boltzmann-BGK equation can lead to a numerical scheme that is not von Neumann stable, whereas for the conservative form

numerical stability can be guaranteed. Section 7.4 is devoted to the derivation of a DLRA scheme which together with a suitable truncation strategy is shown to be numerically stable. Numerical experiments both in 1D and 2D, given in Section 7.5, confirm the derived results before Section 7.6 provides a brief summary and conclusion. The results of this chapter closely follow the presentation in [BEKK24b].

7.1 Linear Boltzmann-BGK equation with multiplicative splitting

We start from the Boltzmann-BGK equation given in (2.11) and restrict it to a 1D setting of the form

$$\partial_t f(t,x,v) + v\partial_x f(t,x,v) = \sigma\left(M[f](t,x,v) - f(t,x,v)\right), \tag{7.1a}$$

where $f(t,x,v)$ denotes the distribution function depending on the time $t \in \mathbb{R}^+$, the spatial variable $x \in \Omega_x \subseteq \mathbb{R}$ and the velocity variable $v \in \mathbb{R}$. The collision frequency of the particles is set to a constant scalar value σ. In the definition of the Maxwellian equilibrium distribution $M[f]$ as provided in (2.8), the number density $n(t,x)$ can be replaced by the mass density $\rho(t,x)$, which under the assumption of a unity mass m are the same. We refer to this quantity as the *density* $\rho(t,x)$. An evolution equation for the density is obtained by integrating (7.1a) with respect to v, resulting in

$$\partial_t \rho(t,x) = -\partial_x \int v f(t,x,v)\,\mathrm{d}v. \tag{7.1b}$$

Following the considerations presented in [EHY21], we employ the multiplicative decomposition

$$f(t,x,v) = M[f](t,x,v)\, g(t,x,v), \tag{7.2}$$

which is advantageous for the construction of an efficient DLRA scheme as g is low-rank even if this is not the case for the Maxwellian. In this thesis, we consider an isothermal Maxwellian without drift, i.e.

$$M[f](t,x,v) = \frac{\rho(t,x)}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right).$$

This results in a linear model, which we call the *linear Boltzmann-BGK equation*. This linear model has been extensively studied in the PDE community [Eva21, CCEY20, AAC16] as well as from a numerical point of view [BCHR20]. In the following considerations, we provide a stability analysis in the context of DLRA simulation using a multiplicative decomposition as proposed in (7.2). The stability analysis for the simplified problem provides insight into the numerical scheme that has been used in the literature [EHY21] dealing with the Boltzmann-BGK equation. In particular, our analysis explains why

such multiplicative schemes need to take relatively small time step sizes even though the collision operator is treated implicitly.
We first insert the multiplicative approach (7.2) into the definition of the density and obtain

$$\rho(t,x) = \frac{\rho(t,x)}{\sqrt{2\pi}} \int g(t,x,v)\, e^{-v^2/2} \mathrm{d}v,$$

which can be equivalently rewritten as the identity

$$1 = \frac{1}{\sqrt{2\pi}} \int g(t,x,v)\, e^{-v^2/2} \mathrm{d}v. \tag{7.3}$$

Then we insert the multiplicative approach (7.2) into equations (7.1a) and (7.1b), yielding

$$\begin{aligned} \partial_t g(t,x,v) =& -v\partial_x g(t,x,v) + \sigma(1 - g(t,x,v)) - \frac{g(t,x,v)}{\rho(t,x)} \partial_t \rho(t,x) \\ & - v \frac{g(t,x,v)}{\rho(t,x)} \partial_x \rho(t,x), \end{aligned} \tag{7.4a}$$

$$\partial_t \rho(t,x) = -\frac{1}{\sqrt{2\pi}} \partial_x \int \rho(t,x)\, g(t,x,v)\, v e^{-v^2/2} \mathrm{d}v. \tag{7.4b}$$

This set of equations is called the *advection form* of the multiplicative system. It corresponds to the way the equations are treated in [EHY21]. We can rewrite equation (7.4a) into a *conservative form*, leading to the system

$$\begin{aligned} \partial_t g(t,x,v) =& -\frac{v}{\rho(t,x)} \partial_x (\rho(t,x)\, g(t,x,v)) + \sigma(1 - g(t,x,v)) \\ & - \frac{g(t,x,v)}{\rho(t,x)} \partial_t \rho(t,x), \end{aligned} \tag{7.5a}$$

$$\partial_t \rho(t,x) = -\frac{1}{\sqrt{2\pi}} \partial_x \int \rho(t,x)\, g(t,x,v)\, v e^{-v^2/2} \mathrm{d}v. \tag{7.5b}$$

Note that for both systems we omit initial and boundary conditions for now. In the following considerations, we discretize both sets of equations (7.4) and (7.5) to compare them in terms of numerical stability. We derive a stable DLRA scheme and give a concrete hyperbolic CFL condition. Similar to Chapter 6, we first discretize the equations and then apply a DLRA approach.

7.2 Discretization of the multiplicative system

In this section, we provide a full discretization of both versions (7.4) and (7.5) of the multiplicative system. Sections 7.2.1 and 7.2.2 discretize equations (7.4) and (7.5) in velocity and space, leading to semi-discrete systems. In Section 7.2.3 a temporal discretization is presented, rendering fully discrete schemes.

7.2.1 Velocity discretization

For the discretization in the velocity space a nodal approach as described in Section 3.3.1 is employed. We prescribe a certain number of grid points N_v and determine the quadrature nodes $v_1, ..., v_{N_v}$ and weights $\omega_1, ..., \omega_{N_v}$ using the Gauss-Hermite quadrature rule. This choice accounts for the special structure of equations (7.4b) and (7.5b) and enables an approximation of the following integrals as

$$\int_{\mathbb{R}} e^{-v^2} g(t,x,v)\,\mathrm{d}v \approx \sum_{k=1}^{N_v} \omega_k g(t,x,v_k).$$

An approximation of the velocity-dependent distribution function $g(t,x,v)$ is obtained from an evaluation at each grid point, i.e. by computing

$$g_k(t,x) \approx g(t,x,v_k) \qquad \text{for} \quad k = 1, ..., N_v.$$

Considering the advection form (7.4), this leads to the system

$$\begin{aligned} \partial_t g_k(t,x) =& -v_k \partial_x g_k(t,x) + \sigma(1 - g_k(t,x)) - \frac{g_k(t,x)}{\rho(t,x)} \partial_t \rho(t,x) \\ & - v_k \frac{g_k(t,x)}{\rho(t,x)} \partial_x \rho(t,x), \end{aligned} \tag{7.6a}$$

$$\partial_t \rho(t,x) = -\frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} \partial_x (\rho(t,x) g_k(t,x)) v_k \omega_k e^{v_k^2/2}, \tag{7.6b}$$

which is discretized in the velocity variable. Analogously, for the conservative system (7.5) the following set of equations is derived:

$$\partial_t g_k(t,x) = -\frac{v_k}{\rho(t,x)} \partial_x (\rho(t,x) g_k(t,x)) + \sigma(1 - g_k(t,x)) - \frac{g_k(t,x)}{\rho(t,x)} \partial_t \rho(t,x), \tag{7.7a}$$

$$\partial_t \rho(t,x) = -\frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} \partial_x (\rho(t,x) g_k(t,x)) v_k \omega_k e^{v_k^2/2}. \tag{7.7b}$$

7.2.2 Spatial discretization

Regarding the discretization of the spatial domain Ω_x, we construct a uniform spatial grid with N_x grid cells and equidistant spacing $\Delta x = \frac{1}{N_x}$. Spatially dependent quantities are approximated as

$$\rho_j(t) \approx \rho(t, x_j) \qquad \text{and} \qquad g_{jk}(t) \approx g_k(t, x_j) \qquad \text{for} \quad j = 1..., N_x.$$

Assuming periodic boundary conditions, first-order spatial derivatives ∂_x are approximated using the centered FD method. For stability reasons, a diffusion term involving second-order derivatives ∂_{xx} is added. This term is also approximated by the centered FD method. We employ the tridiagonal spatial stencil matrices $\mathbf{D}^x \in \mathbb{R}^{N_x \times N_x}$ given in

(3.8) and $\mathbf{D}^{xx} \in \mathbb{R}^{N_x \times N_x}$ defined in (3.11).

We insert the proposed spatial discretizations into the advection form (7.6) and add a stabilizing second-order term for $\partial_x g$. This corresponds to the method used in [EHY21] for the non-linear isothermal Boltzmann-BGK equation and leads to the semi-discrete time-continuous system

$$\begin{aligned}\dot{g}_{jk}(t) = &-\sum_{i=1}^{N_x} D^x_{ji} g_{ik}(t)\, v_k + \frac{\Delta x}{2}\sum_{i=1}^{N_x} D^{xx}_{ji} g_{ik}(t)\,|v_k| \\ &+ \sigma\left(1 - g_{jk}(t)\right) - \frac{g_{jk}(t)}{\rho_j(t)}\dot{\rho}_j(t) - \frac{g_{jk}(t)}{\rho_j(t)}\sum_{i=1}^{N_x} D^x_{ji}\rho_i(t)\, v_k,\end{aligned} \tag{7.8a}$$

$$\begin{aligned}\dot{\rho}_j(t) = &-\frac{1}{\sqrt{2\pi}}\sum_{i=1}^{N_x}\sum_{k=1}^{N_v} D^x_{ji}\rho_i(t)\, g_{ik}(t)\, v_k \omega_k e^{v_k^2/2} \\ &+ \frac{\Delta x}{2\sqrt{2\pi}}\sum_{i=1}^{N_x}\sum_{k=1}^{N_v} D^{xx}_{ji}\rho_i(t)\, g_{ik}(t)\,|v_k|\,\omega_k e^{v_k^2/2}.\end{aligned} \tag{7.8b}$$

For the conservative form (7.7) the second-order stabilization term is applied to $\partial_x(\rho g)$. We obtain the semi-discrete time-continuous system

$$\begin{aligned}\dot{g}_{jk}(t) = &-\frac{1}{\rho_j(t)}\sum_{i=1}^{N_x} D^x_{ji}\rho_i(t)\, g_{ik}(t)\, v_k + \frac{\Delta x}{2}\frac{1}{\rho_j(t)}\sum_{i=1}^{N_x} D^{xx}_{ji}\rho_i(t)\, g_{ik}(t)\,|v_k| \\ &+ \sigma\left(1 - g_{jk}(t)\right) - \frac{g_{jk}(t)}{\rho_j(t)}\dot{\rho}_j(t),\end{aligned} \tag{7.9a}$$

$$\begin{aligned}\dot{\rho}_j(t) = &-\frac{1}{\sqrt{2\pi}}\sum_{i=1}^{N_x}\sum_{k=1}^{N_v} D^x_{ji}\rho_i(t)\, g_{ik}(t)\, v_k \omega_k e^{v_k^2/2} \\ &+ \frac{\Delta x}{2\sqrt{2\pi}}\sum_{i=1}^{N_x}\sum_{k=1}^{N_v} D^{xx}_{ji}\rho_i(t)\, g_{ik}(t)\,|v_k|\omega_k e^{v_k^2/2}.\end{aligned} \tag{7.9b}$$

Note that due to the different structure of the equations the stabilization term in (7.8a) is applied to $\partial_x g$, whereas in (7.9a) it is added for $\partial_x(\rho g)$. This marks an important difference between both presented schemes.

7.2.3 Temporal discretization

The temporal discretization has to be carefully derived to obtain a possibly numerically stable scheme. We start with the semi-discrete advection form presented in (7.8) and perform an explicit Euler step for the transport part in (7.8a) as well as in (7.8b). The potentially stiff collision term is treated implicitly. This is a reasonable approach and for instance explained in Section 3.1.2. For approximating the time derivative $\partial_t \rho$ the

corresponding difference quotient is used. We obtain the fully discrete scheme

$$g_{jk}^{n+1} = g_{jk}^n - \Delta t \sum_{i=1}^{N_x} D_{ji}^x g_{ik}^n v_k + \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} D_{ji}^{xx} g_{ik}^n |v_k| \tag{7.10a}$$
$$+ \sigma \Delta t \left(1 - g_{jk}^{n+1}\right) - \frac{g_{jk}^{n+1}}{\rho_j^n} \left(\rho_j^{n+1} - \rho_j^n\right) - \Delta t \frac{g_{jk}^n}{\rho_j^n} \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n v_k,$$

$$\rho_j^{n+1} = \rho_j^n - \Delta t \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{N_x} \sum_{k=1}^{N_v} D_{ji}^x \rho_i^n g_{ik}^n v_k \omega_k e^{v_k^2/2} \tag{7.10b}$$
$$+ \Delta t \frac{\Delta x}{2\sqrt{2\pi}} \sum_{i=1}^{N_x} \sum_{k=1}^{N_v} D_{ji}^{xx} \rho_i^n g_{ik}^n |v_k| \omega_k e^{v_k^2/2},$$

which describes one time step from time t_n to time $t_{n+1} = t_n + \Delta t$. Considering the conservative form (7.9), we again perform an explicit Euler step for the transport part in (7.9a) as well as in (7.9b). The collision term is treated implicitly and a factor $\frac{\rho_j^{n+1}}{\rho_j^n}$ is added. The special form of this factor will be explained later in the proof of numerical stability. As done before, the time derivative $\partial_t \rho$ is approximated by its difference quotient. This leads to the fully discretized equations

$$g_{jk}^{n+1} = g_{jk}^n - \Delta t \frac{1}{\rho_j^n} \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n g_{ik}^n v_k + \Delta t \frac{\Delta x}{2} \frac{1}{\rho_j^n} \sum_{i=1}^{N_x} D_{ji}^{xx} \rho_i^n g_{ik}^n |v_k| \tag{7.11a}$$
$$+ \sigma \Delta t \frac{\rho_j^{n+1}}{\rho_j^n} \left(1 - g_{jk}^{n+1}\right) - \frac{g_{jk}^{n+1}}{\rho_j^n} \left(\rho_j^{n+1} - \rho_j^n\right),$$

$$\rho_j^{n+1} = \rho_j^n - \Delta t \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{N_x} \sum_{k=1}^{N_v} D_{ji}^x \rho_i^n g_{ik}^n v_k \omega_k e^{v_k^2/2} \tag{7.11b}$$
$$+ \Delta t \frac{\Delta x}{2\sqrt{2\pi}} \sum_{i=1}^{N_x} \sum_{k=1}^{N_v} D_{ji}^{xx} \rho_i^n g_{ik}^n |v_k| \omega_k e^{v_k^2/2}.$$

Note that the discretizations for ρ given in (7.10b) and (7.11b) are exactly the same. The main differences between the naive discretization of the advection form (7.10) and the proposed scheme (7.11) for the conservative form are the stabilization of $\partial_x (\rho g)$ in (7.11a), opposed to a stabilization of $\partial_x g$ as done in (7.10a), and the additional factor $\frac{\rho_j^{n+1}}{\rho_j^n}$ in the collision term of (7.11a).

7.3 Numerical stability

Although the derivation of the equations proposed in (7.10) and (7.11) is similar, both systems differ drastically in terms of numerical stability. We first introduce the following

notations for the fully discrete setting.

Definition 7.1 (Fully discrete solution and Maxwellian)**.** The *fully discrete solution* f *at time* t_n is given by $\mathbf{f}^n = \left(f_{jk}^n\right) \in \mathbb{R}^{N_x \times N_v}$ with entries

$$f_{jk}^n = \frac{1}{\sqrt{2\pi}} \rho_j^n g_{jk}^n e^{-v_k^2/2}.$$

The *fully discrete Maxwellian at time* t_n is denoted by $\mathbf{M}^n = \left(M_{jk}^n\right) \in \mathbb{R}^{N_x \times N_v}$ with entries

$$M_{jk}^n = \frac{1}{\sqrt{2\pi}} \rho_j^n e^{-v_k^2/2}.$$

In this section, both fully discrete schemes presented are compared. In Section 7.3.1 it is shown that the advection form (7.10) is generally not von Neumann stable whereas for the conservative form (7.11) a proof of numerical stability is established in Section 7.3.2.

7.3.1 Advection form

We begin with the fully discretized advection form (7.10), which is comparable to the discretization chosen in the article [EHY21] as the term $\partial_x (Mg)$ is split up into the sum of $M\partial_x g$ and $g\partial_x M$. In [EHY21], numerical experiments are given but no explicit stability analysis is conducted. In the following part, we provide an example showing that numerical stability in the sense of von Neumann cannot be guaranteed.

Theorem 7.2. *There exist initial values* $\mathbf{g}^n = \left(g_{jk}^n\right) \in \mathbb{R}^{N_x \times N_v}$ *and* $\boldsymbol{\rho}^n = \left(\rho_j^n\right) \in \mathbb{R}^{N_x}$ *such that the advection form* (7.10) *of the linear Boltzmann-BGK equation for* $\sigma = 0$ *is not von Neumann stable.*

Proof. Let us assume a solution g_{jk}^n that is constant in space and velocity, e.g. $g_{jk}^n \equiv 1$. For this solution the terms containing $\mathbf{D}^x\mathbf{g}^n$ and $\mathbf{D}^{xx}\mathbf{g}^n$ in (7.10a) are zero. Let us further assume that there is no collisionality, i.e. $\sigma = 0$. We insert this information into (7.10a) and derive

$$g_{jk}^{n+1} = 1 - \frac{g_{jk}^{n+1}}{\rho_j^n}\left(\rho_j^{n+1} - \rho_j^n\right) - \Delta t \frac{1}{\rho_j^n} \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n v_k.$$

After rearranging the equation, we obtain

$$\rho_j^{n+1} g_{jk}^{n+1} = \rho_j^n - \Delta t \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n v_k.$$

Multiplication with $\frac{1}{\sqrt{2\pi}} e^{-v_k^2/2}$ leads to

$$f_{jk}^{n+1} = f_{jk}^n - \Delta t \sum_{i=1}^{N_x} D_{ji}^x f_{ik}^n v_k. \tag{7.12}$$

This expression corresponds to a discretization of the linear advection equation of the form $\partial_t f + v\partial_x f = 0$ with an explicit Euler step forward in time and a centered FD method in space. According to Remark 3.16 this discretization is not von Neumann stable. □

Indeed, it can be shown that the discretization given in (7.12) is not von Neumann stable but stable in the sense of Definition 3.8 for relatively small time step sizes [LeV07]. This matches our numerical insights gained from [EHY21], where the spatial discretization is comparable to (7.10) and small time step sizes are required.

7.3.2 Conservative form

Having found out that for a certain choice of the initial values the system of equations (7.10) is not von Neumann stable, we now consider equations (7.11) in terms of numerical stability. We observe that the advection terms are treated explicitly, whereas the collision term is treated implicitly. As explained in Section 3.1.2, this leads to a removal of the potential stiffness caused by a large number of collisions. We seek a rigorous proof of stability under a classic hyperbolic CFL condition, which will be derived in the following norm.

Definition 7.3 (Stability norm)**.** For $\mathbf{f}^n = (f_{jk}^n) \in \mathbb{R}^{N_x \times N_v}$, the *$\mathscr{H}$-norm* is defined as

$$\|\mathbf{f}^n\|_{\mathscr{H}}^2 = \sqrt{2\pi} \sum_{j=1}^{N_x} \sum_{k=1}^{N_v} \left(f_{jk}^n\right)^2 \omega_k e^{3v_k^2/2}.$$

This corresponds to a Frobenius norm $\|\cdot\|_F$ with weights $\sqrt{2\pi}\omega_k e^{3v_k^2/2}$.

The choice of this norm is inspired by the analysis in [AAC16], where hypocoercivity for the linear Boltzmann-BGK equation is shown. Different from the considerations in [AAC16], we use a fully discrete analogue to the considered weighted L^2-norm which also takes the Gauss-Hermite quadrature into account. Note that the factor $\sqrt{2\pi}$ does not affect the stability but is added for consistency.
At each time step the fully discrete distribution function f and the fully discrete density ρ are required to fulfill the discrete counterpart of its definition given in Definition 2.5, namely the identity

$$\rho_j^n = \sum_{k=1}^{N_v} f_{jk}^n \omega_k e^{v_k^2} \qquad \text{for all} \quad n \in \mathbb{N}.$$

With relation (7.3) this identity can be rewritten in a equivalent formulation as

$$1 = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} g_{jk}^n \omega_k e^{v_k^2/2} \qquad \text{for all} \quad n \in \mathbb{N}. \tag{7.13}$$

We are able to show that the equality given in (7.13) holds for the conservative equations (7.11) under a suitable choice of the initial condition.

Lemma 7.4. *Let us assume that the initial condition for g satisfies*

$$1 = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} g_{jk}^0 \omega_k e^{v_k^2/2} \qquad \text{for all } \; j \in \{1, ..., N_x\}\,.$$

Then, for all $n \in \mathbb{N}$, the equality given in (7.13) holds.

Proof. The proof follows by induction. For the induction assumption let us assume that the relation $1 = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} g_{jk}^n \omega_k e^{v_k^2/2}$ holds for one $n \in \mathbb{N}$. For the induction step we begin with equation (7.11a), put the terms containing g_{jk}^{n+1} to the left-hand side and multiply with ρ_j^{n+1}. This results in

$$(1 + \sigma\Delta t)\, \rho_j^{n+1} g_{jk}^{n+1} = \rho_j^n g_{jk}^n - \Delta t \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n g_{ik}^n v_k + \Delta t \frac{\Delta x}{2} \sum_{i=1}^{N_x} D_{ji}^{xx} \rho_i^n g_{ik}^n |v_k| + \sigma\Delta t \rho_j^{n+1}.$$

Multiplication with $\frac{1}{\sqrt{2\pi}} \omega_k e^{v_k^2/2}$ and summation over k leads to

$$\begin{aligned}
&(1 + \sigma\Delta t)\, \rho_j^{n+1} \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} g_{jk}^{n+1} \omega_k e^{v_k^2/2} \\
&= \frac{\rho_j^n}{\sqrt{2\pi}} \sum_{k=1}^{N_v} g_{jk}^n \omega_k e^{v_k^2/2} - \Delta t \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{N_x} \sum_{k=1}^{N_v} D_{ji}^x \rho_i^n g_{ik}^n v_k \omega_k e^{v_k^2/2} \\
&\quad + \Delta t \frac{\Delta x}{2\sqrt{2\pi}} \sum_{i=1}^{N_x} \sum_{k=1}^{N_v} D_{ji}^{xx} \rho_i^n g_{ik}^n |v_k| \omega_k e^{v_k^2/2} + \sigma\Delta t \rho_j^{n+1} \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} \omega_k e^{v_k^2/2}.
\end{aligned}$$

We insert the induction assumption as well as $\frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} \omega_k e^{v_k^2/2} = 1$. Then, together with equation (7.11b), this establishes

$$(1 + \sigma\Delta t)\, \rho_j^{n+1} \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{N_v} g_{jk}^{n+1} \omega_k e^{v_k^2/2} = (1 + \sigma\Delta t) \rho_j^{n+1}.$$

Canceling with $(1 + \sigma\Delta t)\, \rho_j^{n+1}$ gives the desired equality for $n + 1$, and completes the proof. □

Also the following inequality is indispensable to show numerical stability of the conservative system (7.11).

Lemma 7.5. *Under the time step restriction* $\max_k (|v_k|)\, \Delta t \leq \Delta x$ *it holds*

$$\begin{aligned}
&\Delta t \left\| \mathbf{D}^x \mathbf{f}^{n+1} \operatorname{diag}(v_k) - \frac{\Delta x}{2} \mathbf{D}^{xx} \mathbf{f}^{n+1} \operatorname{diag}(|v_k|) \right\|_{\mathscr{H}}^2 \\
&- \Delta x \left\| \mathbf{D}^+ \mathbf{f}^{n+1} \operatorname{diag}\left(|v_k|^{1/2} \right) \right\|_{\mathscr{H}}^2 \leq 0.
\end{aligned} \tag{7.14}$$

Proof. We employ a Fourier analysis similar to [KEC23] and use Lemma 5.7 introducing

the matrices $\mathbf{E}$ and $\mathbf{\Lambda}^\gamma$ that diagonalize the stencil matrices according to

$$\mathbf{D}^\gamma \mathbf{E} = \mathbf{E}\mathbf{\Lambda}^\gamma \qquad \text{with} \quad \gamma \in \{x, xx, +\},$$

where $\mathbf{E}$ are unitary and $\mathbf{\Lambda}^\gamma$ are diagonal matrices. Let us denote $\widehat{\mathbf{f}}^{n+1} = \left(\widehat{f}_{\alpha k}^{n+1}\right) \in \mathbb{C}^{N_x \times N_v}$ with entries $\widehat{f}_{\alpha k}^{n+1} = \sum_{j=1}^{N_x} E_{\alpha j} f_{jk}^{n+1}$. With Parseval's identity given in Proposition 3.14 we obtain

$$\begin{aligned}
&\Delta t \left\| \mathbf{D}^x \mathbf{f}^{n+1} \operatorname{diag}(v_k) - \frac{\Delta x}{2} \mathbf{D}^{xx} \mathbf{f}^{n+1} \operatorname{diag}(|v_k|) \right\|_{\mathscr{H}}^2 - \Delta x \left\| \mathbf{D}^+ \mathbf{f}^{n+1} \operatorname{diag}\left(|v_k|^{1/2}\right) \right\|_{\mathscr{H}}^2 \\
&= \Delta t \left\| \mathbf{D}^x \mathbf{f}^{n+1} \operatorname{diag}\left(v_k \omega_k^{1/2} e^{3v_k^2/4}\right) - \frac{\Delta x}{2} \mathbf{D}^{xx} \mathbf{f}^{n+1} \operatorname{diag}\left(|v_k| \omega_k^{1/2} e^{3v_k^2/4}\right) \right\|_F^2 \\
&\quad - \Delta x \left\| \mathbf{D}^+ \mathbf{f}^{n+1} \operatorname{diag}\left(|v_k|^{1/2} \omega_k^{1/2} e^{3v_k^2/4}\right) \right\|_F^2 \\
&= \Delta t \left\| \mathbf{\Lambda}^x \widehat{\mathbf{f}}^{n+1} \operatorname{diag}\left(v_k \omega_k^{1/2} e^{3v_k^2/4}\right) - \frac{\Delta x}{2} \mathbf{\Lambda}^{xx} \widehat{\mathbf{f}}^{n+1} \operatorname{diag}\left(|v_k| \omega_k^{1/2} e^{3v_k^2/4})\right) \right\|_F^2 \\
&\quad - \Delta x \left\| \mathbf{\Lambda}^+ \widehat{\mathbf{f}}^{n+1} \operatorname{diag}\left(|v_k|^{1/2} \omega_k^{1/2} e^{3v_k^2/4}\right) \right\|_F^2 \\
&= 2 \sum_{\alpha=1}^{N_x} \sum_{k=1}^{N_v} \left(\Delta t \frac{|v_k|^2}{(\Delta x)^2} |1 - \cos(\nu_\alpha)| - \frac{|v_k|}{\Delta x} |1 - \cos(\nu_\alpha)| \right) \omega_k e^{3v_k^2/2} \left| \widehat{f}_{\alpha k}^{n+1} \right|^2 .
\end{aligned}$$

A sufficient condition to ensure negativity is that for each index k it must hold

$$\Delta t \frac{|v_k|^2}{(\Delta x)^2} |1 - \cos(\nu_\alpha)| \leq \frac{|v_k|}{\Delta x} |1 - \cos(\nu_\alpha)| .$$

Hence, for $\max_k(|v_k|)\Delta t \leq \Delta x$, equation (7.14) holds and we have proven the lemma. □

Using the above results, numerical stability of the conservative form of the equations proposed (7.11) in the $\mathscr{H}$-norm can be shown.

Theorem 7.6 (Energy stability of the fully discrete system)**.** *Under the time step restriction* $\max_k(|v_k|)\,\Delta t \leq \Delta x$ *the fully discrete system* (7.11) *is numerically stable in the* $\mathscr{H}$*-norm, i.e. it holds*

$$\left\| \mathbf{f}^{n+1} \right\|_{\mathscr{H}}^2 \leq \left\| \mathbf{f}^n \right\|_{\mathscr{H}}^2 .$$

Proof. We multiply (7.11a) with $\rho_j^{n+1} \rho_j^n g_{jk}^{n+1}$ and put the last term of the equation from the right-hand to the left-hand side. This results in

$$\begin{aligned}
\left(\rho_j^{n+1} g_{jk}^{n+1}\right)^2 = \rho_j^n g_{jk}^n \rho_j^{n+1} g_{jk}^{n+1} &- \Delta t \rho_j^{n+1} g_{jk}^{n+1} \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n g_{ik}^n v_k \\
&+ \Delta t \frac{\Delta x}{2} \rho_j^{n+1} g_{jk}^{n+1} \sum_{i=1}^{N_x} D_{ji}^{xx} \rho_i^n g_{ik}^n |v_k| + \sigma \Delta t \rho_j^{n+1} g_{jk}^{n+1} \left(\rho_j^{n+1} - \rho_j^{n+1} g_{jk}^{n+1}\right).
\end{aligned}$$

Multiplication with $2\left(\frac{1}{\sqrt{2\pi}}e^{-v_k^2/2}\right)^2$ leads to

$$2\left(f_{jk}^{n+1}\right)^2 = 2f_{jk}^n f_{jk}^{n+1} - 2\Delta t f_{jk}^{n+1}\sum_{i=1}^{N_x} D_{ji}^x f_{ik}^n v_k + \Delta t\Delta x f_{jk}^{n+1}\sum_{i=1}^{N_x} D_{ji}^{xx} f_{ik}^n |v_k| + 2\sigma\Delta t f_{jk}^{n+1}\left(M_{jk}^{n+1} - f_{jk}^{n+1}\right).$$

Note that it holds

$$2f_{jk}^n f_{jk}^{n+1} = \left(f_{jk}^{n+1}\right)^2 + \left(f_{jk}^n\right)^2 - \left(f_{jk}^{n+1} - f_{jk}^n\right)^2. \tag{7.15}$$

We insert this relation and obtain

$$\left(f_{jk}^{n+1}\right)^2 = \left(f_{jk}^n\right)^2 - \left(f_{jk}^{n+1} - f_{jk}^n\right)^2 - 2\Delta t f_{jk}^{n+1}\sum_{i=1}^{N_x} D_{ji}^x f_{ik}^n v_k + \Delta t\Delta x f_{jk}^{n+1}\sum_{i=1}^{N_x} D_{ji}^{xx} f_{ik}^n |v_k| + 2\sigma\Delta t f_{jk}^{n+1}\left(M_{jk}^{n+1} - f_{jk}^{n+1}\right).$$

In the next step, we multiply with $\sqrt{2\pi}\omega_k e^{3v_k^2/2}$ and sum over j and k. This yields

$$\begin{aligned}\left\|\mathbf{f}^{n+1}\right\|_{\mathscr{H}}^2 = \left\|\mathbf{f}^n\right\|_{\mathscr{H}}^2 &- \sqrt{2\pi}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(f_{jk}^{n+1} - f_{jk}^n\right)^2\omega_k e^{3v_k^2/2} \\ &- 2\sqrt{2\pi}\Delta t\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1} D_{ji}^x f_{ik}^n v_k\omega_k e^{3v_k^2/2} \\ &+ \sqrt{2\pi}\Delta t\Delta x\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1} D_{ji}^{xx} f_{ik}^n |v_k|\,\omega_k e^{3v_k^2/2} \\ &+ 2\sqrt{2\pi}\sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1}\left(M_{jk}^{n+1} - f_{jk}^{n+1}\right)\omega_k e^{3v_k^2/2}.\end{aligned} \tag{7.16}$$

According to Lemma 7.4, we can use the equality $\sum_{k=1}^{N_v} f_{jk}^{n+1}\omega_k e^{v_k^2} = \rho_j^{n+1}$. Hence, we can conclude that the term $2\sqrt{2\pi}\sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=1}^{N_v} M_{jk}^{n+1}\left(M_{jk}^{n+1} - f_{jk}^{n+1}\right)\omega_k e^{3v_k^2/2}$, which is added in the next step, is equal to zero. Lemma 3.1 implies that also the term $2\sqrt{2\pi}\Delta t\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1} D_{ji}^x f_{ik}^{n+1} v_k\omega_k e^{3v_k^2/2}$, which is subtracted in the next step, is equal to zero. Furthermore, we add an artificial zero to the equation given in (7.16) by adding and subtracting the expression $\sqrt{2\pi}\Delta t\Delta x\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1} D_{ji}^{xx} f_{ik}^{n+1} |v_k|\,\omega_k e^{3v_k^2/2}$.

This leads to

$$\begin{aligned}
\left\|\mathbf{f}^{n+1}\right\|_{\mathscr{H}}^2 = \ & \|\mathbf{f}^n\|_{\mathscr{H}}^2 - \sqrt{2\pi}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(f_{jk}^{n+1}-f_{jk}^n\right)^2\omega_k e^{3v_k^2/2} \\
& - 2\sqrt{2\pi}\Delta t\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1}D_{ji}^x\left(f_{ik}^n - f_{ik}^{n+1}\right)v_k\omega_k e^{3v_k^2/2} \qquad \text{(I)} \\
& + \sqrt{2\pi}\Delta t\Delta x\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1}D_{ji}^{xx}\left(f_{ik}^n - f_{ik}^{n+1}\right)|v_k|\,\omega_k e^{3v_k^2/2} \qquad \text{(II)} \\
& + \sqrt{2\pi}\Delta t\Delta x\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1}D_{ji}^{xx}f_{ik}^{n+1}\,|v_k|\,\omega_k e^{3v_k^2/2} \qquad \text{(III)} \\
& - 2\sqrt{2\pi}\sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(f_{jk}^{n+1}-M_{jk}^{n+1}\right)^2\omega_k e^{3v_k^2/2}.
\end{aligned}$$

Now we analyze the terms (I), (II) and (III) separately. Let us start with (I) and (II) and apply Young's inequality proposed in Lemma 5.6. For the sum (I) + (II) this leads to

$$\begin{aligned}
& - 2\sqrt{2\pi}\Delta t\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1}D_{ji}^x\left(f_{ik}^n - f_{ik}^{n+1}\right)v_k\omega_k e^{3v_k^2/2} \\
& + \sqrt{2\pi}\Delta t\Delta x\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1}D_{ji}^{xx}\left(f_{ik}^n - f_{ik}^{n+1}\right)|v_k|\omega_k e^{3v_k^2/2} \\
= & \sum_{i=1}^{N_x}\sum_{k=1}^{N_v}\left(-\sqrt{2}\sqrt[4]{2\pi}\left(f_{ik}^n - f_{ik}^{n+1}\right)\sqrt{\omega_k}e^{3v_k^2/4}\right) \\
& \cdot\left(\sqrt{2}\sqrt[4]{2\pi}\Delta t\sum_{j=1}^{N_x}\left(D_{ji}^x f_{jk}^{n+1}v_k - \frac{\Delta x}{2}D_{ji}^{xx}f_{jk}^{n+1}|v_k|\right)\sqrt{\omega_k}e^{3v_k^2/4}\right) \\
\leq & \sqrt{2\pi}\sum_{i=1}^{N_x}\sum_{k=1}^{N_v}\left(f_{ik}^n - f_{ik}^{n+1}\right)^2\omega_k e^{3v_k^2/2} \\
& + \sqrt{2\pi}\,(\Delta t)^2\sum_{i=1}^{N_x}\sum_{k=1}^{N_v}\left(\sum_{j=1}^{N_x}\left(D_{ji}^x f_{jk}^{n+1}v_k - \frac{\Delta x}{2}D_{ji}^{xx}f_{jk}^{n+1}|v_k|\right)\right)^2\omega_k e^{3v_k^2/2}.
\end{aligned}$$

For (III) we exploit the properties of the spatial stencil matrices given in Lemma 3.1. We derive the equality

$$\begin{aligned}
& \sqrt{2\pi}\Delta t\Delta x\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v} f_{jk}^{n+1}D_{ji}^{xx}f_{ik}^{n+1}\,|v_k|\,\omega_k e^{3v_k^2/2} \\
= & -\sqrt{2\pi}\Delta t\Delta x\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(\sum_{i=1}^{N_x}D_{ji}^+ f_{ik}^{n+1}|v_k|^{1/2}\right)^2\omega_k e^{3v_k^2/2}.
\end{aligned}$$

We insert both relations and obtain

$$\begin{aligned}\|\mathbf{f}^{n+1}\|_{\mathscr{H}}^2 \leq \|\mathbf{f}^n\|_{\mathscr{H}}^2 &+ \sqrt{2\pi}\,(\Delta t)^2 \sum_{i=1}^{N_x}\sum_{k=1}^{N_v}\left(\sum_{j=1}^{N_x} D_{ji}^x f_{jk}^{n+1} v_k - \frac{\Delta x}{2} D_{ji}^{xx} f_{jk}^{n+1}\,|v_k|\right)^2 \omega_k e^{3v_k^2/2} \\ &- \sqrt{2\pi}\Delta t \Delta x \sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(\sum_{i=1}^{N_x} D_{ji}^{+} f_{ik}^{n+1}|v_k|^{1/2}\right)^2 \omega_k e^{3v_k^2/2} \\ &- 2\sqrt{2\pi}\sigma\Delta t \sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(f_{jk}^{n+1} - M_{jk}^{n+1}\right)^2 \omega_k e^{3v_k^2/2}.\end{aligned}$$

Together with Lemma 7.5 we can conclude that under the CFL condition $\max_k(|v_k|)\,\Delta t \leq \Delta x$ it holds $\left\|\mathbf{f}^{n+1}\right\|_{\mathscr{H}}^2 \leq \|\mathbf{f}^n\|_{\mathscr{H}}^2$. Hence, under this time step restriction the proposed fully discrete system (7.11) is numerically stable in the $\mathscr{H}$-norm. □

In principle, we have shown energy stability according to Section 3.2.4 for equations (7.11). However, as the considered $\mathscr{H}$-norm is not directly related to the physical energy of the system, we refer to it as *numerical stability* in the sense that the solution remains bounded over time.

7.4 Stable DLRA scheme for multiplicative linear Boltzmann-BGK

In practical applications, the implementation of the full system given in (7.11) may lead to prohibitive numerical costs, especially when computing in higher-dimensional settings. To reduce computational and memory demands, we apply a DLRA approach to the distribution function g. We first rewrite the conservative form of the equations. In (7.11a), we put all terms containing g_{jk}^{n+1} to the left-hand side and multiply the equation with $\frac{\rho_j^n}{\rho_j^{n+1}}$. Then equations (7.11) can be written in the equivalent form

$$\begin{aligned} g_{jk}^{n+1}(1+\sigma\Delta t) = \frac{\rho_j^n}{\rho_j^{n+1}} g_{jk}^n &- \Delta t \frac{1}{\rho_j^{n+1}} \sum_{i=1}^{N_x} D_{ji}^x\,(\rho_i^n g_{ik}^n)\,v_k \qquad (7.17a) \\ &+ \Delta t \frac{\Delta x}{2} \frac{1}{\rho_j^{n+1}} \sum_{i=1}^{N_x} D_{ji}^{xx}\,(\rho_i^n g_{ik}^n)\,|v_k| + \sigma\Delta t, \\ \rho_j^{n+1} = \rho_j^n &- \Delta t \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{N_x}\sum_{k=1}^{N_v} D_{ji}^x \rho_i^n g_{ik}^n v_k \omega_k e^{v_k^2/2} \qquad (7.17b) \\ &+ \Delta t \frac{\Delta x}{2\sqrt{2\pi}} \sum_{i=1}^{N_x}\sum_{k=1}^{N_v} D_{ji}^{xx} \rho_i^n g_{ik}^n |v_k| \omega_k e^{v_k^2/2}. \end{aligned}$$

We propose a numerically stable DLRA implementation that uses the rank-adaptive augmented BUG integrator presented in [CKL22] for equation (7.17a) together with additional basis augmentations and a suitable truncation strategy. Note that the scattering term $1 + \sigma\Delta t$ is only applied in the S-step as it does not affect the span of the basis functions derived in the K- and L-step. In detail, the DLRA scheme works as follows.

We first substitute $g_{jk}^n = \sum_{m,\eta=1}^r X_{jm}^n S_{m\eta}^n V_{k\eta}^n$ into the update equation (7.17b) and obtain

$$\begin{aligned}\rho_j^{n+1} = \rho_j^n - \Delta t \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n \sum_{m,\eta=1}^{r} X_{im}^n S_{m\eta}^n \sum_{k=1}^{N_v} V_{k\eta}^n v_k \omega_k e^{v_k^2/2} \\ + \Delta t \frac{\Delta x}{2\sqrt{2\pi}} \sum_{i=1}^{N_x} D_{ji}^{xx} \rho_i^n \sum_{m,\eta=1}^{r} X_{im}^n S_{m\eta}^n \sum_{k=1}^{N_v} V_{k\eta}^n |v_k| \omega_k e^{v_k^2/2}.\end{aligned} \tag{7.18a}$$

For the K-step, we introduce the notation $K_{j\eta}^n = \sum_{m=1}^r X_{jm}^n S_{m\eta}^n$ and solve

$$\begin{aligned}K_{jp}^{n+1} = \frac{\rho_j^n}{\rho_j^{n+1}} K_{jp}^n - \Delta t \frac{1}{\rho_j^{n+1}} \sum_{i=1}^{N_x} D_{ji}^x \rho_i^n \sum_{\eta=1}^{r} K_{i\eta}^n \sum_{k=1}^{N_v} V_{k\eta}^n v_k V_{kp}^n \\ + \Delta t \frac{\Delta x}{2} \frac{1}{\rho_j^{n+1}} \sum_{i=1}^{N_x} D_{ji}^{xx} \rho_i^n \sum_{\eta=1}^{r} K_{i\eta}^n \sum_{k=1}^{N_v} V_{k\eta}^n |v_k| V_{kp}^n + \sigma \Delta t \sum_{k=1}^{N_v} V_{kp}^n.\end{aligned} \tag{7.18b}$$

We derive the updated basis $\widehat{\mathbf{X}}^{n+1}$ of rank $2r$ from a QR-decomposition of the augmented quantity $\widehat{\mathbf{X}}^{n+1} = \mathrm{qr}\left(\left[\mathbf{K}^{n+1}, \mathbf{X}^n\right]\right)$. In addition, we augment the basis according to

$$\widehat{\widehat{\mathbf{X}}}^{n+1} = \mathrm{qr}\left(\left[\widehat{\mathbf{X}}^{n+1}, \left(\boldsymbol{\rho}^{n+1}\right)^2 \widehat{\mathbf{X}}^{n+1}\right]\right). \tag{7.18c}$$

This basis augmentation ensures the exactness of the corresponding projection operators in the proof of stability of the proposed scheme. Its explicit form will be made clear later. We compute and store $\widehat{\widehat{\mathbf{M}}} = \widehat{\widehat{\mathbf{X}}}^{n+1,\top} \mathbf{X}^n$. Note that for this scheme we perform full rank updates, leading to an increase from rank $2r$ to $4r$. Quantities of rank $2r$ are denoted with one single hat and quantities of rank $4r$ with double hats.

For the L-step, which can be computed in parallel with the K-step, we write $L_{km}^n = \sum_{\eta=1}^r S_{m\eta}^n V_{k\eta}^n$ and solve

$$\begin{aligned}L_{kp}^{n+1} = \sum_{m=1}^{r} L_{km}^n \sum_{j=1}^{N_x} X_{jm}^n \frac{\rho_j^n}{\rho_j^{n+1}} X_{jp}^n - \Delta t \sum_{m=1}^{r} v_k L_{km}^n \sum_{i=1}^{N_x} X_{im}^n \rho_i^n \sum_{j=1}^{N_x} D_{ji}^x \frac{1}{\rho_j^{n+1}} X_{jp}^n \\ + \Delta t \frac{\Delta x}{2} \sum_{m=1}^{r} |v_k| L_{km}^n \sum_{i=1}^{N_x} X_{im}^n \rho_i^n \sum_{j=1}^{N_x} D_{ji}^{xx} \frac{1}{\rho_j^{n+1}} X_{jp}^n + \sigma \Delta t \sum_{j=1}^{N_x} X_{jp}^n.\end{aligned} \tag{7.18d}$$

We derive the updated basis $\widehat{\mathbf{V}}^{n+1}$ of rank $2r$ from a QR-decomposition of the augmented

quantity $\widehat{\mathbf{V}}^{n+1} = \mathrm{qr}\left(\left[\mathbf{L}^{n+1}, \mathbf{V}^n\right]\right)$. In addition, we augment the basis according to

$$\widehat{\widehat{\mathbf{V}}}^{n+1} = \mathrm{qr}\left(\left[\widehat{\mathbf{V}}^{n+1}, \omega e^{\mathbf{v}^2/2}\widehat{\mathbf{V}}^{n+1}\right]\right) \tag{7.18e}$$

leading to a new augmented basis $\widehat{\widehat{\mathbf{V}}}^{n+1}$ of rank $4r$. This basis augmentation again ensures the exactness of the corresponding projection operators and will be made clear later in the proof of numerical stability. We compute and store $\widehat{\widehat{\mathbf{N}}} = \widehat{\widehat{\mathbf{V}}}^{n+1,\top}\mathbf{V}^n$.
For the S-step, the previously computed solutions from the K- and L-step are used. We introduce the notation $\widetilde{S}^n_{m\eta} = \sum_{j,k=1}^r \widehat{\widehat{M}}_{mj} S^n_{jk} \widehat{\widehat{N}}_{\eta k}$ and solve

$$\begin{aligned}
\widehat{\widehat{S}}^{n+1}_{qp} = {} & \frac{1}{1+\sigma\Delta t}\sum_{j=1}^{N_x} \widehat{\widehat{X}}^{n+1}_{jq} \frac{\rho^n_j}{\rho^{n+1}_j} \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}^{n+1}_{jm} \widetilde{S}^n_{m\eta} \sum_{k=1}^{N_v} \widehat{\widehat{V}}^{n+1}_{k\eta} \widehat{\widehat{V}}^{n+1}_{kp} \\
& - \frac{\Delta t}{1+\sigma\Delta t}\sum_{j=1}^{N_x} \widehat{\widehat{X}}^{n+1}_{jq} \frac{1}{\rho^{n+1}_j} \sum_{i=1}^{N_x} D^x_{ji}\rho^n_i \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}^{n+1}_{im} \widetilde{S}^n_{m\eta} \sum_{k=1}^{N_v} \widehat{\widehat{V}}^{n+1}_{k\eta} v_k \widehat{\widehat{V}}^{n+1}_{kp} \\
& + \frac{\Delta t}{1+\sigma\Delta t}\frac{\Delta x}{2}\sum_{j=1}^{N_x} \widehat{\widehat{X}}^{n+1}_{jq} \frac{1}{\rho^{n+1}_j} \sum_{i=1}^{N_x} D^{xx}_{ji}\rho^n_i \sum_{m,\eta=1}^{4r} \widehat{\widehat{X}}^{n+1}_{im} \widetilde{S}^n_{m\eta} \sum_{k=1}^{N_v} \widehat{\widehat{V}}^{n+1}_{k\eta} |v_k| \widehat{\widehat{V}}^{n+1}_{kp} \\
& + \frac{\sigma\Delta t}{1+\sigma\Delta t}\sum_{j=1}^{N_x} \widehat{\widehat{X}}^{n+1}_{jq} \sum_{k=1}^{N_v} \widehat{\widehat{V}}^{n+1}_{kp}.
\end{aligned} \tag{7.18f}$$

The last step consists in truncating the augmented quantities $\widehat{\widehat{\mathbf{X}}}^{n+1}$, $\widehat{\widehat{\mathbf{V}}}^{n+1}$ and $\widehat{\widehat{\mathbf{S}}}^{n+1}$ from rank $4r$ to a new rank r_{n+1}. We use a modification of the truncation strategy described in Section 4.2.2 that ensures that the equality $\frac{1}{\sqrt{2\pi}}\sum_{m,\eta=1}^r \sum_{k=1}^{N_v} X^n_{jm} S^n_{m\eta} V^n_{k\eta} \omega_k e^{v_k^2/2} = 1$ stays valid in each time step and works as follows:

(i) We set $\mathbf{Z} = (Z_k) \in \mathbb{R}^{N_v}$ with entries $Z_k = \frac{1}{\sqrt{2\pi}}\omega_k e^{v_k^2/2}$ and $\mathbf{z} = \frac{\mathbf{Z}}{\|\mathbf{Z}\|_E}$, where $\|\cdot\|_E$ denotes the Euclidean norm. Further, we set $\mathbf{H}_1 = \widehat{\widehat{\mathbf{X}}}^{n+1}\widehat{\widehat{\mathbf{S}}}^{n+1}\widehat{\widehat{\mathbf{V}}}^{n+1,\top}\mathbf{z}\mathbf{z}^\top$ and $\mathbf{H}_2 = \widehat{\widehat{\mathbf{X}}}^{n+1}\widehat{\widehat{\mathbf{S}}}^{n+1}\widehat{\widehat{\mathbf{V}}}^{n+1,\top}\left(\mathbf{I} - \mathbf{z}\mathbf{z}^\top\right)$ in order to ensure

$$\begin{aligned}
\mathbf{1} &= \widehat{\widehat{\mathbf{X}}}^{n+1}\widehat{\widehat{\mathbf{S}}}^{n+1}\widehat{\widehat{\mathbf{V}}}^{n+1,\top}\mathbf{Z} \\
&= \left(\widehat{\widehat{\mathbf{X}}}^{n+1}\widehat{\widehat{\mathbf{S}}}^{n+1}\widehat{\widehat{\mathbf{V}}}^{n+1,\top}\mathbf{z}\mathbf{z}^\top + \widehat{\widehat{\mathbf{X}}}^{n+1}\widehat{\widehat{\mathbf{S}}}^{n+1}\widehat{\widehat{\mathbf{V}}}^{n+1,\top}\left(\mathbf{I} - \mathbf{z}\mathbf{z}^\top\right)\right)\mathbf{Z} \\
&= \left(\mathbf{H}_1 + \mathbf{H}_2\right)\mathbf{Z},
\end{aligned}$$

where $\mathbf{I} \in \mathbb{R}^{N_v \times N_v}$ represents the identity matrix and $\mathbf{1} \in \mathbb{R}^{N_x}$ the vector containing the value one at each entry. $\mathbf{H}_1$ is a matrix of rank one and for $\mathbf{H}_2$ it holds that $\mathbf{H}_2\mathbf{Z} = 0$.

(ii) We compute $\mathbf{X}^{\mathbf{H}_1}\mathbf{S}^{\mathbf{H}_1}\mathbf{V}^{\mathbf{H}_1,\top} = \mathrm{svd}\left(\widehat{\widehat{\mathbf{S}}}^{n+1}\widehat{\widehat{\mathbf{V}}}^{n+1,\top}\mathbf{z}\mathbf{z}^\top\right)$ from an SVD, where $\mathbf{X}^{\mathbf{H}_1} \in \mathbb{R}^{4r}, \mathbf{S}^{\mathbf{H}_1} \in \mathbb{R}$, and $\mathbf{V}^{\mathbf{H}_1} \in \mathbb{R}^{N_v}$.

(iii) We compute $\widehat{\widehat{\mathbf{P}}}\mathbf{\Sigma}\widehat{\widehat{\mathbf{Q}}}^\top = \mathrm{svd}\left(\widehat{\widehat{\mathbf{S}}}^{n+1}\right)$ from an SVD, where $\widehat{\widehat{\mathbf{P}}}, \widehat{\widehat{\mathbf{Q}}} \in \mathbb{R}^{4r\times 4r}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{4r\times 4r}$ is the diagonal matrix containing the singular values

$\sigma_1, ..., \sigma_{4r}$. The new rank $\widetilde{r} \leq 4r$ is determined such that

$$\left(\sum_{j=\widetilde{r}+1}^{4r} \sigma_j^2 \right)^{1/2} \leq \vartheta,$$

where ϑ denotes a prescribed tolerance parameter. We set $\mathbf{S}^{\mathbf{H}_2} \in \mathbb{R}^{\widetilde{r}\times\widetilde{r}}$ to be the matrix containing the $\widetilde{r}$ largest singular values of $\widehat{\widehat{\mathbf{S}}}^{n+1}$ and the matrices $\widehat{\widehat{\mathbf{P}}}^{\mathbf{H}_2}, \widehat{\widehat{\mathbf{Q}}}^{\mathbf{H}_2} \in \mathbb{R}^{4r\times\widetilde{r}}$ to contain the first $\widetilde{r}$ columns of $\widehat{\widehat{\mathbf{P}}}$ and $\widehat{\widehat{\mathbf{Q}}}$, respectively. Finally, we compute $\mathbf{X}^{\mathbf{H}_2} = \widehat{\widehat{\mathbf{X}}}^{n+1}\widehat{\widehat{\mathbf{P}}}^{\mathbf{H}_2} \in \mathbb{R}^{N_x\times\widetilde{r}}$ and $\mathbf{V}^{\mathbf{H}_2} = \left(\mathbf{I} - \mathbf{z}\mathbf{z}^\top\right)^\top \widehat{\widehat{\mathbf{V}}}^{n+1}\widehat{\widehat{\mathbf{Q}}}^{\mathbf{H}_2} \in \mathbb{R}^{N_v\times\widetilde{r}}$.

(iv) We combine both parts and perform a QR-decomposition to obtain

$$\mathbf{X}^{n+1}\mathbf{R}^1 = \mathrm{qr}\left(\left[\widehat{\widehat{\mathbf{X}}}^{n+1}\mathbf{X}^{\mathbf{H}_1}, \mathbf{X}^{\mathbf{H}_2}\right]\right) \quad \text{and} \quad \mathbf{V}^{n+1}\mathbf{R}^2 = \mathrm{qr}\left(\left[\mathbf{V}^{\mathbf{H}_1}, \mathbf{V}^{\mathbf{H}_2}\right]\right).$$

(v) We compute

$$\mathbf{S}^{n+1} = \mathbf{R}^1 \begin{bmatrix} \mathbf{S}^{\mathbf{H}_1} & 0 \\ 0 & \mathbf{S}^{\mathbf{H}_2} \end{bmatrix} \mathbf{R}^{2,\top}.$$

Then the new rank r_{n+1} is given by $r_{n+1} = \widetilde{r} + 1$.

Altogether, this leads to the updated solution $\mathbf{g}^{n+1} = \mathbf{X}^{n+1}\mathbf{S}^{n+1}\mathbf{V}^{n+1,\top}$ after one time step at time $t_{n+1} = t_n + \Delta t$. To provide an overview of the structure of the proposed DLRA scheme, its working principle is visualized in Algorithm 3. Note that the notation using brackets refers to a simplification of the algorithm that is explained later in Section 7.5.1.

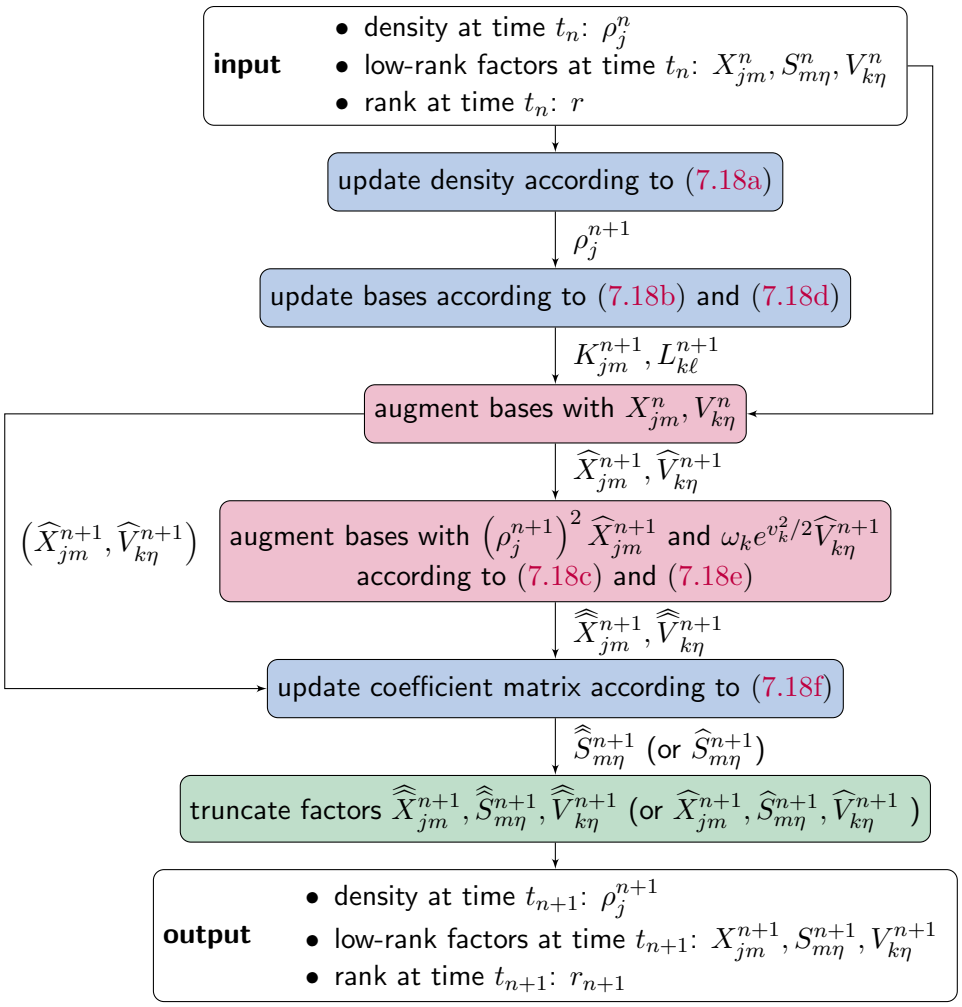
Proof of stability of the proposed multiplicative low-rank scheme. It can be shown that the DLRA scheme proposed in (7.18) preserves the numerical stability of the full conservative system presented in (7.11), which has been shown in Theorem 7.6. The rewriting of equations (7.11) into (7.17), the basis augmentations in (7.18c) and (7.18e) and the implementation of the described truncation strategy are crucial for the proof. We begin with the following definition.

Definition 7.7 (Low-rank approximation of the fully discrete solution)**.** The *low-rank approximation of the fully discrete solution* f *at time* t_n is given by $\mathbf{f}^n = \left(f_{jk}^n\right) \in \mathbb{R}^{N_x\times N_v}$ with entries

$$f_{jk}^n = \frac{1}{\sqrt{2\pi}}\rho_j^n \sum_{m,\eta=1}^{r} X_{jm}^n S_{m\eta}^n V_{k\eta}^n e^{-v_k^2/2}.$$

Note that in this notation we do not distinguish between the full solution $\mathbf{f}^n$ and its low-rank approximation $\mathbf{f}_r^n$ at time t_n. Then we can show that the DLRA scheme (7.18) is numerically stable.

Algorithm 3 Flowchart of the (simplified) stable multiplicative DLRA scheme (7.18).



Theorem 7.8 (Numerical stability of the proposed multiplicative DLRA scheme)**.** *Under the time step restriction* $\max_k(|v_k|)\Delta t \leq \Delta x$ *the fully discrete DLRA scheme* (7.18) *is numerically stable in the* $\mathscr{H}$*-norm, i.e. it holds*

$$\left\|\mathbf{f}^{n+1}\right\|_{\mathscr{H}}^2 \leq \left\|\mathbf{f}^{n}\right\|_{\mathscr{H}}^2 .$$

Proof. We begin with the S-step given in (7.18f), multiply it with $\widehat{\widehat{X}}^{n+1}_{\alpha q}\widehat{\widehat{V}}^{n+1}_{\beta p}$ and sum over q and p. For simplicity, the notation $g^{n+1}_{\alpha\beta} := \sum_{q,p=1}^{4r}\widehat{\widehat{X}}^{n+1}_{\alpha q}\widehat{\widehat{S}}^{n+1}_{qp}\widehat{\widehat{V}}^{n+1}_{\beta p}$ as well as the projection operators $P^{X^{n+1}}_{\alpha j} = \sum_{q=1}^{4r}\widehat{\widehat{X}}^{n+1}_{\alpha q}\widehat{\widehat{X}}^{n+1}_{jq}$ and $P^{V^{n+1}}_{k\beta} = \sum_{p=1}^{4r}\widehat{\widehat{V}}^{n+1}_{kp}\widehat{\widehat{V}}^{n+1}_{\beta p}$ are introduced. This leads to

$$\begin{aligned}
g^{n+1}_{\alpha\beta} = &\sum_{j=1}^{N_x} P^{X^{n+1}}_{\alpha j}\frac{\rho^n_j}{\rho^{n+1}_j}\sum_{k=1}^{N_v} g^n_{jk}P^{V^{n+1}}_{k\beta}\\
&- \frac{\Delta t}{1+\sigma\Delta t}\sum_{j=1}^{N_x} P^{X^{n+1}}_{\alpha j}\frac{1}{\rho^{n+1}_j}\sum_{i=1}^{N_x} D^x_{ji}\rho^n_i\sum_{k=1}^{N_v} g^n_{ik}v_k P^{V^{n+1}}_{k\beta}\\
&+ \frac{\Delta t}{1+\sigma\Delta t}\frac{\Delta x}{2}\sum_{j=1}^{N_x} P^{X^{n+1}}_{\alpha j}\frac{1}{\rho^{n+1}_j}\sum_{i=1}^{N_x} D^{xx}_{ji}\rho^n_i\sum_{k=1}^{N_v} g^n_{ik}\left|v_k\right| P^{V^{n+1}}_{k\beta}\\
&+ \frac{\sigma\Delta t}{1+\sigma\Delta t}\sum_{j=1}^{N_x} P^{X^{n+1}}_{\alpha j}\sum_{k=1}^{N_v} P^{V^{n+1}}_{k\beta}.
\end{aligned}$$

We multiply with $(1+\sigma\Delta t)\frac{2}{\sqrt{2\pi}}\left(\rho^{n+1}_\alpha\right)^2 g^{n+1}_{\alpha\beta}\omega_\beta e^{v_\beta^2/2}$ and sum over α and β. Then,

$$\begin{aligned}
&(1+\sigma\Delta t)\frac{2}{\sqrt{2\pi}}\sum_{\alpha=1}^{N_x}\sum_{\beta=1}^{N_v}\left(\rho^{n+1}_\alpha g^{n+1}_{\alpha\beta}\right)^2\omega_\beta e^{v_\beta^2/2}\\
&= \frac{2}{\sqrt{2\pi}}\sum_{j,\alpha=1}^{N_x} P^{X^{n+1}}_{\alpha j}\frac{\rho^n_j}{\rho^{n+1}_j}\left(\rho^{n+1}_\alpha\right)^2\sum_{k=1}^{N_v} g^n_{jk}\sum_{\beta=1}^{N_v} P^{V^{n+1}}_{k\beta}g^{n+1}_{\alpha\beta}\omega_\beta e^{v_\beta^2/2}\\
&\quad - \frac{2\Delta t}{\sqrt{2\pi}}\sum_{j,\alpha=1}^{N_x} P^{X^{n+1}}_{\alpha j}\frac{1}{\rho^{n+1}_j}\left(\rho^{n+1}_\alpha\right)^2\sum_{i=1}^{N_x} D^x_{ji}\rho^n_i\sum_{k=1}^{N_v} g^n_{ik}v_k\sum_{\beta=1}^{N_v} P^{V^{n+1}}_{k\beta}g^{n+1}_{\alpha\beta}\omega_\beta e^{v_\beta^2/2}\\
&\quad + \Delta t\frac{\Delta x}{\sqrt{2\pi}}\sum_{j=1,\alpha}^{N_x} P^{X^{n+1}}_{\alpha j}\frac{1}{\rho^{n+1}_j}\left(\rho^{n+1}_\alpha\right)^2\sum_{i=1}^{N_x} D^{xx}_{ji}\rho^n_i\sum_{k=1}^{N_v} g^n_{ik}\left|v_k\right|\sum_{\beta=1}^{N_v} P^{V^{n+1}}_{k\beta}g^{n+1}_{\alpha\beta}\omega_\beta e^{v_\beta^2/2}\\
&\quad + \frac{2\sigma\Delta t}{\sqrt{2\pi}}\sum_{j,\alpha=1}^{N_x} P^{X^{n+1}}_{\alpha j}\left(\rho^{n+1}_\alpha\right)^2\sum_{\beta=1}^{N_v} g^{n+1}_{\alpha\beta}\omega_\beta e^{v_\beta^2/2}\sum_{k=1}^{N_v} P^{V^{n+1}}_{k\beta}.
\end{aligned}$$

Using the basis augmentations given in (7.18c) and (7.18e), we can deduce that the equalities

$$\sum_{\alpha=1}^{N_x} P^{X^{n+1}}_{\alpha j}\left(\rho^{n+1}_\alpha\right)^2 g^{n+1}_{\alpha\beta} = \left(\rho^{n+1}_j\right)^2 g^{n+1}_{j\beta} \quad \text{and} \quad \sum_{\beta=1}^{N_v} P^{V^{n+1}}_{k\beta}g^{n+1}_{j\beta}\omega_\beta e^{v_\beta^2/2} = g^{n+1}_{jk}\omega_k e^{v_k^2/2}$$

hold. We insert these relations and, to be consistent in notation, change the summation indices on the left-hand side from α to j and from β to k. This leads to

$$
\begin{aligned}
&(1+\sigma\Delta t)\,\frac{2}{\sqrt{2\pi}}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(\rho_j^{n+1}g_{jk}^{n+1}\right)^2\omega_k e^{v_k^2/2}\\
&=\frac{2}{\sqrt{2\pi}}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\rho_j^n g_{jk}^n\rho_j^{n+1}g_{jk}^{n+1}\omega_k e^{v_k^2/2}-\frac{2\Delta t}{\sqrt{2\pi}}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\rho_j^{n+1}g_{jk}^{n+1}\sum_{i=1}^{N_x}D_{ji}^x\rho_i^n g_{ik}^n v_k\omega_k e^{v_k^2/2}\\
&+\Delta t\frac{\Delta x}{\sqrt{2\pi}}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\rho_j^{n+1}g_{jk}^{n+1}\sum_{i=1}^{N_x}D_{ji}^{xx}\rho_i^n g_{ik}^n\left|v_k\right|\omega_k e^{v_k^2/2}\\
&+\frac{2\sigma\Delta t}{\sqrt{2\pi}}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(\rho_j^{n+1}\right)^2 g_{jk}^{n+1}\omega_k e^{v_k^2/2}.
\end{aligned}
$$

We rearrange the equation, insert the notations from Definition 7.1 and use the relation stated in (7.15), yielding

$$
\begin{aligned}
\left\|\mathbf{f}^{n+1}\right\|_{\mathscr{H}}^2 = \;&\|\mathbf{f}^n\|_{\mathscr{H}}^2-\sqrt{2\pi}\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}\left(f_{jk}^{n+1}-f_{jk}^n\right)^2\omega_k e^{3v_k^2/2}\\
&-2\sqrt{2\pi}\Delta t\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v}f_{jk}^{n+1}D_{ji}^x f_{ik}^n v_k\omega_k e^{3v_k^2/2}\\
&+\sqrt{2\pi}\Delta t\Delta x\sum_{i,j=1}^{N_x}\sum_{k=1}^{N_v}f_{jk}^{n+1}D_{ji}^{xx}f_{ik}^n\left|v_k\right|\omega_k e^{3v_k^2/2}\\
&+2\sqrt{2\pi}\sigma\Delta t\sum_{j=1}^{N_x}\sum_{k=1}^{N_v}f_{jk}^{n+1}\left(M_{jk}^{n+1}-f_{jk}^{n+1}\right)\omega_k e^{3v_k^2/2},
\end{aligned}
$$

which is exactly expression (7.16) dealt with in the proof of Theorem 7.6. As the truncation step is specifically designed to leave these expressions unchanged, we can conclude analogously to the proof of Theorem 7.6 that the proposed DLRA scheme decreases the $\mathscr{H}$-norm and hence is numerically stable under the time step restriction $\max_k(|v_k|)\Delta t \leq \Delta x$. □

7.5 Numerical results

To validate our theoretical considerations, we compare the solution of the full equations (7.17) to the solution obtained by the DLRA scheme given in (7.18) for different numerical test examples. Section 7.5.1 reconsiders the 1D plane source problem, before in Section 7.5.2 a 1D tanh problem with more challenging initial distributions is considered. In Sections 7.5.3 and 7.5.4 2D test examples are presented in order to investigate the computational benefit of the DLRA scheme in higher-dimensional settings.

7.5.1 1D plane source

We begin with the 1D plane source test problem, which has already been treated in Sections 5.6.1 and 6.6.1 for the (multiplicative) Su-Olson problem. We consider the spatial domain $\Omega_x = [-10, 10]$ and choose the initial density ρ to be the cutoff Gaussian

$$\rho\left(t=0,x\right) = \max\left(10^{-4}, \frac{1}{\sqrt{2\pi\sigma_{\mathrm{IC}}^2}} \exp\left(-\frac{x^2}{2\sigma_{\mathrm{IC}}^2}\right)\right)$$

with constant deviation $\sigma_{\mathrm{IC}} = 0.3$. The initial distribution function g is assumed to be constant in space and velocity and we prescribe

$$g\left(t=0,x,v\right) = 1.$$

We consider a relatively large collisionality by choosing $\sigma = 10$. For the low-rank computations an initial rank of $r = 20$ is prescribed. As computational parameters we use $N_x = 1000$ grid cells in the spatial as well as $N_v = 500$ grid points in the velocity domain. Based on this choice, we obtain $\max_k\left(|v_k|\right) \approx 31.05$, which is adjusted to the next larger integer. The time step size is determined by $\Delta t = C_{\mathrm{CFL}} \cdot \frac{\Delta x}{32}$ with $C_{\mathrm{CFL}} = 0.99$, according to the corresponding CFL condition.

Practical implementations show that the basis augmentations to rank $4r$ performed in (7.18c) and (7.18e), which are needed for the theoretical proof of numerical stability, may not be necessary for numerical examples and that the standard basis augmentations to rank $2r$ provide similar solutions while being significantly faster. For this reason, we propose to leave out the basis augmentations presented in (7.18c) and (7.18e) in practical applications. In this case, all quantities with double hats related to rank $4r$ decrease to quantities of rank $2r$ with one single hat. The simplified scheme with rank $2r$ is also visualized (in brackets) in the flowchart of Algorithm 3.

In Figure 7.1 we compare the results for the solution $f\left(t,x,v\right)$ computed with the multiplicative full solver, the simplified multiplicative DLRA scheme with rank $2r$ and the basis augmented multiplicative DLRA scheme with rank $4r$ at different times up to $t_{\mathrm{End}} = 6$. We observe that the reduced as well as the augmented multiplicative DLRA algorithm capture the main characteristics of the full reference system. This is also true for the computational results for the density $\rho\left(t,x\right)$ displayed in Figure 7.2. Figure 7.3 shows the evolution of the rank in time, which for a chosen tolerance parameter of $\vartheta = 10^{-5}\left\|\mathbf{\Sigma}\right\|_F$ increases up to $r = 75$ before it significantly decreases over time. Note that the evolution of the rank for the reduced as well as for the basis augmented multiplicative DLRA algorithm show good agreement as the new rank is displayed after the corresponding truncation step. Further, the evolution of the norm $\left\|\mathbf{f}\right\|_{\mathscr{H}}^2$ in time is illustrated. As expected from the theoretical results, its value decreases smoothly over time for all considered systems. Additionally, we display the quantities $\kappa^+ := \max_j\left(\frac{1}{\sqrt{2\pi}}\sum_{k=1}^{N_v} g_{jk}\omega_k e^{v_k^2/2}\right)$ and $\kappa^- := \min_j\left(\frac{1}{\sqrt{2\pi}}\sum_{k=1}^{N_v} g_{jk}\omega_k e^{v_k^2/2}\right)$. According to Lemma 7.4, it is essential that they are both equal to 1, which for the DLRA schemes is ensured by the adjusted truncation step. It can be observed that this property is fulfilled up to order $\mathcal{O}\left(10^{-10}\right)$.

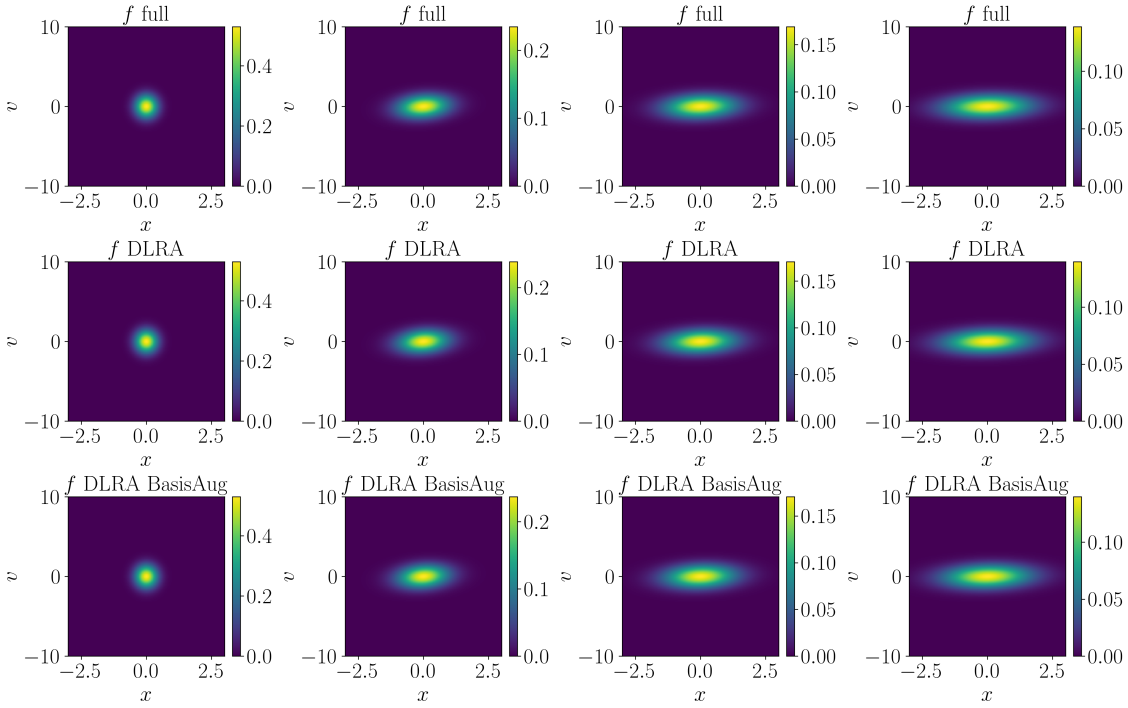


Figure 7.1: Numerical results for the solution $f(t, x, v)$ of the 1D plane source problem at time $t = 0$ (first column), $t = 2$ (second column), $t = 4$ (third column), and $t = 6$ (fourth column), computed with the multiplicative full solver (first row), the reduced multiplicative DLRA scheme (second row), and the basis augmented multiplicative DLRA scheme (third row).

7.5.2 1D tanh

For the next 1D test problem, a more challenging initial density distribution is considered. The initial density ρ is chosen to be

$$\rho(t = 0, x) = \begin{cases} \tanh(x) & \text{for } x < -1, \\ 1 & \text{for } x \in [-1, 1], \\ \coth(x) - 2 & \text{for } x > 1. \end{cases}$$

The initial distribution function g is assigned to be constant in space and velocity and we prescribe

$$g(t = 0, x, v) = 1.$$

All other initial settings and computational parameters remain unchanged from the previous test example given in Section 7.5.1.

In Figure 7.4 a comparison of the numerical results for the solution $f(t, x, v)$ computed with the three different solvers up to $t_{\text{End}} = 6$ is given. We observe that both DLRA algorithms are able to reproduce the full solution. This is also true for the computational results for the density $\rho(t, x)$ displayed in Figure 7.5. In addition, Figure 7.6 shows the evolution of the rank in time, which for a chosen tolerance parameter of $\vartheta = 10^{-5} \|\mathbf{\Sigma}\|_F$

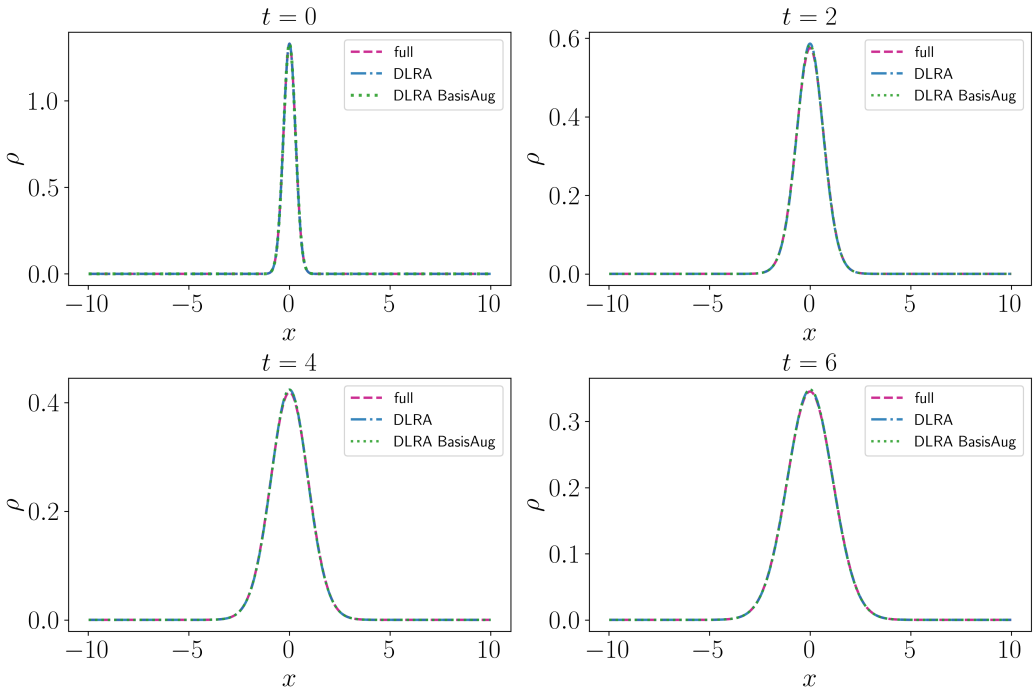


Figure 7.2: Numerical results for the density $\rho(t, x)$ of the 1D plane source problem at time $t = 0$, $t = 2$, $t = 4$, and $t = 6$, computed with the multiplicative full solver, the reduced multiplicative DLRA scheme, and the basis augmented multiplicative DLRA scheme.

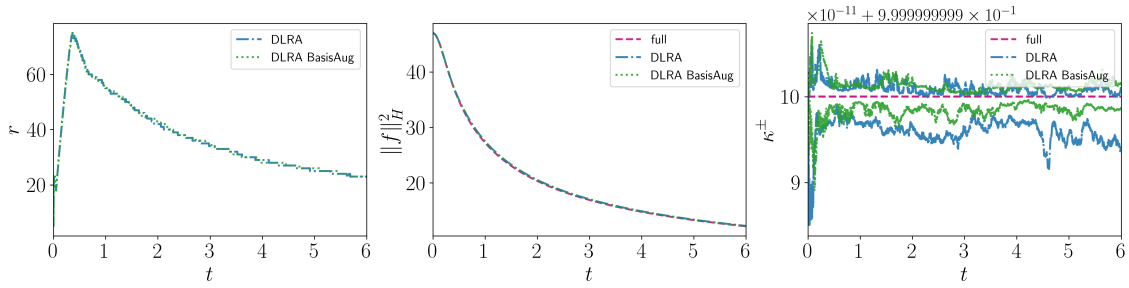


Figure 7.3: Left: Evolution of the rank in time for the 1D plane source problem for the reduced multiplicative DLRA scheme and the basis augmented multiplicative DLRA scheme. **Middle:** Evolution of the $\mathscr{H}$-norm in time for the multiplicative full solver, the reduced multiplicative DLRA scheme, and the basis augmented multiplicative DLRA scheme. **Right:** Evolution of $\kappa^{\pm}$ in time for the multiplicative full solver, the reduced multiplicative DLRA scheme, and the basis augmented multiplicative DLRA scheme. The line corresponding to the full system has the constant value 1.

increases up to $r = 93$ before it significantly decreases over time. Again, the evolution of the rank for the reduced and for the basis augmented multiplicative DLRA algorithm nearly coincide. Further, the evolution of the norm $\|\mathbf{f}\|_{\mathscr{H}}^2$ in time is illustrated. Its value decreases smoothly over time for all considered systems. Additionally, we display the quantities κ^+ and κ^- defined in Section 7.5.1, which are required to be equal to 1. This property is fulfilled up to order $\mathcal{O}\left(10^{-8}\right)$ for all schemes.

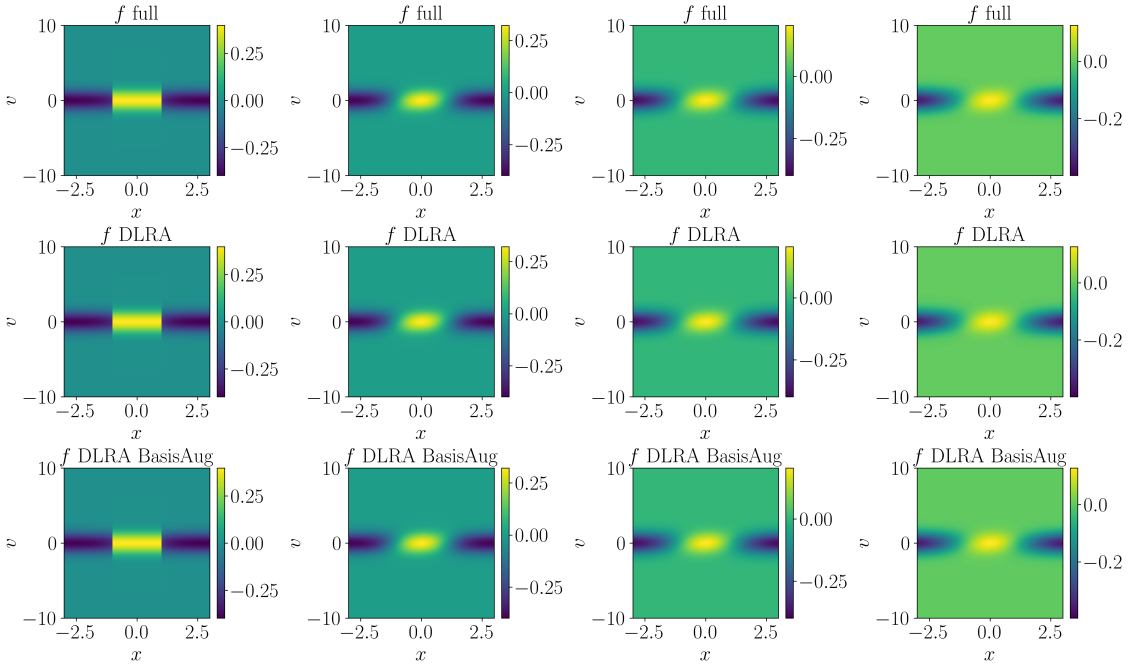


Figure 7.4: Numerical results for the solution $f(t,x,v)$ of the 1D $\tanh$ problem at time $t=0$ (first column), $t=2$ (second column), $t=4$ (third column), and $t=6$ (fourth column), computed with the multiplicative full solver (first row), the reduced multiplicative DLRA scheme (second row), and the basis augmented multiplicative DLRA scheme (third row).

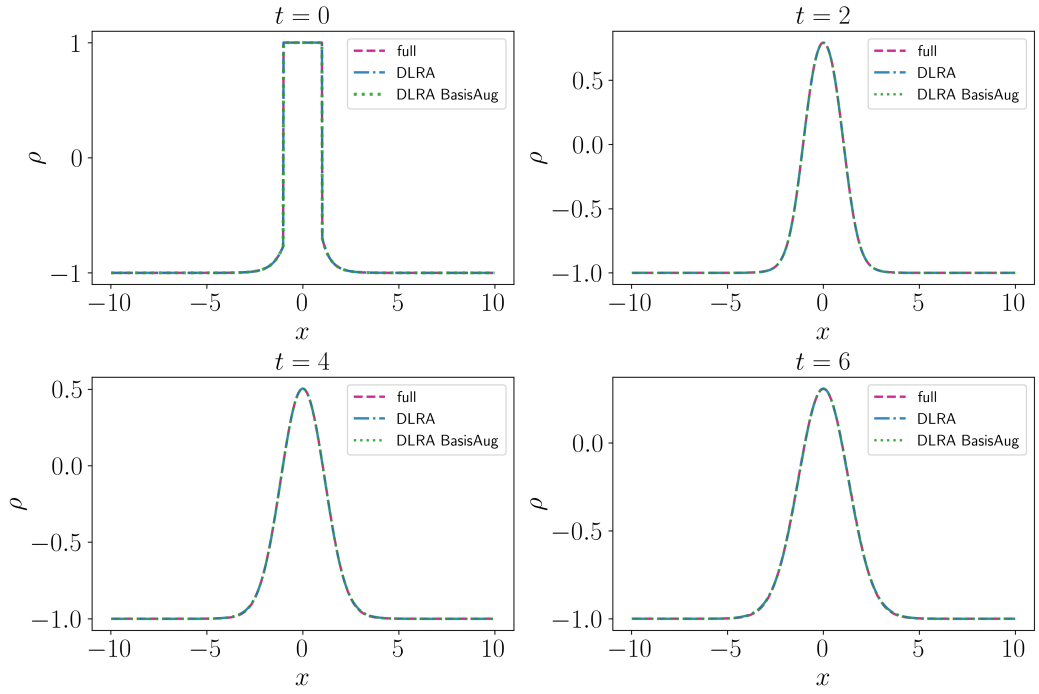


Figure 7.5: Numerical results for the density $\rho(t,x)$ of the 1D $\tanh$ problem at time $t=0$, $t=2$, $t=4$, and $t=6$, computed with the multiplicative full solver, the reduced multiplicative DLRA scheme, and the basis augmented multiplicative DLRA scheme.

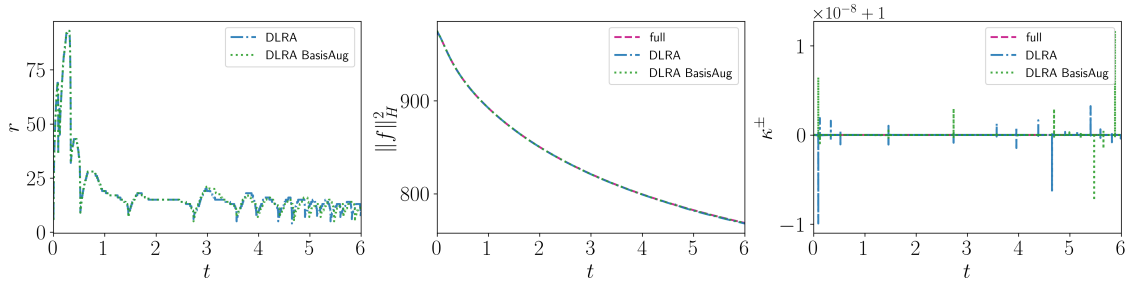


Figure 7.6: Left: Evolution of the rank in time for the 1D tanh problem for the reduced multiplicative DLRA scheme and the basis augmented multiplicative DLRA scheme. **Middle:** Evolution of the $\mathscr{H}$-norm in time for the multiplicative full solver, the reduced multiplicative DLRA scheme, and the basis augmented multiplicative DLRA scheme. **Right:** Evolution of $\kappa^{\pm}$ in time for the multiplicative full solver, the reduced multiplicative DLRA scheme, and the basis augmented multiplicative DLRA scheme. The line corresponding to the full system has the constant value 1.

7.5.3 2D plane source

To highlight the computational advantages of the DLRA scheme, a 2D version of the plane source problem considered in Section 7.5.1 is presented. The corresponding 2D conservative form of the equations established in (7.5) is given by

$$\partial_t g\left(t,\mathbf{x},\mathbf{v}\right) = -\frac{\mathbf{v}}{\rho\left(t,\mathbf{x}\right)}\cdot\nabla_{\mathbf{x}}\left(\rho\left(t,\mathbf{x}\right)g\left(t,\mathbf{x},\mathbf{v}\right)\right)+\sigma\left(1-g\left(t,\mathbf{x},\mathbf{v}\right)\right)-\frac{g\left(t,\mathbf{x},\mathbf{v}\right)}{\rho\left(t,\mathbf{x}\right)}\partial_t\rho\left(t,\mathbf{x}\right),$$
$$\partial_t\rho\left(t,\mathbf{x}\right) = -\frac{1}{2\pi}\nabla_{\mathbf{x}}\cdot\int\rho\left(t,\mathbf{x}\right)g\left(t,\mathbf{x},\mathbf{v}\right)\mathbf{v}e^{-|\mathbf{v}|^2/2}\mathrm{d}\mathbf{v},$$

where $\mathbf{x}=(x,y)\in\Omega_{\mathbf{x}}\subseteq\mathbb{R}^2$ and $\mathbf{v}=(v,w)\in\mathbb{R}^2$. For the numerical experiments, we consider the spatial domain $\Omega_{\mathbf{x}}=[-3,3]\times[-3,3]$. The initial density ρ is chosen to be the cutoff Gaussian

$$\rho\left(t=0,\mathbf{x}\right)=\frac{1}{4\pi}\max\left(10^{-1},\frac{10^2}{4\pi\sigma_{\mathrm{IC}}^2}\exp\left(-\frac{|\mathbf{x}|^2}{4\sigma_{\mathrm{IC}}^2}\right)\right)$$

with constant deviation $\sigma_{\mathrm{IC}}=0.3$. The initial distribution function g is assumed to be constant in space and velocity and we prescribe

$$g\left(t=0,\mathbf{x},\mathbf{v}\right)=1.$$

A large collisionality of $\sigma=100$ is chosen. For the low-rank computations an initial rank of $r=20$ is considered. Computations are performed on a spatial grid with $N_x=N_y=128$ grid cells in both spatial directions. For the velocity grid, $N_v=N_w=32$ grid points are prescribed in both velocity directions. Based on this choice, we obtain $\max_k\left(|\mathbf{v}_k|\right)\approx 10.08$, which is adjusted to the next larger integer. The time step size is determined by $\Delta t=C_{\mathrm{CFL}}\cdot\frac{\Delta x}{11}$ with a CFL number of $C_{\mathrm{CFL}}=0.7$ in order to guarantee numerical stability. We compare the solution of the 2D conservative full system corresponding to (7.17) to the solution obtained from the 2D DLRA scheme corresponding to (7.18). The extension to two dimensions is straightforward.

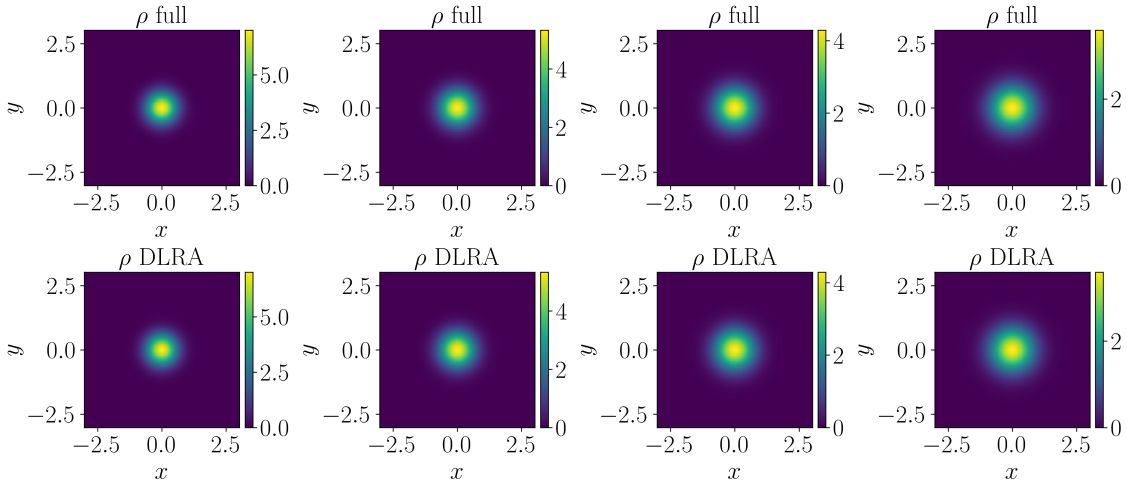


Figure 7.7: Numerical results for the density $\rho(t, \mathbf{x})$ of the 2D plane source problem at time $t = 0$ (first column), $t = 1$ (second column), $t = 2$ (third column, and $t = 3$ (fourth column), computed with the multiplicative full solver (first row) and the reduced multiplicative DLRA scheme (second row).

Figure 7.7 displays the density $\rho(t, x)$ at different times up to $t_{\text{End}} = 3.0$ computed with the multiplicative full solver and the reduced multiplicative DLRA scheme with rank $2r$. Note that we refrain from computations with the basis augmented $4r$ scheme as in two space and velocity dimensions this would lead to extremely increased computational costs while obtaining good agreement also for the reduced multiplicative DLRA scheme with rank $2r$. We observe that at all times the solution of the reduced DLRA scheme matches the solution of the full system. To determine the evolution of the rank, we use a tolerance parameter of $\vartheta = 10^{-5}\|\mathbf{\Sigma}\|_F$. In Figure 7.8 we observe an increase of the rank up to $r = 73$ before it decreases over time. Further, the evolution of the norm $\|\mathbf{f}\|^2_{\mathscr{H}}$ in time is displayed. As expected from 1D theoretical results, its value decreases smoothly over time for all considered systems. In addition, we plot the quantities $\kappa^+ := \max_j \left(\frac{1}{2\pi} \sum_{k=1}^{N_v} \sum_{\ell=1}^{N_w} g(t, \mathbf{x}_j, v_k, w_\ell) \omega_k \omega_\ell e^{(v_k^2 + w_\ell^2)/2} \right)$ and $\kappa^- := \min_j \left(\frac{1}{2\pi} \sum_{k=1}^{N_v} \sum_{\ell=1}^{N_w} g(t, \mathbf{x}_j, v_k, w_\ell) \omega_k \omega_\ell e^{(v_k^2 + w_\ell^2)/2} \right)$. According to 1D theoretical results, it is essential that they are both equal to 1. This property is fulfilled up to order $\mathcal{O}\left(10^{-10}\right)$. For this setup, the computational benefit of the DLRA method compared to the full solver is significant. The scheme is implemented in Julia v1.11 and performed on a MacBook Pro with M1 chip, resulting in a decrease of run time by a factor of approximately 10 from 2315 seconds to 235 seconds, confirming the computational advantages of the DLRA scheme.

7.5.4 2D beam

As a second 2D test example, we consider a beam in the spatial domain $\Omega_{\mathbf{x}} = [-5, 5] \times [-5, 5]$ starting at the point $(0, 0)$ in the middle of the spatial plane and moving to the

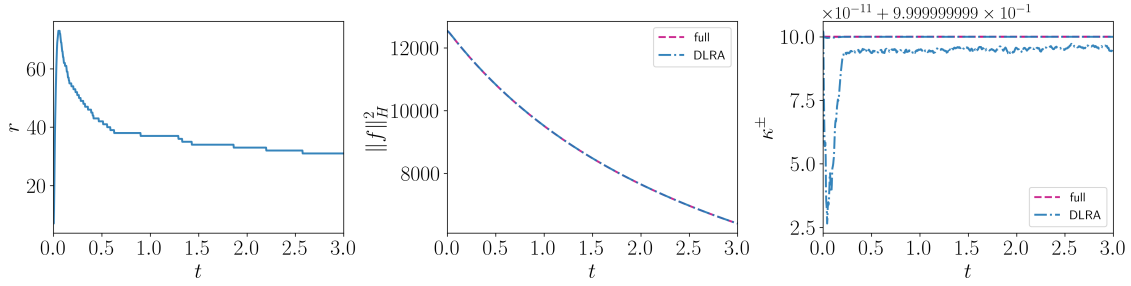


Figure 7.8: Left: Evolution of the rank in time for the 2D plane source problem for the reduced multiplicative DLRA scheme. **Middle:** Evolution of the $\mathscr{H}$-norm in time for the multiplicative full solver and the reduced multiplicative DLRA scheme. **Right:** Evolution of $\kappa^{\pm}$ in time for the multiplicative full solver and the reduced multiplicative DLRA scheme. The line corresponding to the full system has the constant value 1.

bottom left. As initial conditions we prescribe the density ρ to be the cutoff Gaussian

$$\rho\left(t=0,\mathbf{x}\right)=\frac{1}{4\pi}\max\left(10^{-1},\frac{10^{2}}{4\pi\sigma_{\mathrm{IC},\rho}^{2}}\exp\left(-\frac{\left|\mathbf{x}\right|^{2}}{4\sigma_{\mathrm{IC},\rho}^{2}}\right)\right),$$

where $\sigma_{\mathrm{IC},\rho}=0.2$ denotes a constant deviation, and the distribution function g to be

$$g\left(t=0,\mathbf{x},\mathbf{v}\right)=\frac{K}{4\pi}\max\left(10^{-14},\frac{10^{6}}{4\pi\sigma_{\mathrm{IC},g}^{2}}\exp\left(-\frac{\left|\mathbf{v}-\mathbf{v}_{\mathrm{beam}}\right|^{2}}{4\sigma_{\mathrm{IC},g}^{2}}\right)\right)$$

with constant deviation $\sigma_{\mathrm{IC},g}=0.01$ and K being a normalization constant such that the 2D analogue to Lemma 7.4 is fulfilled. The beam velocity $\mathbf{v}_{\mathrm{beam}}$ is set to

$$\mathbf{v}_{\mathrm{beam}}=\begin{pmatrix}-1\\-1\end{pmatrix}$$

and the collisionality to a constant value of $\sigma=1.5$. All other initial settings and computational parameters remain unchanged from the previous test example given in Section 7.5.3.

Figure 7.9 displays the numerical results for the density $\rho\left(t,x\right)$ at different times up to $t_{\mathrm{End}}=3.0$ computed with the multiplicative full solver and the reduced multiplicative DLRA scheme with rank $2r$. At all displayed time steps the DLRA solution captures the solution of the full system. In Figure 7.10 the evolution of the rank in time is shown. We use a tolerance parameter of $\vartheta=10^{-4}\left\|\mathbf{\Sigma}\right\|_{F}$ and allow a maximal rank of $r=200$. Due to the choice of σ, the solution of the problem is not low-rank and a very high rank is required for an accurate approximation. For this reason, we observe an increase of the rank up to the maximal allowed value. Also, the evolution of the norm $\left\|\mathbf{f}\right\|_{\mathscr{H}}^{2}$ in time is illustrated. As expected, it decreases smoothly over time for all considered systems. In addition, we plot the quantities κ^{+} and κ^{-} defined in Section 7.5.3. It is essential that they are both equal to 1. This property is fulfilled up to order $\mathcal{O}\left(10^{-9}\right)$. Due to the high rank, the computational benefits of the DLRA scheme are diminished compared to the previous test case. The scheme is implemented in Julia v1.11 and performed on a MacBook Pro

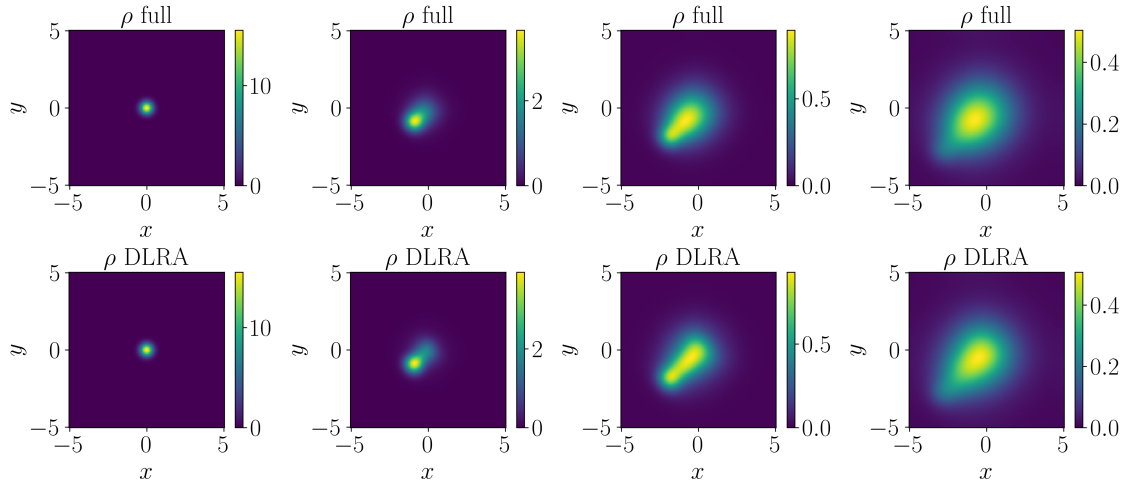


Figure 7.9: Numerical results for the density $\rho(t, \mathbf{x})$ of the 2D beam problem at time $t = 0$ (first column), $t = 1$ (second column), $t = 2$ (third column, and $t = 3$ (fourth column), computed with the multiplicative full solver (first row) and the reduced multiplicative DLRA scheme (second row).

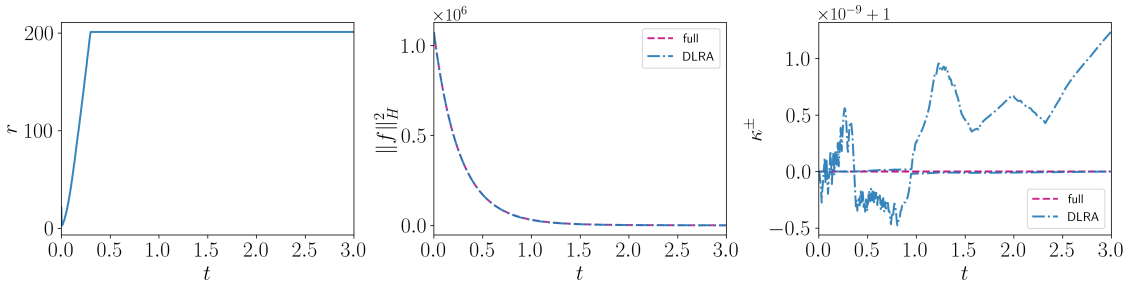


Figure 7.10: Left: Evolution of the rank in time for the 2D beam problem for the reduced multiplicative DLRA scheme. The rank increases up to the maximal allowed value of $r = 200$. **Middle:** Evolution of the $\mathscr{H}$-norm in time for the multiplicative full solver and the reduced multiplicative DLRA scheme. **Right:** Evolution of $\kappa^{\pm}$ in time for the multiplicative full solver and the reduced multiplicative DLRA scheme. The line corresponding to the full system has the constant value 1.

with M1 chip, resulting in a decrease of run time by a factor of approximately 1.5 from 1220 seconds to 845 seconds. This example illustrates the relation between the choice of σ and the low-rank structure of the solution. It is expected that for larger values of σ the solution becomes low-rank and hence the computational benefits of the DLRA scheme are enhanced.

7.6 Summary and conclusion

We have proposed a multiplicative DLRA discretization for the linear Boltzmann-BGK problem that is numerically stable. The main research contributions are:

(i) *A multiplicative splitting of the distribution function:* As the Maxwellian equilibrium distribution M is generally not a low-rank function, we have considered a multiplica-

tive splitting $f = Mg$ of the distribution function. The remaining function g can be considered as a deviation from the equilibrium distribution and in [EHY21] it is shown to be of low rank. For deriving an efficient and stable DLRA scheme the spatial discretization had to be chosen in a conservative form and additional basis augmentations have been required.

(ii) *A stable numerical scheme with rigorous mathematical proofs:* We have shown that a stable discretization had to be carefully derived to obtain a rigorous analytical proof of stability under a specifically designed truncation strategy. A classic hyperbolic CFL condition has been deduced, enabling the choice of an optimal time step size and thereby reducing the computational effort.

(iii) *A rank-adaptive augmented integrator:* We have implemented the rank-adaptive augmented BUG integrator introduced in [CKL22], which is flexible to additional basis augmentations. Compared to the projector-splitting integrator proposed in [LO14], which is used for the non-linear isothermal Boltzmann-BGK equation in [EHY21], this choice allows to adaptively determine the rank in each step, avoiding the a priori determination of a certain fixed rank.

(iv) *Numerical test examples confirming the theoretical properties:* We have presented a number of numerical test examples in both 1D and 2D which validate the stability and the accuracy of the DLRA scheme while showing a significant reduction of computational and memory requirements compared to the full method.

Altogether, the insights gained in this chapter can be helpful for future work as the employed multiplicative splitting is attached to the investigation of more complicated equations, e.g. the non-linear Boltzmann-BGK equation treated in [EHY21], for which we propose to reconsider the chosen discretization in terms of stabilization.

Part II

Application of DLRA to inverse problems

8

Numerical solution of parameter identification inverse problems

Many practical applications involve non-observable quantities that shall be inferred from related observations and measurements. In medical imaging, for example, a classic problem consists in the non-intrusive reconstruction of properties of an examined tissue from measurements [Nat86]. In geophysics, information on the Earth's history is collected from lake and sea sediment analyses [LO84] or subsurface structures are analyzed by seismic imaging for the detection of oil and gas deposits [Nol87]. Image reconstruction and image deblurring techniques allow for the reconstruction of sharp images in projectors and cameras [Gro93, BBM21]. In wave propagation, the characteristics of antennas such as reflective surface mesh shapes are estimated from radiation patterns [BLA86]. A wide variety of further applications can be found in [BK89, Gro93, Kir21, Vog02] and the references therein. All aforementioned settings are examples of *inverse problems*.

In Section 8.1 we give an introduction to the theory of inverse problems. Section 8.2 provides methods for the numerical optimization with PDEs, which can be applied for the reconstruction of unknown parameters.

8.1 Inverse problems

We have already introduced some descriptive examples for inverse problems. The goal of this section consists in considering them from a mathematical point of view. In Section 8.1.1 we formalize the definition of inverse problems. Section 8.1.2 introduces *PDE parameter identification inverse problems*, relating inverse problems and PDEs.

8.1.1 General formulation

We begin with a formal, general definition of direct and inverse problems in a noise-free setting.

Definition 8.1 (Direct and inverse problems, [Gro93, Kir21])**.** Let X and Y be normed vector spaces (typically Banach or Hilbert spaces) and $F : X \to Y$ be an operator acting on a *cause* $x \in X$ so that for the *effect* $y \in Y$ the relation $F(x) = y$ holds. The corresponding problems are classified as follows:

Direct problem:	Given x and F, evaluate $F(x) = y$.
Inverse problem of causation:	Given y and F, solve $F(x) = y$ for x.
Inverse problem of model identification:	Given x and y, solve $F(x) = y$ for F.

Hence, a direct problem consists in evaluating the consequences of a given cause whereas for an inverse problem the unknown cause or unknown model parameters for a given observation must be determined.
Hadamard's concept of *well-posedness* was originally introduced for direct problems in [Had02, Had23]. It directly translates to inverse problems.

Definition 8.2 (Well-posedness and ill-posedness, [Kir21])**.** Let $F : X \to Y$ with X and Y being normed vector spaces. The inverse problem $F(x) = y$ is called *well-posed* if the following properties are satisfied:

(i) There exists a solution to the problem (existence).

(ii) There is at most one solution to the problem (uniqueness).

(iii) The solution continuously depends on the data (stability).

If at least one of the conditions is violated, the corresponding problem is called *ill-posed*.

The first two properties depend directly on the vector spaces X and Y and on the operator F. The stability property relies on the choice of the norms and, following [Bal19], can be considered to be subjective as, according to the setting, different conditions for the acceptance of the translation of errors from the measurements y to the data x are possible.

Regularization of ill-posed problems. While direct problems are usually well-posed, the corresponding inverse problems are often ill-posed [BBM21]. This ill-posedness can be overcome by regularization techniques. Common strategies are *Tikhonov regularization*, filtering and truncated *SVD approaches*, iterative methods such as *Landweber iteration* or the *conjugate gradient method*, stochastic models such as *Bayesian inversion*, and *projection methods*. A precise description of regularization and related techniques can be found in [Lou89, Kir21, Gro93].

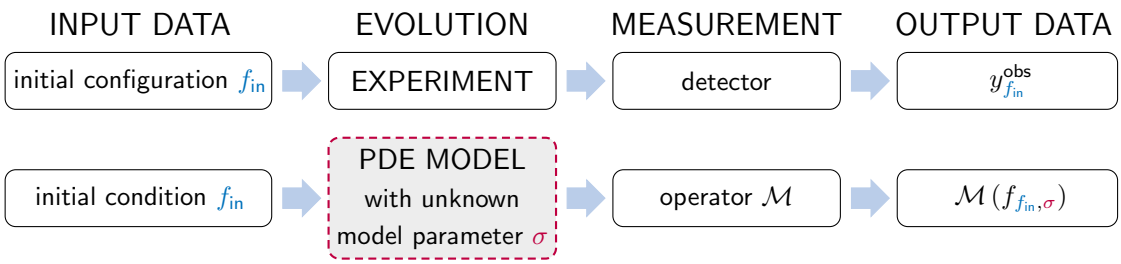


Figure 8.1: Illustration of the structure of a parameter identification inverse problem. The first row is related to real-world experiments. The second row is related to the mathematical simulation using a PDE model. Controllable (known) quantities are depicted in blue. Unknown quantities are depicted in purple. The PDE model, which is known except for the parameter σ, is depicted in gray.

8.1.2 PDE parameter identification

This thesis is concerned with inverse problems related to model identification, specifically *parameter identification inverse problems* in physical processes that are described using kinetic PDEs. Let σ be the unknown model parameter and $F : U \to Y$ be a forward operator with $F(\sigma) = y$. Usually, σ is chosen from an *admissible parameter set* $U_{\text{ad}} \subseteq U$ that incorporates physically motivated constraints on the model parameter [BK89]. The problem is equipped with input data f_{in} from the initial configuration, to which we account by denoting $F_{f_{\text{in}}}(\sigma) = y_{f_{\text{in}}}$. The output data $y_{f_{\text{in}}}$ is assumed to be known from observations or measurements. Then the parameter identification inverse problem reads:

$$\text{Given } f_{\text{in}} \text{ and } y_{f_{\text{in}}}, \qquad \text{solve } F_{f_{\text{in}}}(\sigma) = y_{f_{\text{in}}} \qquad \text{for } \sigma.$$

On a mathematical level, the evolution characterized by the forward operator is described by a kinetic PDE model depending on the unknown model parameter σ. The PDE model is equipped with an initial condition f_{in} and we denote $f_{f_{\text{in}},\sigma}$ for the solution of the PDE. The output data is generated by measurements of the distribution function $\mathcal{M}(f_{f_{\text{in}},\sigma})$, where $\mathcal{M}$ denotes a measurement operator, which is assumed to be known. As shown in Figure 8.1, one seeks to align the experimental output data $y^{\text{obs}}_{f_{\text{in}}}$ which is obtained from detector observations in real-world experiments and the synthetic data $\mathcal{M}(f_{f_{\text{in}},\sigma})$ generated from the mathematical PDE model. The illustration is inspired by the one given in [Hel25].

Measurement models and noise. For simplicity, in the above setting we assume that, both in real-world experiments and in the PDE model, the output data can be perfectly derived from the measurements taken with the detector and from the measurement operator $\mathcal{M}$, respectively. For practical applications this is clearly not realistic since all physical measurements are affected by errors. A common choice consists in adding *statistical noise of Gaussian type* to the output data [Tar05]. Also the construction of the measurement operator $\mathcal{M}$ and the choice of an optimal set of experiments are challenging questions which are subject to current research. More information can be found for example in the review articles [Ren10, HJM24, Ale21].

Output least squares minimization. A common strategy to align the results from the experimental observations and the computed solutions from the PDE model is the *output least squares minimization* [Vog02, Gro93, BK89]. In this approach, one solves the optimization problem

$$\arg\min_{\sigma\in U} J(\sigma) \qquad \text{with} \quad J(\sigma) = \frac{1}{2}\left\|\mathcal{M}(f_{f_{\text{in}},\sigma}) - y_{f_{\text{in}}}^{\text{obs}}\right\|_Y^2, \tag{8.1}$$

where $f_{f_{\text{in}},\sigma}$ is generated from the solution of the PDE with the considered parameter value σ. The functional J measures the mismatch between the observed data $y_{f_{\text{in}}}^{\text{obs}}$ and the computed data $\mathcal{M}(f_{f_{\text{in}},\sigma})$ that is obtained when solving the PDE. Note that in (8.1) usually a regularizing penalty term is added. We refrain from this but emphasize that this extension is straightforward. In addition, we do not pose additional physically motivated constraints on the parameter σ and the PDE solution $f_{f_{\text{in}},\sigma}$ but emphasize their importance for showing theoretical results as for instance done in [HPUU08].

8.2 Numerical optimization with PDEs

In this section, techniques for solving optimization problems such as (8.1) are presented. We consider the following general non-linear minimization problem

$$\arg\min_{(f,\sigma)\in X\times U} \widetilde{J}(f,\sigma) \qquad \text{subject to} \quad G(f,\sigma) = 0, \tag{8.2}$$

where $\widetilde{J} : X \times U \to \mathbb{R}_0^+$ and $G : X \times U \to Z$ are continuous with Banach space Z and reflexive Banach spaces X and U. The *state variable* f is dependent on the *control variable* σ through the equation $G(f,\sigma) = 0$, which is also called the *state equation*. Section 8.2.1 reformulates the optimization problem (8.2) in a reduced form and introduces the *adjoint state method*, which allows for an efficient calculation of the gradient in gradient-based iterative solution schemes. Section 8.2.2 considers the optimization parameters, for which the size of the parameter space can be significantly reduced by *spline approximation*. In Section 8.2.3 the *gradient descent method* for the reduced minimization problem with spline-approximated optimization parameters is given.

8.2.1 Adjoint state method for a gradient-based solution

For an efficient gradient-based iterative solution of (8.2) the *adjoint state method* can be applied. We first recall the following theoretical background.

Definition 8.3 (Lagrangian, [HPUU08])**.** The *Lagrange function* or *Lagrangian* $\mathfrak{L} : X \times U \times Z^* \to \mathbb{R}$ for the minimization problem (8.2) is defined by

$$\mathfrak{L}(f,\sigma,g) = \widetilde{J}(f,\sigma) + \langle g, G(f,\sigma)\rangle_{Z^*,Z}, \tag{8.3}$$

where Z^* denotes the dual space of Z and $\langle \cdot, \cdot \rangle_{Z,Z^*}$ the corresponding dual pairing. The quantity $g \in Z^*$ is called a *Lagrange multiplier* or an *adjoint state*.

The Lagrangian formulation allows to rewrite constrained optimization problems so that methods from unconstrained optimization can be applied. Additional information can be found in standard textbooks on numerical optimization such as [NW06, UU12] and more specifically for numerical optimization in PDE settings in [BS00, Trö10]. Further, we generalize the notion of differentiability, allowing to consider continuous infinite-dimensional optimization settings.

Definition 8.4 (Fréchet differentiability, [HPUU08])**.** Let $F : U \subseteq X \to Y$ be an operator between Banach spaces X and Y and let $U \subseteq X$ be a non-empty open subset. The operator F is called *Fréchet differentiable at* $x \in U$ if there exists a linear and bounded operator $F'(x) \in \mathcal{L}(X, Y)$ such that

$$\left\| F(x+h) - F(x) - F'(x)\, h \right\|_Y = o\left(\|h\|_X \right) \qquad \text{for} \quad \|h\|_X \to 0.$$

If F is Fréchet differentiable at every $x \in V$ with $V \subseteq U$ open, F is called *Fréchet differentiable on* V and $F' : V \to \mathcal{L}(X, Y)$, $x \mapsto F'(x)$ is called the *Fréchet derivative of* F *on* V.

For the derivation of the adjoint state method we follow the explanations in [HPUU08, Ple06]. Let $\widetilde{J}$ and G be continuously Fréchet differentiable and the solution operator $\sigma \in U \mapsto f(\sigma) = f_\sigma \in X$ be uniquely defined and continuously Fréchet differentiable. We first rewrite the full constrained optimization problem (8.2) as its corresponding *reduced problem* by inserting $f(\sigma)$. We obtain the formulation

$$\min_{\sigma \in U} \; J(\sigma) := \widetilde{J}(f(\sigma), \sigma), \tag{8.4}$$

for which $J : U \to \mathbb{R}_0^+$ is called the *reduced functional*. In order to apply a gradient-based optimization procedure, we are interested in the computation of its Fréchet derivative $J'(\sigma)$. A direct evaluation gives

$$J'(\sigma) = f'(\sigma)^* \, \partial_f \widetilde{J}(f(\sigma), \sigma) + \partial_\sigma \widetilde{J}(f(\sigma), \sigma), \tag{8.5}$$

where $f'(\sigma)^*$ denotes the adjoint of $f'(\sigma)$. From a numerical point of view the computation of $f'(\sigma)$ tends to be quite expensive and we seek a more sophisticated approach. We insert $f(\sigma)$ into the Lagrangian (8.3) for the minimization problem (8.2). This yields

$$\mathfrak{L}(f(\sigma), \sigma, g) = \widetilde{J}(f(\sigma), \sigma) + \langle g, G(f(\sigma), \sigma) \rangle_{Z^*,Z} \tag{8.6}$$

with arbitrary $g \in Z^*$. On the solution manifold with $f = f(\sigma)$ the state equation is fulfilled and we obtain the equality

$$J(\sigma) = \widetilde{J}(f(\sigma), \sigma) = \widetilde{J}(f(\sigma), \sigma) + \langle g, G(f(\sigma), \sigma) \rangle_{Z^*,Z} = \mathfrak{L}(f(\sigma), \sigma, g).$$

Differentiating this expression with respect to σ leads to

$$J'(\sigma) = \partial_f \mathfrak{L}(f(\sigma), \sigma, g) f'(\sigma) + \partial_\sigma \mathfrak{L}(f(\sigma), \sigma, g). \tag{8.7}$$

To avoid the computation of $f'(\sigma)$ appearing in the first term, we choose a special $g_\sigma \in Z^*$ such that

$$\partial_f \mathfrak{L}(f(\sigma), \sigma, g_\sigma) = 0. \tag{8.8}$$

From (8.6) we can conclude that this is exactly the case if

$$\partial_f G(f(\sigma), \sigma)^* g_\sigma = -\partial_f \widetilde{J}(f(\sigma), \sigma).$$

This equation is called the *adjoint equation*. With the special choice g_σ, we obtain from (8.7) that the Fréchet derivative of $J(\sigma)$ can be computed as

$$J'(\sigma) = \partial_\sigma \mathfrak{L}(f(\sigma), \sigma, g_\sigma) = \partial_\sigma \widetilde{J}(f(\sigma), \sigma) + \partial_\sigma G(f(\sigma), \sigma)^* g_\sigma. \tag{8.9}$$

Numerically, the evaluation of this expression is usually much less expensive than the direct evaluation of (8.5).

Summary of the adjoint state method. We summarize the adjoint state method for an efficient computation of $J'(\sigma)$ as follows:

(i) Set up the Lagrangian $\mathfrak{L}$ for the problem as done in (8.3).

(ii) Compute the adjoint state g_σ by solving (8.8).

(iii) Compute $J'(\sigma)$ by evaluating (8.9).

8.2.2 Spline approximation of the optimization parameters

In the optimization problem (8.2) and its reduced formulation (8.4) the function σ is chosen from a reflexive Banach space U. Let us denote $\sigma = \sigma(x)$ and assume that for a numerical solution a relatively fine grid in the spatial variable x is prescribed. When evaluating the scattering coefficient $\sigma(x)$ at each point of the spatial grid and taking these values as the parameters to be optimized, there are several computational disadvantages. For instance, a huge parameter space is obtained and very rough functions are part of the ansatz space. To avoid this, we consider the parametrization of $\sigma(x)$ by *splines*. A lot of profound literature on splines is available [dB78, Sch07, Sch15, HH13] but the topic is also covered in several introductory textbooks on numerical analysis such as [SB02, QSS02]. We restrict our considerations to a 1D setting and consider the spatial domain $\Omega_x = [a, b]$. Let $\Delta = \{a = \tau_0 < \tau_1 < ... < \tau_{N_c} = b\}$ be a partition of the interval $[a, b]$ with $N_c + 1$ pairwise different knots.

Definition 8.5 (Spline (function), [QSS02])**.** A *spline (function)* $s_k(x)$ of degree k on Δ is a function $s_k : [a,b] \to \mathbb{R}$ with the following properties:

(i) $s_k(x) \in C^{k-1}[a,b]$, i.e. the function $s_k(x)$ is $k-1$ times continuously differentiable on the interval $[a,b]$.

(ii) On every subinterval $[\tau_i, \tau_{i+1}], i = 0, ..., N_c - 1$, the function $s_k(x)$ coincides with a polynomial of degree at most k, i.e. $s_k(x)\big|_{[\tau_i,\tau_{i+1}]} \in \mathbb{P}_k$.

The set of all spline functions of degree k on Δ is denoted by $S_{k,\Delta}$.

The properties given in the definition are not sufficient to uniquely characterize a spline function of degree k. It can be shown that $N_c + k$ degrees of freedom are left and thus $\dim(S_{k,\Delta}) = N_c + k$ [QSS02]. Finding a suitable basis representation for $s_k(x)$ is crucial for numerical applications as intuitive choices can for instance lead to ill-conditioning or require a large number of numerical operations for the evaluation of $s_k(x)$ [Cox72, Sch07]. We introduce the following set of spline functions.

Definition 8.6 (B-spline (function), [QSS02, SB02])**.** The normalized *B-spline (function)* $B_{i,k+1}(x)$ of degree k on Δ is defined as

$$B_{i,k+1}(x) = (\tau_{i+k+1} - \tau_i)\, h[\tau_i, ..., \tau_{i+k+1}],$$

where

$$h(z) = (z-x)_+^k = \begin{cases} (z-x)^k & \text{for } z \geq x, \\ 0 & \text{for } z < x, \end{cases} \qquad z \in \mathbb{R},$$

and $h[\tau_i, ..., \tau_{i+k+1}]$ are the *divided differences* of the real function h, which are recursively defined by $h[\tau_i] = h(\tau_i)$ and

$$h[\tau_i, ..., \tau_{i+k+1}] = \frac{h[\tau_{i+1}, ..., \tau_{i+k+1}] - h[\tau_i, ..., \tau_{i+k}]}{\tau_{i+k+1} - \tau_i}.$$

A recursive computation of the normalized B-spline functions is possible through the following recursion formula.

Lemma 8.7 (Cox-de Boor recursion formula, [Cox72, dB72])**.** *The normalized B-spline $B_{i,k+1}(x)$ of degree k on Δ can be obtained from the recursion*

$$B_{i,1}(x) = \begin{cases} 1 & \text{if} \quad x \in [\tau_i, \tau_{i+1}], \\ 0 & \text{else}, \end{cases}$$

$$B_{i,k+1}(x) = \frac{x - \tau_i}{\tau_{i+k} - \tau_i} B_{i,k}(x) + \frac{\tau_{i+k+1} - x}{\tau_{i+k+1} - \tau_{i+1}} B_{i+1,k}(x), \qquad k \geq 1.$$

Proof. See for instance [SB02]. □

The normalized B-spline functions exhibit useful properties making them well-suited for numerical applications.

Lemma 8.8 (Properties of B-splines, [SB02])**.** *The normalized B-spline functions introduced in Definition 8.6 fulfill the following properties:*

(i) $B_{i,k+1}(x) = 0 \qquad$ *for* $x \notin [\tau_i, \tau_{i+k+1}]$,

(ii) $B_{i,k+1}(x) > 0 \qquad$ *for* $\tau_i < x < \tau_{i+k+1}$

(iii) For all x with $\inf\{\tau_i\} < x < \sup\{\tau_i\}$ it holds $\sum_i B_{i,k+1}(x) = 1$, and the sum contains only finitely many non-zero terms.

Proof. See for instance [SB02]. □

From the recursion formula given in Lemma 8.7 we can conclude that with respect to the partition Δ only $N_c - k$ linearly independent normalized B-splines of order k can be constructed. This can be overcome by considering an *extended partition* Δ_{ext}, i.e. by adding $2k$ knots such that

$$\begin{aligned} &\tau_{-k} \leq \tau_{-k+1} \leq ... \leq \tau_{-1} \leq \tau_0 = a, \\ &b = \tau_{N_c} \leq \tau_{N_c+1} \leq ... \leq \tau_{N_c+k}. \end{aligned} \tag{8.10}$$

Then the normalized B-splines $B_{i,k+1}(x)$ for $i = -k, ..., -1$ and $i = N_c - k, ..., N_c - 1$ can also be constructed and we obtain a unique basis representation of $s_k(x)$ in terms of normalized B-splines.

Theorem 8.9 (B-spline basis of $S_{k,\Delta}$, [Sch07])**.** *The normalized B-spline basis functions $B_{i,k+1}(x)$ of degree k on the extended partition Δ_{ext} constitute a basis of $S_{k,\Delta}$, i.e. for each spline $s_k(x) \in S_{k,\Delta}$ there exists a unique representation*

$$s_k(x) = \sum_{i=-k}^{N_c-1} c_i B_{i,k+1}(x) \qquad \text{with} \quad c_i \in \mathbb{R},$$

and the real numbers c_i with $i = -k, ..., N_c - 1$ are called the B-spline coefficients.

Proof. See for instance [Sch07]. □

Periodic splines. In many practical applications periodic functions are of importance. For these, a suitable approximation can be derived in terms of *periodic splines*.

Definition 8.10 (Periodic spline (function), [Sch07])**.** A spline function $s_k(x) \in S_{k,\Delta}$ of degree k on Δ which satisfies

$$s_k^{(j)}(a) = s_k^{(j)}(b) \qquad \text{for} \quad j = 0, 1, ..., k-1$$

is called a *periodic spline (function)*. The set of all periodic spline functions of degree k on Δ is denoted by $\mathring{S}_{k,\Delta}$.

For the representation of periodic spline functions with normalized B-splines we set the $2k$ additional knots from (8.10) to be periodically extended such that

$$\tau_{-i} = \tau_{N_c-i} - b + a \qquad \text{and} \qquad \tau_{N_c+i} = \tau_i + b - a \qquad \text{for} \quad i = 1, ..., k. \tag{8.11}$$

Then we can give the definition of *periodic B-splines*. We restrict the definition to the case $N_c > k$ but an extension for $N_c \leq k$ is also possible [Sch07].

Definition 8.11 (Periodic B-spline (function), [Sch07])**.** With the periodically extended knots given in (8.11) the normalized *periodic B-spline (function)* $\mathring{B}_{i,k+1}(x)$ of degree k on Δ is defined as

$$\mathring{B}_{i,k+1}(x) = B_{i,k+1}(x) \qquad \text{for} \quad i = 0, ..., N_c - 1 - k$$

and

$$\mathring{B}_{i,k+1}(x) = \begin{cases} B_{i,k+1}(x) & \text{for} \quad a \leq x < \tau_{k+1+i}, \\ B_{N_c+i,k+1}(x) & \text{for} \quad \tau_{k+1+i} \leq x \leq b, \end{cases} \qquad \text{for} \quad i = -k, ..., -1.$$

Having the normalized periodic B-spline functions at hand, we can use them to express a periodic spline $s_k \in \mathring{S}_{k,\Delta}$.

Theorem 8.12 (Periodic B-spline basis of $\mathring{S}_{k,\Delta}$, [Sch07])**.** *The normalized periodic B-spline basis functions $\mathring{B}_{i,k+1}(x)$ of degree k on the extended partition Δ_{ext} with knots as given in (8.11) constitute a basis of $\mathring{S}_{k,\Delta}$, i.e. for each spline $s_k(x) \in \mathring{S}_{k,\Delta}$ there exists a unique representation*

$$s_k(x) = \sum_{i=-k}^{N_c-1-k} c_i \mathring{B}_{i,k+1}(x) \qquad \textit{with} \quad c_i \in \mathbb{R}. \tag{8.12}$$

Proof. See for instance [Sch07]. □

With the definition of the periodic B-spline functions we can also rewrite the basis representation (8.12) of $s_k(x)$ with non-periodic normalized B-splines.

Corollary 8.13 (B-spline basis of $\mathring{S}_{k,\Delta}$, [Sch07])**.** *The unique representation of a periodic spline $s_k(x) \in \mathring{S}_{k,\Delta}$ given in (8.12) can be equivalently written in terms of non-periodic normalized B-splines as*

$$s_k(x) = \sum_{i=-k}^{N_c-1} c_i B_{i,k+1}(x) \qquad \textit{with} \quad c_{N_c-1-i} = c_{-i-1} \qquad \textit{for} \quad i = 0, ..., k-1.$$

Proof. See for instance [Sch07]. □

Spline interpolation. A common field of application of spline functions is *spline interpolation*. Let $\sigma(x)$ be a prescribed function that can be evaluated on the considered interval

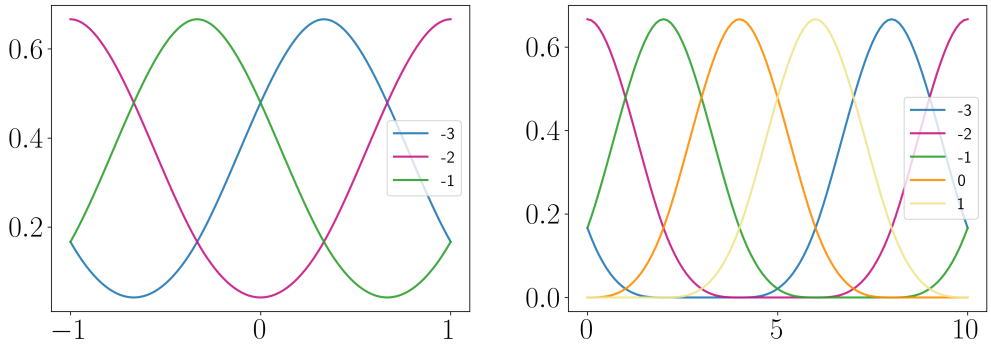


Figure 8.2: Normalized cubic periodic B-spline basis functions $\mathring{B}_{i,4}(x)$ for $i = -3, ..., N_c - 4$ for $N_c = 3$ (left) and $N_c = 5$ (right) on different spatial domains.

$\Omega_x = [a, b]$ and denote $\sigma_0 = \sigma(\tau_0)$, $\sigma_1 = \sigma(\tau_1)$, ..., $\sigma_{N_c} = \sigma(\tau_{N_c})$. An *interpolating spline function* $s_k(x) \in S_{k,\Delta}$ of degree k on Δ additionally satisfies

$$s_k(\tau_i) = \sigma_i \qquad \text{for} \quad i = 0, ..., N_c - 1.$$

For spline interpolation, especially *cubic splines* play an important role as they provide C^2-approximations, which are particularly useful in practical applications. For more information on (cubic) spline interpolation we refer to literature such as [dB78, SB02]. Coming back to the model parameter function $\sigma(x) \in U$ of the optimization problem (8.2) and its reduced formulation (8.4), it can be shown that for U being the Sobolev space $W^{q,2}$ a suitable interpolation of $\sigma(x)$ with cubic B-splines can be given. Corresponding error estimates can be found in [BK89].

Spline approximation of σ. In this thesis, we assume that the model parameter function is well-approximated by a representation with normalized cubic periodic B-splines with equally spaced knots, i.e. we set

$$\sigma(x) \approx \sum_{i=-3}^{N_c-4} c_i \mathring{B}_{i,4}(x) \qquad \text{with} \quad \mathbf{c} = (c_i) \in \mathbb{R}^{N_c}. \tag{8.13}$$

This assumption is justified by considering for instance a suitable interpolation of a given parameter function $\sigma(x)$. The set of normalized cubic periodic B-splines for $N_c = 3$ and for $N_c = 5$ on different spatial domains is illustrated in Figure 8.2. The reduced optimization problem (8.4) is solved on the smaller parameter space $\mathbb{R}^{N_c}$ of dimension N_c and translates to

$$\underset{\mathbf{c}\in\mathbb{R}^{N_c}}{\arg\min} \; J(\mathbf{c}) = \widetilde{J}(f(\mathbf{c}), \mathbf{c}), \tag{8.14}$$

for which $J : \mathbb{R}^{N_c} \to \mathbb{R}_0^+$ is called the *cost function*.

8.2.3 Gradient descent method

For the solution of the reduced optimization problem (8.14) with cost function $J \in C^1\left(\mathbb{R}^{N_c}\right)$ a gradient-based approach is pursued. A simple method consists in applying the *gradient descent method*, which is a standard technique in unconstrained numerical optimization and has been extensively treated in literature, e.g. in [GK99, Nes04, NW06, WR22]. We begin with the definition of a *descent direction*.

Definition 8.14 (Descent direction, [WR22])**.** A vector $\mathbf{d} \in \mathbb{R}^{N_c}$ is called a *descent direction* of J at $\mathbf{c}$ if, for all $\eta > 0$ sufficiently small, it holds

$$J\left(\mathbf{c} + \eta \mathbf{d}\right) < J\left(\mathbf{c}\right).$$

Let $\mathbf{c}^0 \in \mathbb{R}^{N_c}$ be a starting value of the optimization procedure. Then the gradient descent scheme generates an iterative sequence according to

$$\mathbf{c}^{n+1} = \mathbf{c}^n - \eta^n \nabla_{\mathbf{c}} J\left(\mathbf{c}^n\right) \qquad \text{for} \quad n = 0, 1, ..., \tag{8.15}$$

where $\eta^n > 0$ denotes an adaptively chosen step size. The descent direction $\mathbf{d}^n = -\nabla_{\mathbf{c}} J\left(\mathbf{c}^n\right)$ is the direction of steepest descent, explaining why this method is also called the *steepest descent method*. Without additional assumptions on the cost function J convergence of the scheme (8.15) cannot be guaranteed. We introduce the following concepts.

Definition 8.15 (L-smoothness and m-strong convexity, [WR22])**.** The function $J \in C^1\left(\mathbb{R}^{N_c}\right)$ is called *L-smooth* if its gradient is Lipschitz continuous, i.e. if there exists a constant $L \geq 0$ such that for all $\mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}^{N_c}$ it holds

$$\left\|\nabla J_{\mathbf{c}}\left(\mathbf{c}_1\right) - \nabla J_{\mathbf{c}}\left(\mathbf{c}_2\right)\right\|_E \leq L \left\|\mathbf{c}_1 - \mathbf{c}_2\right\|_E,$$

where $\|\cdot\|_E$ denotes the Euclidean norm. The function $J \in C^1\left(\mathbb{R}^{N_c}\right)$ is called *m-strongly convex* if there exists a constant $m > 0$ such that for all $\mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}^{N_c}$ it holds

$$J_{\mathbf{c}}\left(\mathbf{c}_2\right) \geq J_{\mathbf{c}}\left(\mathbf{c}_1\right) + \nabla J_{\mathbf{c}}\left(\mathbf{c}_1\right)^{\top}\left(\mathbf{c}_2 - \mathbf{c}_1\right) + \frac{m}{2}\left\|\mathbf{c}_2 - \mathbf{c}_1\right\|_E^2.$$

The constant m is called the *modulus of convexity*.

Then, convergence of the gradient descent method to a unique global minimum $\mathbf{c}^* \in \mathbb{R}^{N_c}$ can be established.

Theorem 8.16 (Convergence of the gradient descent method, [Nes04])**.** *Let $J \in C^1\left(\mathbb{R}^{N_c}\right)$ be L-smooth and m-strongly convex. Then, for $\eta^n \equiv \eta \leq \frac{2}{m+L}$, the gradient descent method (8.15) generates a sequence $\{\mathbf{c}^n\}$ such that*

$$\left\|\mathbf{c}^n - \mathbf{c}^*\right\|_E^2 \leq \left(1 - \frac{2\eta m L}{m + L}\right)^n \left\|\mathbf{c}^0 - \mathbf{c}^*\right\|_E^2,$$

where $\mathbf{c}^* \in \mathbb{R}^{N_c}$ denotes the unique global minimum. Optimal convergence is obtained for $\eta = \frac{2}{m+L}$.

Proof. See for instance [Nes04]. □

L-smoothness and m-strong convexity are strong restrictions on the cost function J. The assumption of m-strong convexity can be weakened when imposing the *Polyak-Łojasiewicz inequality* instead. Further reading can be found in [Pol63, KNS16].

Step size strategies. There are different strategies on how to determine the step size η^n in each step of iterative descent schemes. We focus on *line search methods* that determine the length of the step size when the descent direction $\mathbf{d}^n$ is given. Often, a step size is considered acceptable if the *Armijo condition*

$$J\left(\mathbf{c}^n + \eta^n \mathbf{d}^n\right) \leq J\left(\mathbf{c}^n\right) + h_1 \eta^n \nabla_{\mathbf{c}} J\left(\mathbf{c}^n\right)^\top \mathbf{d}^n \tag{8.16}$$

with constant $h_1 \in (0,1)$ is satisfied. To find an appropriate step size in practical applications, one can choose an initial guess $\bar{\eta}$ and a step size reduction factor $p \in (0,1)$ and determine the first value in the sequence $\bar{\eta}, p\bar{\eta}, p^2\bar{\eta}, p^3\bar{\eta}, \ldots$ such that condition (8.16) is fulfilled. This approach is called *backtracking line search* with Armijo condition [Arm66]. Alternative step size conditions are for instance the (strong) *Wolfe conditions*, the *Goldstein conditions* or methods involving the Hessian if the cost function J is in $C^2\left(\mathbb{R}^{N_c}\right)$. More information on step size strategies is available in [NW06, WR22].

Alternative gradient-based iterative methods. There are various alternatives to the simple gradient descent method described in (8.15), exhibiting faster convergence properties or requiring less computational effort. *Stochastic gradient descent methods* reduce the computational complexity when the evaluation of multiple gradients instead of only one gradient is required. For twice continuously differentiable cost functions the convergence can be improved by taking evaluations of the Hessian into account. This leads to *Newton-* and *quasi-Newton methods*. *Conjugate gradient methods* are especially useful for the solution of large linear and non-linear systems of equations. A more detailed description of alternative gradient-based iterative methods is given in [NW06, GK99].

9

An adaptive DLRA optimizer for parameter identification inverse problems

A classic inverse problem arising in medical imaging is the reconstruction of properties of an examined tissue from measurements without doing harm to the human body. We consider a model using the time-dependent *radiative transfer equation (RTE)* with an unknown *scattering coefficient* incorporating properties of the background medium. The associated inverse problem considers the reconstruction of the scattering coefficient from measurements. This parameter identification inverse problem shall be solved by using PDE constrained optimization. Similar to recent papers [LWY23, CLL18, HKLT25, ELWY24], we deploy a gradient-based approach for which in each iteration the evaluation of both the forward and the adjoint problem is required. Obviously, this can numerically become very costly, especially in higher-dimensional settings, which is overcome by using a *dynamical low-rank approximation (DLRA)* approach. We pursue the following strategy: "first optimize, then discretize, then low-rank", i.e. we first perform the optimization in a continuous setting before the resulting equations are discretized and the method of DLRA is applied.

The structure of this chapter is as follows. Section 9.1 recalls the 1D formulation of the RTE and the associated inverse transport problem. In Section 9.2 we apply a PDE constrained optimization procedure for the solution of the parameter identification inverse problem and derive the adjoint equations using a Lagrangian reformulation. For the scattering coefficient a normalized cubic periodic B-spline approximation is introduced and a gradient descent step for updating its coefficients is formulated. Section 9.3 is devoted to the discretization of the forward and the adjoint equations as well as of the gradient in angle, space, and time, leading to a fully discrete gradient descent scheme. In Section 9.4 the method of DLRA is applied to the forward and adjoint equations and a backtracking line search method for an adaptive refinement of the gradient descent step size and the DLRA rank tolerance is presented. Numerical results given in Section 9.5 confirm the accuracy and efficiency of the DLRA scheme compared to the solutions computed with a full solver. Finally, Section 9.6 provides a brief summary and conclusion. The results of this chapter closely follow the presentation in [BEKK25a].

9.1 Radiative transfer equation

In optical tomography, the propagation of near-infrared light through tissue can be modeled by using the RTE [RBH07, KNBH02, KH02]. Neglecting boundary effects, the time-dependent form of this kinetic PDE can be given in 1D slab geometry as

$$\begin{cases} \partial_t f(t,x,\mu) + \mu \partial_x f(t,x,\mu) & = \sigma(x)\left(\frac{1}{|\Omega_\mu|}\langle f(t,x,\mu)\rangle_\mu - f(t,x,\mu)\right), \\ f(t=0,x,\mu) & = f_{\text{in}}(x,\mu), \end{cases} \tag{9.1}$$

where $f(t,x,\mu) : \mathbb{R}_0^+ \times \Omega_x \times \Omega_\mu \to \mathbb{R}_0^+$ denotes the distribution function which describes the repartition of photons in phase space. Here, t stands for the time variable, $x \in \Omega_x \subseteq \mathbb{R}$ for the space variable and $\mu \in \Omega_\mu = [-1,1]$ for the angular variable. An integration over the corresponding domain is denoted by brackets $\langle\cdot\rangle$ and $|\Omega_\mu|$ measures the length of the domain Ω_μ. The function $\sigma(x)$ represents the properties of the background medium, indicating the probability of particles at position x to be scattered into a new direction. We refer to it as the *scattering coefficient*. At the initial time $t = 0$ the function $f_{\text{in}}(x,\mu)$ is prescribed for the distribution function. The inverse problem associated with the RTE presented in (9.1) consists in reconstructing the scattering coefficient $\sigma(x)$ from output data that is generated from measurements. For more general information on the inverse transport problem we refer to the review articles [Bal09, Ste03].

9.2 PDE constrained optimization

In practical applications, optical tomography commonly relies on a multitude of measurements from different positions. To be close to realistic settings, we take a number of N_{IC} measurements for the reconstruction of the scattering coefficient $\sigma(x)$ into account. Furthermore, we assume the measurements to be generated by a measurement operator $\mathcal{M}$ acting on the angle-averaged solution of the RTE at the final time $t = T$, which has been computed with the corresponding initial condition $f_{\text{in},m}$ for $m = 1, ..., N_{\text{IC}}$. The restriction to angle-averaged measurements is a common choice for modeling real-world problems [BJ09, CLW18]. For simplicity, the experimentally observed data y_m^{obs} is assumed to be close to the measurements of the angle-averaged solution that is obtained with the ground truth parameter, i.e.

$$y_m^{\text{obs}}(x) \approx \mathcal{M}\left(\langle f_{\sigma,m}(t=T,x,\mu)\rangle_\mu\right) \qquad \text{for} \quad m = 1, ..., N_{\text{IC}},$$

where $f_{\sigma,m}(t,x,\mu)$ denotes a solution of

$$\begin{cases} \partial_t f_m(t,x,\mu) + \mu \partial_x f_m(t,x,\mu) & = \sigma(x)\left(\frac{1}{|\Omega_\mu|}\langle f_m(t,x,\mu)\rangle_\mu - f_m(t,x,\mu)\right), \\ f_m(t=0,x,\mu) & = f_{\text{in},m}(x,\mu), \end{cases} \tag{9.2}$$

computed with the corresponding scattering coefficient $\sigma(x)$. For notational brevity, we refrain from explicitly denoting the dependence of the distribution function and of the experimentally observed data on the initial condition $f_{\text{in},m}$.

For the solution of the PDE parameter identification inverse problem associated with (9.2) one tries to minimize the square loss between the measurements of the angle-averaged distribution function satisfying (9.2) and the experimentally observed data, i.e. one tries to solve the minimization problem

$$\begin{aligned} &\underset{f_1,\dots f_{N_{\text{IC}}},\sigma}{\arg\min} \quad \widetilde{J}(f_1,\dots,f_{N_{\text{IC}}},\sigma) \\ &\text{with} \quad \widetilde{J}(f_1,\dots,f_{N_{\text{IC}}},\sigma) = \frac{1}{2}\sum_{m=1}^{N_{\text{IC}}} \left\langle \left| \mathcal{M}\left(\langle f_m(t=T,x,\mu)\rangle_\mu\right) - y_m^{\text{obs}}(x) \right|^2 \right\rangle_x , \qquad (9.3) \\ &\text{subject to} \quad (9.2). \end{aligned}$$

This can be reformulated in a reduced form by inserting the solution $f_{\sigma,m}$ simulated from (9.2). Then the reduced minimization problem is given by

$$\underset{\sigma}{\arg\min} \; J(\sigma) \quad \text{with} \quad J(\sigma) = \frac{1}{2}\sum_{m=1}^{N_{\text{IC}}} \left\langle \left| \mathcal{M}\left(\langle f_{\sigma,m}(t=T,x,\mu)\rangle_\mu\right) - y_m^{\text{obs}}(x) \right|^2 \right\rangle_x . \qquad (9.4)$$

Note that this setup is close to realistic applications in the sense as described above. For real-word applications we point out that the considered setting with one spatial and one angular variable may not be sufficient. In addition, it is assumed that there is no noise in the measurements, which in practical applications is clearly infeasible. Nevertheless, the results gained from the considered setup provide valuable insights into the combination of parameter identification and DLRA and can be directly extended to higher-dimensional settings.

Section 9.2.1 derives the adjoint equations associated with the forward problem (9.2), before in Section 9.2.2 an explicit gradient descent step is formulated.

9.2.1 Lagrangian formulation and adjoint problem

For the derivation of the adjoint problem as described in Section 8.2.1, we reformulate the PDE constrained minimization problem (9.3) using the method of Lagrange multipliers. We aim for a solution of

$$\underset{f_1,\dots,f_{N_{\text{IC}}},g_1,\dots,g_{N_{\text{IC}}},\lambda_1,\dots\lambda_{N_{\text{IC}}},\sigma}{\arg\min} \quad \mathfrak{L}(f_1,\dots f_{N_{\text{IC}}},g_1,\dots,g_{N_{\text{IC}}},\lambda_1,\dots,\lambda_{N_{\text{IC}}},\sigma),$$

where

$$\mathfrak{L} = \widetilde{J}\left(f_1, ..., f_{N_{\mathrm{IC}}}, \sigma\right) + \sum_{m=1}^{N_{\mathrm{IC}}} \left\langle g_m, \partial_t f_m + \mu \partial_x f_m - \sigma\left(x\right)\left(\frac{1}{|\Omega_\mu|}\left\langle f_m\right\rangle_\mu - f_m\right)\right\rangle_{t,x,\mu} + \sum_{m=1}^{N_{\mathrm{IC}}} \left\langle \lambda_m, f_m\left(t=0,x,\mu\right) - f_{\mathrm{in},m}\left(x,\mu\right)\right\rangle_{x,\mu},$$

and $g_m\left(t,x,\mu\right)$ and $\lambda_m\left(x,\mu\right)$ are the *Lagrange multipliers* with respect to $f_m\left(t,x,\mu\right)$ and $f_{\mathrm{in},m}\left(x,\mu\right)$ for $m = 1, ..., N_{\mathrm{IC}}$, respectively. Applying integration by parts and assuming periodic boundary conditions, the Lagrangian can be rewritten as

$$\begin{aligned}\mathfrak{L} = \widetilde{J}\left(f_1, ..., f_{N_{\mathrm{IC}}}, \sigma\right) &+ \sum_{m=1}^{N_{\mathrm{IC}}} \left\langle f_m, -\partial_t g_m - \mu \partial_x g_m - \sigma\left(x\right)\left(\frac{1}{|\Omega_\mu|}\left\langle g_m\right\rangle_\mu - g_m\right)\right\rangle_{t,x,\mu} \\ &+ \sum_{m=1}^{N_{\mathrm{IC}}} \left\langle g_m(t=T,x,\mu), f_m(t=T,x,\mu)\right\rangle_{x,\mu} \\ &- \sum_{m=1}^{N_{\mathrm{IC}}} \left\langle g_m\left(t=0,x,\mu\right), f_m\left(t=0,x,\mu\right)\right\rangle_{x,\mu} \\ &+ \sum_{m=1}^{N_{\mathrm{IC}}} \left\langle \lambda_m, f_m\left(t=0,x,\mu\right) - f_{\mathrm{in},m}\left(x,\mu\right)\right\rangle_{x,\mu}.\end{aligned}$$

The corresponding *adjoint* or *dual problems* associated with (9.2) can be derived by setting $\partial_{f_m}\mathfrak{L} = 0$ for $m = 1, ..., N_{\mathrm{IC}}$. By straightforward calculation one obtains

$$\begin{cases} -\partial_t g_m\left(t,x,\mu\right) - \mu \partial_x g_m\left(t,x,\mu\right) &= \sigma\left(x\right)\left(\frac{1}{|\Omega_\mu|}\left\langle g_m\left(t,x,\mu\right)\right\rangle_\mu - g_m\left(t,x,\mu\right)\right), \\ g_m\left(t=T,x,\mu\right) &= -\left\langle f_m\left(t=T,x,\mu\right)\right\rangle_\mu + y_m^{\mathrm{obs}}\left(x\right). \end{cases} \tag{9.5}$$

The notation $g_{\sigma,m}\left(t,x,\mu\right)$ indicates that $g_m\left(t,x,\mu\right)$ fulfills equations (9.5). On the solution manifold with $f_{\sigma,m}\left(t,x,\mu\right)$ satisfying the PDE constraints (9.2) and $g_{\sigma,m}\left(t,x,\mu\right)$ satisfying the adjoint equations (9.5) for $m = 1, ..., N_{\mathrm{IC}}$, the following equality holds

$$\mathfrak{L}\left(f_{\sigma,1}, ..., f_{\sigma,N_{\mathrm{IC}}}, g_{\sigma,1}, ..., g_{\sigma,N_{\mathrm{IC}}}, \lambda_1, ..., \lambda_{N_{\mathrm{IC}}}, \sigma\right) = \widetilde{J}\left(f_{\sigma,1}, ..., f_{\sigma,N_{\mathrm{IC}}}, \sigma\right) = J\left(\sigma\right),$$

translating to the derivatives such that

$$\begin{aligned} J'\left(\sigma\right) &= \frac{\mathrm{d}\mathfrak{L}\left(f_{\sigma,1}, ..., f_{\sigma,N_{\mathrm{IC}}}, g_{\sigma,1}, ..., g_{\sigma,N_{\mathrm{IC}}}, \lambda_1, ..., \lambda_{N_{\mathrm{IC}}}, \sigma\right)}{\mathrm{d}\sigma} \\ &= \sum_{m=1}^{N_{\mathrm{IC}}} \left(\frac{\partial \mathfrak{L}}{\partial f_{\sigma,m}}\frac{\partial f_{\sigma,m}}{\partial \sigma} + \frac{\partial \mathfrak{L}}{\partial g_{\sigma,m}}\frac{\partial g_{\sigma,m}}{\partial \sigma} + \frac{\partial \mathcal{L}}{\partial \lambda_m}\frac{\partial \lambda_m}{\partial \sigma} + \frac{\partial \mathfrak{L}}{\partial \sigma}\right) = \sum_{m=1}^{N_{\mathrm{IC}}} \frac{\partial \mathfrak{L}}{\partial \sigma}, \end{aligned} \tag{9.6}$$

where the first three terms vanish since (9.2) and (9.5) are fulfilled. This can be used for an efficient computation of the gradient in the following gradient descent step.

9.2.2 Optimization parameters and gradient descent step

To avoid computational disadvantages from a large optimization parameter space, the function $\sigma(x)$ is approximated using splines. We follow the considerations performed in Section 8.2.2 and use a number N_c of normalized cubic periodic B-spline basis functions $\mathring{B}_{i,4}(x)$ with equally spaced knots as illustrated in Figure 8.2. As given in (8.13), we obtain the representation

$$\sigma(x) \approx \sum_{i=-3}^{N_c-4} c_i \mathring{B}_{i,4}(x) \qquad \text{with} \quad \mathbf{c} = (c_i) \in \mathbb{R}^{N_c}, \tag{9.7}$$

where the vector $\mathbf{c}$ contains the coefficients of the approximation. With prescribed initial values $\mathbf{c}^0 \in \mathbb{R}^{N_c}$ the gradient descent step for the solution of the minimization problem (9.4) updates the coefficients from $\mathbf{c}^n$ to $\mathbf{c}^{n+1}$ in each step by determining

$$\mathbf{c}^{n+1} = \mathbf{c}^n - \eta^n \nabla_{\mathbf{c}} J(\mathbf{c}^n) \qquad \text{for} \quad n = 0, 1, ..., \tag{9.8}$$

where $\eta^n > 0$ denotes an adaptively chosen step size. From expression (9.6) we can derive an explicit formulation for the components of the gradient of the cost function as

$$\frac{\partial J(\mathbf{c})}{\partial c_i} = \sum_{m=1}^{N_{\mathrm{IC}}} \frac{\partial \mathfrak{L}}{\partial \sigma} \frac{\partial \sigma}{\partial c_i} = \sum_{m=1}^{N_{\mathrm{IC}}} \left(-\frac{1}{|\Omega_\mu|} \langle \langle f_{\sigma,m} \rangle_\mu, \langle g_{\sigma,m} \rangle_\mu \rangle_t + \langle f_{\sigma,m}, g_{\sigma,m} \rangle_{t,\mu} \right) \mathring{B}_{i,4}, \tag{9.9}$$

depending on the solutions of the forward and the adjoint equations as well as on the B-spline basis functions and allowing for an efficient computation of the gradient in the gradient descent step (9.8). Note that from now on we write $f_m(t, x, \mu)$ instead of $f_{\sigma,m}(t, x, \mu)$ and $g_m(t, x, \mu)$ instead of $g_{\sigma,m}(t, x, \mu)$.

9.3 Discretization

For the numerical implementation we discretize the forward problem (9.2), the adjoint problem (9.5) and the components of the gradient (9.9) in Sections 9.3.1, 9.3.2 and 9.3.3 in angle, space and time. This leads to a fully discrete scheme. Section 9.3.4 summarizes the fully discrete gradient descent method.

9.3.1 Angular discretization

For the discretization in angle we choose a modal approach making use of the normalized rescaled Legendre polynomials $P_\ell(\mu)$ introduced in Section 3.3.2. This is a standard approach which is commonly used for radiative transfer problems and has also already been adapted in the inverse setting [WSA07]. The normalized rescaled Legendre polynomials $P_\ell(\mu)$ constitute a complete set of orthonormal functions on the interval $[-1, 1]$

and satisfy $\langle P_k(\mu), P_\ell(\mu)\rangle_\mu = \delta_{k\ell}$. We expand the distribution functions $f_m(t,x,\mu)$ and $g_m(t,x,\mu)$ for $m = 1,...,N_{IC}$ in terms of the rescaled Legendre polynomials and obtain the approximations

$$f_m(t,x,\mu) \approx \sum_{\ell=0}^{N_\mu-1} u_{\ell m}(t,x) P_\ell(\mu) \quad \text{and} \quad g_m(t,x,\mu) \approx \sum_{\ell=0}^{N_\mu-1} w_{\ell m}(t,x) P_\ell(\mu), \tag{9.10}$$

where $u_\ell(t,x)$ and $w_\ell(t,x)$ are the corresponding expansion coefficients. We insert these representations into the forward problem (9.2) as well as into the adjoint problem (9.5), multiply with $P_k(\mu)$ and integrate over the angular variable μ. With the matrix $\mathbf{A} \in \mathbb{R}^{N_\mu \times N_\mu}$ defined in (3.26), and the orthonormality condition from above we obtain

$$\begin{cases} \partial_t u_{km}(t,x) & = -\sum_{\ell=0}^{N_\mu-1} \partial_x u_{\ell m}(t,x) A_{k\ell} + \sigma(x) u_{km}(t,x)(\delta_{k0}-1), \\ u_{km}(t=0,x) & = u_{\text{in},km}(x), \end{cases} \tag{9.11}$$

for the forward equations and

$$\begin{cases} -\partial_t w_{km}(t,x) & = \sum_{\ell=0}^{N_\mu-1} \partial_x w_{\ell m}(t,x) A_{k\ell} + \sigma(x) w_{km}(t,x)(\delta_{k0}-1), \\ w_{km}(t=T,x) & = \left(-2u_{0m}(t=T,x) + \sqrt{2} y_m^{\text{obs}}(x)\right)\delta_{k0}, \end{cases} \tag{9.12}$$

for the adjoint equations for $m = 1,...,N_{\text{IC}}$. For the angular discretization of the gradient of the cost function we insert the representations (9.10) into (9.9) and derive

$$\begin{aligned} \frac{\partial J(\mathbf{c})}{\partial c_i} \approx \sum_{m=1}^{N_{\text{IC}}} \Bigg(& -\langle u_{0m}(t,x), w_{0m}(t,x)\rangle_t \\ & + \sum_{k=0}^{N_\mu-1} \langle u_{km}(t,x), w_{km}(t,x)\rangle_t \Bigg) \mathring{B}_{i,4}(x). \end{aligned} \tag{9.13}$$

9.3.2 Spatial discretization

The discretization in the spatial variable is performed on a spatial grid with N_x grid cells and equidistant spacing $\Delta x = \frac{1}{N_x}$. Spatially dependent quantities are approximated at the grid points x_j for $j = 1,...,N_x$ and denoted by

$$\begin{aligned} u_{jkm}(t) &\approx u_{km}(t,x_j), & w_{jkm}(t) &\approx w_{km}(t,x_j), \\ \sigma_j &\approx \sigma(x_j), & y_{jm}^{\text{obs}} &\approx y_m^{\text{obs}}(x_j), & \mathring{B}_{ji,4} &\approx \mathring{B}_{i,4}(x_j). \end{aligned}$$

Assuming periodic boundary conditions, first-order spatial derivatives ∂_x are approximated using the centered FD method. For stability reasons, a diffusion term involving second-order derivatives ∂_{xx} is added. This term is also approximated by the centered FD method. We employ the tridiagonal stencil matrices $\mathbf{D}^x \in \mathbb{R}^{N_x \times N_x}$ given in (3.8) and $\mathbf{D}^{xx} \in \mathbb{R}^{N_x \times N_x}$ defined in (3.11). Recall that the symmetric matrix $\mathbf{A}$ is diagonalizable in the form $\mathbf{A} = \mathbf{Q}\mathbf{M}\mathbf{Q}^\top$ with $\mathbf{Q}$ being orthogonal and $\mathbf{M} = \text{diag}(\sigma_0,...,\sigma_{N_\mu-1})$ and that

we have defined $|\mathbf{A}| = \mathbf{Q}|\mathbf{M}|\mathbf{Q}^\top$. Then the spatially discretized forward equations with centered FD method and an additional second-order FD stabilization term are obtained from (9.11) as

$$\begin{cases} \partial_t u_{jkm}(t) & = -\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{x} u_{i\ell m}(t) A_{k\ell} \\ & \quad + \frac{\Delta x}{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{xx} u_{i\ell m}(t) |A|_{k\ell} + \sigma_j u_{jkm}(t)(\delta_{k0}-1), \\ u_{jkm}(t=0) & = u_{\mathrm{in},jkm}, \end{cases} \tag{9.14}$$

and the spatially discretized adjoint equations from (9.12) as

$$\begin{cases} -\partial_t w_{jkm}(t) & = \sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{x} w_{i\ell m}(t) A_{k\ell} \\ & \quad + \frac{\Delta x}{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{xx} w_{i\ell m}(t) |A|_{k\ell} + \sigma_j w_{jkm}(t)(\delta_{k0}-1), \\ w_{jkm}(t=T) & = \left(-2u_{j0m}(t=T) + \sqrt{2}y_{jm}^{\mathrm{obs}}\right)\delta_{k0}. \end{cases} \tag{9.15}$$

Using the expression given in (9.13), we derive the equation

$$\frac{\partial J(\mathbf{c})}{\partial c_i} \approx \sum_{m=1}^{N_{\mathrm{IC}}} \left(-\langle u_{j0m}(t), w_{j0m}(t)\rangle_t + \sum_{k=0}^{N_\mu-1} \langle u_{jkm}(t), w_{jkm}(t)\rangle_t \right) \mathring{B}_{ji,4} \tag{9.16}$$

for the spatial discretization of the gradient of the cost function. The spatially discretized scattering coefficient $\boldsymbol{\sigma} = (\sigma_j) \in \mathbb{R}^{N_x}$ can be computed as

$$\sigma_j \approx \sum_{i=-3}^{N_c-4} c_i \mathring{B}_{ji,4}. \tag{9.17}$$

9.3.3 Temporal discretization

To obtain a fully discrete system, the time interval $[0,T]$ is equidistantly split into a finite number N_t of time cells. An update of the forward equations (9.14) from time t_n to time $t_{n+1} = t_n + \Delta t$ is computed using an explicit Euler step forward in time according to

$$\begin{cases} u_{jkm}^{n+1} & = u_{jkm}^{n} - \Delta t \sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{x} u_{i\ell m}^{n} A_{k\ell} \\ & \quad + \Delta t \frac{\Delta x}{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{xx} u_{i\ell m}^{n} |A|_{k\ell} + \sigma_j \Delta t u_{jkm}^{n}(\delta_{k0}-1), \\ u_{jkm}^{0} & = u_{\mathrm{in},jkm}. \end{cases} \tag{9.18}$$

For the adjoint equations (9.15) we start computations with an end time condition after N_t steps and evolve the solution from time t_n to time $t_{n-1} = t_n - \Delta t$ by an explicit Euler step backwards in time according to

$$\begin{cases} w_{jkm}^{n-1} & = w_{jkm}^{n} + \Delta t \sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{x} w_{i\ell m}^{n} A_{k\ell} \\ & \quad + \Delta t \frac{\Delta x}{2}\sum_{i=1}^{N_x}\sum_{\ell=0}^{N_\mu-1} D_{ji}^{xx} w_{i\ell m}^{n} |A|_{k\ell} + \sigma_j \Delta t w_{jkm}^{n}(\delta_{k0}-1), \\ w_{jkm}^{N_t} & = \left(-2u_{j0m}^{N_t} + \sqrt{2}y_{jm}^{\mathrm{obs}}\right)\delta_{k0}. \end{cases} \tag{9.19}$$

The fully discrete gradient of the cost function can be obtained from (9.16) by approximating the integrals with respect to time by step functions, which yields

$$\frac{\partial J(\mathbf{c})}{\partial c_i} \approx \frac{1}{N_t+1} \sum_{m=1}^{N_{\mathrm{IC}}} \sum_{n=0}^{N_t} \left(-u_{j0m}^n w_{j0m}^n + \sum_{k=0}^{N_\mu - 1} u_{jkm}^n w_{jkm}^n \right) \mathring{B}_{ji,4}. \tag{9.20}$$

9.3.4 Fully discrete optimization scheme

The strategy for the fully discrete gradient descent method for the solution of the PDE parameter identification problem is summarized in Algorithm 4. Note that for the stopping criterion an error estimate `estimated-err` for the deviation of the computed coefficients from the true coefficients is required to run the algorithm. In practical applications, a stopping criterion depending on the amount of progress that is still made in the optimization procedure can be used.

Algorithm 4 Gradient descent method for the PDE parameter identification.

Input: measurements $\mathbf{y}_m^{\mathrm{obs}} = \left(y_{jm}^{\mathrm{obs}}\right) \in \mathbb{R}^{N_x}$ for $m = 1, ..., N_{\mathrm{IC}}$,
initial data $\mathbf{u}_m^0 = \left(u_{jkm}^0\right) \in \mathbb{R}^{N_x \times N_\mu}$ for $m = 1, ..., N_{\mathrm{IC}}$,
initial guess for the coefficients $\mathbf{c}^0 = \left(c_i^0\right) \in \mathbb{R}^{N_c}$,
initial step size $\eta^0 > 0$,
estimated error `estimated-err`,
error tolerance `errtol`,
maximal number of iterations `maxiter`

Output: optimal coefficients $\mathbf{c}^{\mathrm{opt}} = \left(c_i^{\mathrm{opt}}\right) \in \mathbb{R}^{N_c}$ within the prescribed error tolerance

while `estimated-err` $>$ `errtol` **and** $n \leq$ `maxiter` **do**
Compute $\boldsymbol{\sigma}^n = \left(\sigma_j^n\right) \in \mathbb{R}^{N_x}$ from the given coefficients $\mathbf{c}^n$ according to (9.17);
Solve the forward problem according to (9.18) for each $m = 1, ..., N_{\mathrm{IC}}$;
Solve the adjoint problem according to (9.19) for each $m = 1, ..., N_{\mathrm{IC}}$;
Compute the components of the gradient $\frac{\partial J(\mathbf{c}^n)}{\partial c_i^n}$ using (9.20) and the solutions of (9.18) and (9.19);
Update the coefficients according to (9.8): $\mathbf{c}^{n+1} = \mathbf{c}^n - \eta^n \nabla_{\mathbf{c}} J(\mathbf{c}^n)$, where η^n is adaptively determined by line search;
end while

9.4 Adaptive DLRA scheme for the fully discrete optimization problem

For the solution of the PDE parameter identification problem the coefficients $\mathbf{c}$ of the spline approximation (9.17) of $\boldsymbol{\sigma}$ are updated several times in the gradient descent step (9.8). For each iteration the solution of the fully discrete forward equations (9.18) as well as of the fully discrete adjoint equations (9.19) have to be computed and stored in order to compute the fully discretized gradient of the cost function as given in (9.20). This can

be computationally expensive. To reduce computational and memory requirements, the method of DLRA is applied to the fully discrete optimization procedure for the inverse transport problem proposed in Algorithm 4. We reformulate the forward equations (9.18) as well as the adjoint equations (9.19) using the rank-adaptive augmented BUG integrator introduced in [CKL22].

For the forward equations (9.18), the initial low-rank factors $\mathbf{X}_m^{0,\text{for}}, \mathbf{S}_m^{0,\text{for}}$ and $\mathbf{V}_m^{0,\text{for}}$ are obtained by an SVD of $\mathbf{u}_m^0 = \left(u_{jkm}^0\right) \in \mathbb{R}^{N_x \times N_\mu}$, where the number of singular values is truncated to the initial rank r. In each time step, the low-rank factors $\mathbf{X}_m^{n,\text{for}}, \mathbf{S}_m^{n,\text{for}}$ and $\mathbf{V}_m^{n,\text{for}}$ are evolved according to the following scheme.

First, we denote $\mathbf{K}_m^{n,\text{for}} = \mathbf{X}_m^{n,\text{for}} \mathbf{S}_m^{n,\text{for}}$ as well as $\mathbf{L}_m^{n,\text{for}} = \mathbf{V}_m^{n,\text{for}} \mathbf{S}_m^{n,\text{for},\top}$ and solve in parallel the equations

$$\begin{aligned}\mathbf{K}_m^{n+1,\text{for}} &= \mathbf{K}_m^{n,\text{for}} - \Delta t \mathbf{D}^x \mathbf{K}_m^{n,\text{for}} \mathbf{V}_m^{n,\text{for},\top} \mathbf{A}^\top \mathbf{V}_m^{n,\text{for}} + \Delta t \frac{\Delta x}{2} \mathbf{D}^{xx} \mathbf{K}_m^{n,\text{for}} \mathbf{V}_m^{n,\text{for},\top} |\mathbf{A}|^\top \mathbf{V}_m^{n,\text{for}} \\ &\quad + \Delta t \operatorname{diag}(\sigma) \mathbf{K}_m^{n,\text{for}} \mathbf{V}_m^{n,\text{for},\top} \mathbf{H} \mathbf{V}_m^{n,\text{for}}, \qquad (9.21a)\\ \mathbf{L}_m^{n+1,\text{for}} &= \mathbf{L}_m^{n,\text{for}} - \Delta t \mathbf{A} \mathbf{L}_m^{n,\text{for}} \mathbf{X}_m^{n,\text{for},\top} \mathbf{D}^{x,\top} \mathbf{X}_m^{n,\text{for}} + \Delta t \frac{\Delta x}{2} |\mathbf{A}| \mathbf{L}_m^{n,\text{for}} \mathbf{X}_m^{n,\text{for},\top} \mathbf{D}^{xx,\top} \mathbf{X}_m^{n,\text{for}} \\ &\quad + \Delta t \mathbf{H} \mathbf{L}_m^{n,\text{for}} \mathbf{X}_m^{n,\text{for},\top} \operatorname{diag}(\sigma) \mathbf{X}_m^{n,\text{for}}, \qquad (9.21b)\end{aligned}$$

where $\mathbf{H} = \operatorname{diag}([0, -1, ..., -1])$. In the next step, we derive the augmented and time-updated bases $\widehat{\mathbf{X}}_m^{n+1,\text{for}}$ and $\widehat{\mathbf{V}}_m^{n+1,\text{for}}$ from a QR-decomposition of the augmented quantities $\widehat{\mathbf{X}}_m^{n+1,\text{for}} = \operatorname{qr}\left(\left[\mathbf{K}_m^{n+1,\text{for}}, \mathbf{X}_m^{n,\text{for}}\right]\right)$ and $\widehat{\mathbf{V}}_m^{n+1,\text{for}} = \operatorname{qr}\left(\left[\mathbf{L}_m^{n+1,\text{for}}, \mathbf{V}_m^{n,\text{for}}\right]\right)$, respectively. For the S-step, we introduce the notation $\widetilde{\mathbf{S}}_m^{n,\text{for}} = \widehat{\mathbf{X}}_m^{n+1,\text{for},\top} \mathbf{X}_m^{n,\text{for}} \mathbf{S}_m^{n,\text{for}} \mathbf{V}_m^{n,\text{for},\top} \widehat{\mathbf{V}}_m^{n+1,\text{for}}$ and compute

$$\begin{aligned}\widehat{\mathbf{S}}_m^{n+1,\text{for}} &= \widetilde{\mathbf{S}}_m^{n,\text{for}} - \Delta t \widehat{\mathbf{X}}_m^{n+1,\text{for},\top} \mathbf{D}^x \widehat{\mathbf{X}}_m^{n+1,\text{for}} \widetilde{\mathbf{S}}_m^{n,\text{for}} \widehat{\mathbf{V}}_m^{n+1,\text{for},\top} \mathbf{A}^\top \widehat{\mathbf{V}}_m^{n+1,\text{for}} \\ &\quad + \Delta t \frac{\Delta x}{2} \widehat{\mathbf{X}}_m^{n+1,\text{for},\top} \mathbf{D}^{xx} \widehat{\mathbf{X}}_m^{n+1,\text{for}} \widetilde{\mathbf{S}}_m^{n,\text{for}} \widehat{\mathbf{V}}_m^{n+1,\text{for},\top} |\mathbf{A}|^\top \widehat{\mathbf{V}}_m^{n+1,\text{for}} \qquad (9.21c)\\ &\quad + \Delta t \widehat{\mathbf{X}}_m^{n+1,\text{for},\top} \operatorname{diag}(\sigma) \widehat{\mathbf{X}}_m^{n+1,\text{for}} \widetilde{\mathbf{S}}_m^{n,\text{for}} \widehat{\mathbf{V}}_m^{n+1,\text{for},\top} \mathbf{H} \widehat{\mathbf{V}}_m^{n+1,\text{for}}.\end{aligned}$$

Finally, we truncate the time-updated augmented low-rank factors for each $m = 1, ..., N_{\text{IC}}$ to a new rank $r_{n+1} \leq 2r$ by using a suitable truncation strategy such as proposed in Section 4.2.2. Then the time-updated numerical solutions of the forward problem are given by $\mathbf{u}_m^{n+1} = \mathbf{X}_m^{n+1,\text{for}} \mathbf{S}_m^{n+1,\text{for}} \mathbf{V}_m^{n+1,\text{for},\top} \in \mathbb{R}^{N_x \times N_\mu}$.

For the adjoint equations (9.19), we perform an SVD of the end time solutions $\mathbf{w}_m^{N_t} = \left(w_{jkm}^{N_t}\right) \in \mathbb{R}^{N_x \times N_\mu}$, truncate to the prescribed initial rank r, and obtain the low-rank factors $\mathbf{X}_m^{N_t,\text{adj}}, \mathbf{S}_m^{N_t,\text{adj}}$ and $\mathbf{V}_m^{N_t,\text{adj}}$. Then, in each step, the low-rank factors $\mathbf{X}_m^{n,\text{adj}}, \mathbf{S}_m^{n,\text{adj}}$ and $\mathbf{V}_m^{n,\text{adj}}$ are evolved backwards in time as follows.

First, we denote $\mathbf{K}_m^{n,\text{adj}} = \mathbf{X}_m^{n,\text{adj}} \mathbf{S}_m^{n,\text{adj}}$ as well as $\mathbf{L}_m^{n,\text{adj}} = \mathbf{V}_m^{n,\text{adj}} \mathbf{S}_m^{n,\text{adj},\top}$ and solve in parallel the equations

$$\begin{aligned}\mathbf{K}_m^{n-1,\text{adj}} &= \mathbf{K}_m^{n,\text{adj}} + \Delta t \mathbf{D}^x \mathbf{K}_m^{n,\text{adj}} \mathbf{V}_m^{n,\text{adj},\top} \mathbf{A}^\top \mathbf{V}_m^{n,\text{adj}} + \Delta t \frac{\Delta x}{2} \mathbf{D}^{xx} \mathbf{K}_m^{n,\text{adj}} \mathbf{V}_m^{n,\text{adj},\top} |\mathbf{A}|^\top \mathbf{V}_m^{n,\text{adj}} \\ &\quad + \Delta t \operatorname{diag}(\sigma) \mathbf{K}_m^{n,\text{adj}} \mathbf{V}_m^{n,\text{adj},\top} \mathbf{H} \mathbf{V}_m^{n,\text{adj}},\end{aligned}$$

$$\mathbf{L}_m^{n-1,\mathrm{adj}} = \mathbf{L}_m^{n,\mathrm{adj}} + \Delta t \mathbf{A}\mathbf{L}_m^{n,\mathrm{adj}}\mathbf{X}_m^{n,\mathrm{adj},\top}\mathbf{D}^{x,\top}\mathbf{X}_m^{n,\mathrm{adj}} + \Delta t\frac{\Delta x}{2}|\mathbf{A}|\mathbf{L}_m^{n,\mathrm{adj}}\mathbf{X}_m^{n,\mathrm{adj},\top}\mathbf{D}^{xx,\top}\mathbf{X}_m^{n,\mathrm{adj}} + \Delta t\mathbf{H}\mathbf{L}_m^{n,\mathrm{adj}}\mathbf{X}_m^{n,\mathrm{adj},\top}\operatorname{diag}(\sigma)\mathbf{X}_m^{n,\mathrm{adj}}.$$

In the next step, we derive the augmented and time-updated bases $\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj}}$ and $\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj}}$ from a QR-decomposition of the augmented quantities $\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj}} = \mathrm{qr}\left(\left[\mathbf{K}_m^{n-1,\mathrm{adj}}, \mathbf{X}_m^{n,\mathrm{adj}}\right]\right)$ and $\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj}} = \mathrm{qr}\left(\left[\mathbf{L}_m^{n-1,\mathrm{adj}}, \mathbf{V}_m^{n,\mathrm{adj}}\right]\right)$, respectively.
For the S-step, we set $\widetilde{\mathbf{S}}_m^{n,\mathrm{adj}} = \widehat{\mathbf{X}}_m^{n-1,\mathrm{adj},\top}\mathbf{X}_m^{n,\mathrm{adj}}\mathbf{S}_m^{n,\mathrm{adj}}\mathbf{V}_m^{n,\mathrm{adj},\top}\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj}}$ and compute

$$\begin{aligned}\widehat{\mathbf{S}}_m^{n-1,\mathrm{adj}} = \widetilde{\mathbf{S}}_m^{n,\mathrm{adj}} &+ \Delta t\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj},\top}\mathbf{D}^{x}\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj}}\widetilde{\mathbf{S}}_m^{n,\mathrm{adj}}\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj},\top}\mathbf{A}^{\top}\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj}} \\ &+ \Delta t\frac{\Delta x}{2}\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj},\top}\mathbf{D}^{xx}\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj}}\widetilde{\mathbf{S}}_m^{n,\mathrm{adj}}\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj},\top}|\mathbf{A}|^{\top}\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj}} \\ &+ \Delta t\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj},\top}\operatorname{diag}(\sigma)\widehat{\mathbf{X}}_m^{n-1,\mathrm{adj}}\widetilde{\mathbf{S}}_m^{n,\mathrm{adj}}\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj},\top}\mathbf{H}\widehat{\mathbf{V}}_m^{n-1,\mathrm{adj}}.\end{aligned}$$

Finally, we truncate the time-updated augmented low-rank factors for each $m = 1, ..., N_{\mathrm{IC}}$ to a new rank $r_{n+1} \leq 2r$ by using a suitable truncation strategy such as proposed in Section 4.2.2. Then the time-updated numerical solutions of the adjoint problem are given by $\mathbf{w}_m^{n-1} = \mathbf{X}_m^{n-1,\mathrm{adj}}\mathbf{S}_m^{n-1,\mathrm{adj}}\mathbf{V}_m^{n-1,\top,\mathrm{adj}} \in \mathbb{R}^{N_x \times N_\mu}$.
Having determined the low-rank solutions of the forward and the adjoint problems, we can use them to compute the gradient as proposed in (9.20). For the update of the coefficients according to (9.8), we adaptively determine the step size by a backtracking line search approach with Armijo condition similar to [SEKM25] and as described in Algorithm 5.
The line search method works as follows: For a given step size η^n the B-spline coefficients and the scattering coefficient are updated to $\mathbf{c}^{n+1}$ and $\boldsymbol{\sigma}^{n+1}$, respectively. Then the truncation error tolerance ϑ is adjusted using the given step size η^n and the maximal absolute value of $\nabla_{\mathbf{c}} J(\mathbf{c}^n)$. We add some safety parameters h_2 and h_3 as well as a lower bound h_1 for the truncation tolerance. In the next step we compute the value of the cost function J with the low-rank factors of the forward problem at hand and denote it with J^n. Then we solve the forward problem (9.21) with $\boldsymbol{\sigma}^{n+1}$ and the updated value of ϑ and use the obtained low-rank factors for another evaluation of the cost function J, for which the result is denoted by $\bar{J}^n$. While the difference between J^n and $\bar{J}^n$ is larger than a prescribed tolerance depending on the Euclidean norm $\|\cdot\|_E$ of the gradient of the cost function, the gradient descent step size is reduced by the step size reduction factor p and the procedure is repeated.

9.5 Numerical results

We consider the following test examples in one space and one angular dimension to show the computational accuracy and efficiency of the proposed DLRA scheme compared to computations with full solvers for both the forward and the adjoint equations. In Section 9.5.1 initial distributions of Cosine type are treated. Section 9.5.2 presents numerical results for Gaussian initial distributions.

Algorithm 5 Backtracking line search method for the adaptive refinement of the gradient descent step size and the DLRA rank tolerance.

Input: cost function J,
coefficients $\mathbf{c}^n$,
gradient $\nabla_{\mathbf{c}} J(\mathbf{c}^n)$ computed using (9.20),
low-rank factors $\mathbf{X}_m^{n,\text{for}}, \mathbf{S}_m^{n,\text{for}}, \mathbf{V}_m^{n,\text{for}}$ of the forward problem (9.21) for $m = 1, ..., N_{\text{IC}}$,
step size $\eta^n > 0$,
rank error tolerance ϑ,
step size reduction factor $p \in (0, 1)$,
constants h_1, h_2, h_3, h_4

Output: refined step size η^{n+1}, refined rank error tolerance ϑ, updated coefficients $\mathbf{c}^{n+1}$

Update the coefficients according to (9.8): $\mathbf{c}^{n+1} = \mathbf{c}^n - \eta^n \nabla_{\mathbf{c}} J(\mathbf{c}^n)$;
Compute $\boldsymbol{\sigma}^{n+1}$ from the updated coefficients $\mathbf{c}^{n+1}$ according to (9.17);
Update $\vartheta = \max\left(h_1, \min\left(h_2, \eta^n h_3 \left\|\nabla_{\mathbf{c}} J(\mathbf{c}^n)\right\|_\infty\right)\right)$;

Compute $J^n = J\left(\mathbf{X}_1^{n,\text{for}} \mathbf{S}_1^{n,\text{for}} \mathbf{V}_1^{n,\text{for}}, ..., \mathbf{X}_{N_{\text{IC}}}^{n,\text{for}} \mathbf{S}_{N_{\text{IC}}}^{n,\text{for}} \mathbf{V}_{N_{\text{IC}}}^{n,\text{for}}\right)$;
Compute $\overline{\mathbf{X}}_m^{n,\text{for}} \overline{\mathbf{S}}_m^{n,\text{for}} \overline{\mathbf{V}}_m^{n,\text{for}}$ from (9.21) for $m = 1, ..., N_{\text{IC}}$ with $\boldsymbol{\sigma}^{n+1}$ and the updated ϑ;
Compute $\bar{J}^n = J\left(\overline{\mathbf{X}}_1^{n,\text{for}} \overline{\mathbf{S}}_1^{n,\text{for}} \overline{\mathbf{V}}_1^{n,\text{for}}, ..., \overline{\mathbf{X}}_{N_{\text{IC}}}^{n,\text{for}} \overline{\mathbf{S}}_{N_{\text{IC}}}^{n,\text{for}} \overline{\mathbf{V}}_{N_{\text{IC}}}^{n,\text{for}}\right)$;

while $\bar{J}^n > J^n - \eta^n h_4 \left\|\nabla_{\mathbf{c}} J(\mathbf{c}^n)\right\|_E^2$ **do**
  Update $\eta^{n+1} = p\eta^n$;

  Update the coefficients: $\mathbf{c}^{n+1} = \mathbf{c}^{n+1} - \eta^{n+1} \nabla_{\mathbf{c}} J(\mathbf{c}^n)$;
  Compute $\boldsymbol{\sigma}^{n+1}$ from the updated coefficients $\mathbf{c}^{n+1}$ according to (9.17);
  Update $\vartheta = \max\left(h_1, \min\left(h_2, \eta^{n+1} h_3 \left\|\nabla_{\mathbf{c}} J(\mathbf{c}^n)\right\|_\infty\right)\right)$;

  Compute $\overline{\mathbf{X}}_m^{n,\text{for}} \overline{\mathbf{S}}_m^{n,\text{for}} \overline{\mathbf{V}}_m^{n,\text{for}}$ from (9.21) for $m = 1, ..., N_{\text{IC}}$ with $\boldsymbol{\sigma}^{n+1}$ and the updated ϑ;
  Compute $\bar{J}^n = J\left(\overline{\mathbf{X}}_1^{n,\text{for}} \overline{\mathbf{S}}_1^{n,\text{for}} \overline{\mathbf{V}}_1^{n,\text{for}}, ..., \overline{\mathbf{X}}_{N_{\text{IC}}}^{n,\text{for}} \overline{\mathbf{S}}_{N_{\text{IC}}}^{n,\text{for}} \overline{\mathbf{V}}_{N_{\text{IC}}}^{n,\text{for}}\right)$;

  Set $\eta^n = \eta^{n+1}$;
end while

9.5.1 1D cosine

For the first numerical experiment the spatial as well as the angular domain are set to $\Omega_x = \Omega_\mu = [-1, 1]$. We consider $N_{\text{IC}} = 3$ initial distributions of Cosine type of the form

$$u_m(t = 0, x) = 2 + \cos\left(\left(x - \frac{2m}{3}\right)\pi\right) \qquad \text{for} \quad m = 1, 2, 3.$$

The true and the initial B-spline coefficients for the approximation of the scattering coefficient σ are chosen as

$$\mathbf{c}_{\text{true}} = (2.1, 2.0, 2.2)^\top \qquad \text{and} \qquad \mathbf{c}_{\text{init}} = (1.0, 1.5, 3.0)^\top,$$

and we consider normalized cubic periodic B-spline basis functions $\mathring{B}_{i,4}(x)$ with equally spaced knots. For the low-rank computations, we start with an initial rank of $r = 5$ in the forward as well as in the adjoint problem. The maximal allowed value of the rank in each step is restricted to 20. As computational parameters we use $N_x = 100$ cells in the

spatial domain and $N_\mu = 250$ moments in the angular variable. The end time is set to $T = 1.0$ and the time step size of the algorithm is chosen according to $\Delta t = C_{\mathrm{CFL}} \cdot \Delta x$ with a CFL number of $C_{\mathrm{CFL}} = 0.99$. We begin the gradient descent method with a step size of $\eta^0 = 5 \cdot 10^5$ and a truncation error tolerance of $\vartheta = 10^{-2} \|\mathbf{\Sigma}\|_F$, where $\mathbf{\Sigma} \in \mathbb{R}^{2r \times 2r}$ represents the diagonal matrix that is obtained from the SVD in the truncation strategy described in Section 4.2.2 and $\|\cdot\|_F$ denotes the Frobenius norm. For the rescaling of the gradient descent step size and the DLRA rank tolerance we use the step size reduction factor $p = 0.5$ as well as the constants $h_1 = 10^{-3} \|\mathbf{\Sigma}\|_F$ for a lower bound of the rank tolerance and $h_2 = 0.1 \|\mathbf{\Sigma}\|_F$, $h_3 = 0.1 \|\mathbf{\Sigma}\|_F$ as safety parameters. Also $h_4 = 0.5$ is added as a safety parameter to ensure a reasonable difference between J^n and $\bar{J}^n$ in Algorithm 5. The whole gradient descent procedure is conducted until the prescribed error tolerance $\texttt{errtol} = 10^{-4}$ or a maximal number of iterations $\texttt{maxiter} = 500$ is reached.

In Figure 9.1 we compare the solutions of the parameter identification problem computed with the full solvers and the DLRA solvers for both the forward and the adjoint equations. Corresponding to the number of initial conditions, we plot three curves for the scalar flux $\Phi = \frac{1}{\sqrt{2}} \langle f \rangle_\mu$ and display the results that are obtained from the true coefficients and at the end of the gradient descent optimization procedure, evaluated with both the full and the DLRA solver. We observe that the DLRA solution captures well the behavior of the full solution and that they both approach the solutions computed with the true coefficients. In addition, the parameter reconstruction inverse problem for determining σ is accurately resolved with both solvers. It can be observed in the bottom row that beginning with $\boldsymbol{\sigma}_{\mathrm{init}}$ both the full and the DLRA method converge to the true solution $\boldsymbol{\sigma}_{\mathrm{true}}$. The DLRA reconstruction closely resembles the reconstruction computed with the full solvers. Further, the evolution of the rank r in time for the DLRA method is illustrated, where we have averaged the ranks of the forward equations computed with the different initial conditions to obtain r^{for} and the ranks of the adjoint equations computed with the different initial conditions to obtain r^{adj} and finally set $r = \frac{1}{2}\left(r^{\mathrm{for}} + r^{\mathrm{adj}}\right)$. We notice that in the beginning of the optimization process the averaged rank decreases as the initial rank was chosen larger than required. From then on, we observe a relatively monotonous increase until it stays at approximately $r = 9$. This evolution of the rank reflects the fact that in the beginning of the optimization the error tolerance ϑ is chosen quite large as the computed solution is still comparably far away from the true solution. As the optimization algorithm approaches the true coefficients, the DLRA rank tolerance ϑ is decreased, resulting in a higher averaged rank.

For the considered setup, the computational benefit of the DLRA method compared to the solution of the full problem is significant. The scheme is implemented in Julia v1.11 and performed on a MacBook Pro with M1 chip, resulting in a decrease of run time by a factor of approximately 2.5 from 139 seconds to 56 seconds while retaining the accuracy of the computed results. Concerning the memory costs, the solutions of the forward problem and of the adjoint problem have to be stored in order to compute the gradient. For each initial condition, the storage of the solution of the forward problem corresponds to a memory cost of $8\,(N_t + 1)\,N_x N_\mu$, which for the DLRA method can be lowered to $8\,(N_t + 1)\left(rN_x + rN_\mu + r^2\right)$, where r is the maximal averaged rank in the simulation.

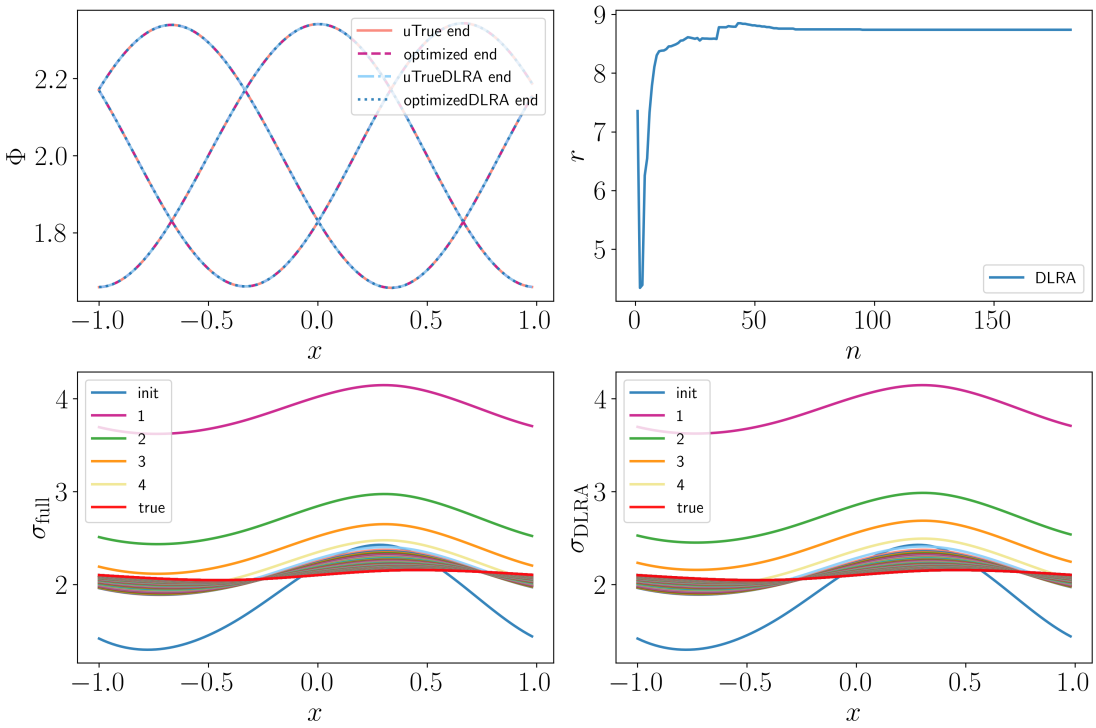


Figure 9.1: Top left: Numerical results for the scalar flux Φ of the 1D Cosine problem, computed with the full solvers and the DLRA solvers, both with the true coefficients and with the optimization gradient descent scheme. **Top right:** Evolution of the averaged rank r for the DLRA method. **Bottom row:** Iterations for the reconstruction of the scattering coefficient σ computed with both the full solvers (left) and the DLRA solvers (right).

9.5.2 1D Gaussian distribution

In a second test example, we prescribe $\Omega_x = [0, 10]$ for the spatial and $\Omega_\mu = [-1, 1]$ for the angular domain. We consider $N_{\text{IC}} = 5$ Gaussian initial distributions of the form

$$u_m\left(t = 0, x\right) = \max\left(10^{-8}, \frac{1}{\sqrt{2\pi\sigma_{\text{IC}}^2}} \exp\left(-\frac{\left(x - x_0\right)^2}{2\sigma_{\text{IC}}^2}\right)\right) \quad \text{for} \quad m = 1, 2, 3, 4, 5,$$

which are centered around equidistantly distributed x_0 and periodically extended on the domain Ω_x. The standard deviation is set to the constant value $\sigma_{\text{IC}} = 0.8$. The true and the initial B-spline coefficients for the approximation of the scattering coefficient σ are chosen as

$$\mathbf{c}_{\text{true}} = \left(2.1, 2.0, 2.2, 2.0, 1.9\right)^\top \quad \text{and} \quad \mathbf{c}_{\text{init}} = \left(2.8, 1.5, 3.0, 2.1, 1.2\right)^\top,$$

and we consider normalized cubic periodic B-spline basis functions $\mathring{B}_{i,4}\left(x\right)$ with equally spaced knots. All other settings and computational parameters remain unchanged from the previous test example given in Section 9.5.1.

In Figure 9.2 we compare the solutions of the parameter identification problem computed with the full solvers and the DLRA solvers for both the forward and the adjoint equations.

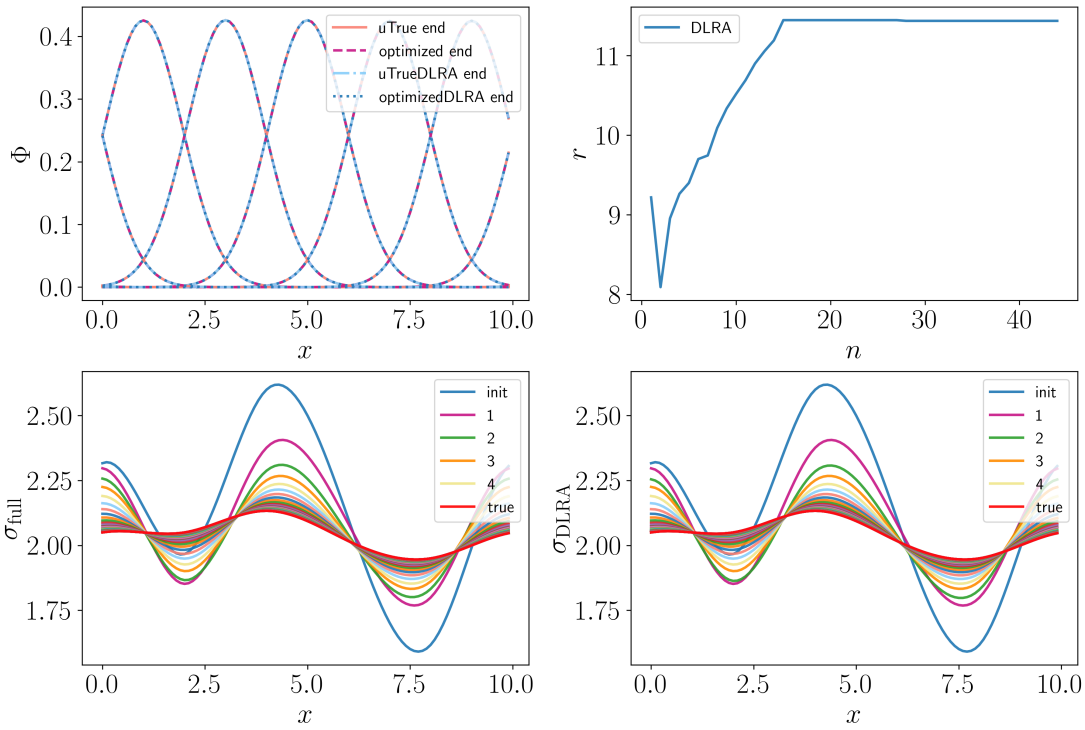


Figure 9.2: Top left: Numerical results for the scalar flux Φ of the 1D Gauss problem, computed with the full solvers and the DLRA solvers, both with the true coefficients and with the optimization gradient descent scheme. **Top right:** Evolution of the averaged rank r for the DLRA method. **Bottom row:** Iterations for the reconstruction of the scattering coefficient σ computed with both the full solvers (left) and the DLRA solvers (right).

Corresponding to the number of initial conditions, we plot five curves for the scalar flux $\Phi = \frac{1}{\sqrt{2}} \langle f \rangle_\mu$ and display the results that are obtained from the true coefficients and at the end of the gradient descent optimization procedure, evaluated with both the full and the DLRA solver. Again we observe that the DLRA solution captures well the behavior of the full solution and that they both approach the solutions computed with the true coefficients. For the reconstruction of the scattering coefficient σ, it can be observed in the bottom row that beginning with $\boldsymbol{\sigma}_{\text{init}}$ both the full and the DLRA method converge to the true solution $\boldsymbol{\sigma}_{\text{true}}$ and that the DLRA reconstruction closely resembles the reconstruction computed with the full solvers. The averaged rank r first decreases as the initial rank was chosen larger than required. From then on, we observe the expected relatively monotonous increase until it stagnates at a value of approximately $r = 11.5$.

The scheme is implemented in Julia v1.11 and performed on a MacBook Pro with M1 chip, resulting in a decrease of run time by a factor of approximately 2 from 11.5 seconds to 6 seconds. Again, for each initial condition, the memory costs reduce from $8\,(N_t + 1)\,N_x N_\mu$ for the full solvers to $8\,(N_t + 1)\,\left(rN_x + rN_\mu + r^2\right)$ for the DLRA solvers, underlining the computational efficiency of the DLRA scheme.

9.6 Summary and conclusion

We have presented a fully discrete DLRA scheme for the reconstruction of the scattering coefficient in the 1D RTE making use of a PDE constrained optimization procedure. The main research contributions are:

(i) *An application of DLRA to a PDE parameter identification inverse problem:* The scattering coefficient $\sigma(x)$ has been determined by PDE constrained optimization, for which after the discretization a DLRA approach has been used. To our knowledge, this is the first research contribution that combines inverse problems and DLRA, allowing for a reduction of computational effort and memory requirements in a computationally demanding setup where in each step of the optimization procedure both the forward and the adjoint equations have to be solved.

(ii) *A setup close to realistic applications:* In most applications, measurements are not able to access the full distribution function but at most angle-averaged quantities, i.e. its moments. We have considered such a setup here, where we have assumed that only the first moment is accessible by measurements. In addition, in optical tomography usually measurements from different positions are taken into account, which we have incorporated by probing as many initial values as coefficients to be reconstructed, enriching the underlying data set.

(iii) *An adaptive gradient descent step size and a rank-adaptive augmented integrator:* For the minimization we have used a gradient descent method which updates the coefficients of a normalized cubic periodic B-spline approximation of $\sigma(x)$. Similar to [SEKM25], the step size has been adaptively chosen by a backtracking line search approach with Armijo condition. Also the rank of the DLRA algorithm has been adaptively determined by using the rank-adaptive augmented BUG integrator presented in [CKL22] combined with an adaptively chosen truncation error tolerance. As a result, this has enabled us to begin the optimization procedure with a comparatively small rank (when the solution is still far from the minimum) and gradually increase the rank as the optimization progresses and more accuracy is required, again enhancing the performance of the DLRA scheme.

(iv) *Numerical test examples showing good agreement:* We have given a number of 1D numerical test examples confirming that for the reconstruction of the scattering coefficient in the inverse transport problem the application of DLRA shows good agreement with the full solution while being significantly faster and saving memory demands.

Altogether, the application of DLRA methods to parameter identification inverse problems provides promising numerical results, motivating for future investigations in this area of research.

10

Conclusion and outlook

The construction of an appropriate numerical scheme for the solution of kinetic PDEs is challenging. Due to the, in general, high dimensionality of kinetic equations, numerical reduction techniques such as DLRA are advantageous to reduce the computational effort and memory requirements. While DLRA has been shown to provide efficient and accurate approximations to the solution of various kinetic equations, it can be considered a rather destructive method regarding conservation properties as the preservation of important physical invariants can often not be guaranteed when reducing the general complexity of the solution. In addition, deriving stability estimates is demanding as the low-rank structure (4.4) imposes non-linear dependencies in the evolution equations for the low-rank factors $\mathbf{X}$, $\mathbf{S}$, and $\mathbf{V}$, even for linear equations. Using suitable time integrators, these nonlinearities can be decoupled by subsequently solving update equations in which all but one low-rank factor is fixed. Then, for linear equations, concepts of von Neumann stability analysis can be used to deduce estimates, for instance on the energy of the system, contributing to the stability considerations.

Part I. Stability analysis for DLRA schemes. In the first part of this thesis, the topic of (energy) stability and conservation properties for DLRA schemes was addressed, beginning with the thermal RTEs with Su-Olson closure. This closure led to a linearized internal energy model, called the Su-Olson problem. In Chapter 5, a "first low-rank, then discretize" approach was pursued and, based on an implicit coupling of the equations, a provably energy stable DLRA scheme was derived. Together with suitable basis augmentations and an adjusted truncation strategy, mass conservation was ensured. In Chapter 6, a multiplicative splitting of the distribution function was imposed. This multiplicative structure gave rise to further complexities such as the question of an adequate discretization of the spatial derivatives. Additional basis augmentations were required for a rigorous proof of energy stability of the DLRA scheme. In addition, the structural order had to be changed into "first discretize, then low-rank" to obtain thorough theoretical results. Again, it was possible to show mass conservation under the same proper treatment as done before. In Chapter 7, the linear Boltzmann-BGK equation was considered. For the translation of knowledge on the construction of efficient DLRA algorithms to more compli-

cated (potentially non-linear) settings, again a multiplicative structure of the distribution function was assumed. The first difficulty consisted in determining a suitable stability norm, which in this case was not directly related to the physical energy of the system. For the proof of numerical stability of the proposed DLRA scheme, additional basis augmentations (different from the ones in the previous chapter) as well as a "first discretize, then low-rank" approach were necessary. Further, a specifically designed truncation strategy had to be implemented. Altogether, a rigorous stability analysis was conducted for three efficient and accurate DLRA schemes for linear equations, giving insights into the structural difficulties of DLRA algorithms and proposing solution strategies.

Part II. Application of DLRA to inverse problems. The second part of this thesis was devoted to the application of the DLRA method to parameter identification inverse problems. In Chapter 9, the combination of DLRA and inverse problems was experimentally examined in numerical test problems for the reconstruction of the scattering coefficient in the RTE. For the numerical optimization, a gradient descent approach on a comparably small parameter space obtained from a spline approximation of the scattering parameter was pursued. In the sense of a "first optimize, then discretize, then low-rank" ansatz, a low-rank solver was implemented for the solution of the fully discrete forward and adjoint equations in each step of an iterative gradient descent scheme. In this chapter, no theoretical results were provided but numerical examples demonstrated good solution properties of the proposed DLRA scheme.

Outlook. For future research, various open questions are left. Concerning stability estimates for DLRA schemes, a general difficult task consists in finding a suitable notion of stability depending on the underlying problem. In addition, the stability estimates get much more complicated when refraining from periodic boundary conditions, which were assumed throughout this thesis. Even though this thesis has proposed strategies for the construction of provably stable DLRA algorithms for linear equations, a direct transition of knowledge to the corresponding non-linear equations is hardly possible. The main reason for that is that in the non-linear case most of the theoretical concepts applied such as the von Neumann stability approach are not available, making the analysis much more difficult. Hence, for non-linear equations a different strategy for the proof of stability must be used. In addition, we have seen that for the multiplicative splitting a discretization of the conservative form of the equations was necessary to obtain a numerically stable algorithm. For the non-linear Boltzmann-BGK equation such a discretization, i.e. by not splitting up the term $\partial_x (Mg)$, is possible but cannot be efficiently implemented as the Maxwellian M is generally not of low rank. Thus, the question of provable stability results for an efficient DLRA scheme for the non-linear Boltzmann-BGK equation with multiplicative splitting remains subject to future research. Moreover, further investigations on the structural order of discretizing and applying the DLRA method are of interest as we have seen that this ordering plays an important role when deriving stability estimates for a multiplicative splitting of the distribution function and had to be changed in Chapter 6 and 7 compared to Chapter 5. Regarding conservation properties, we have used that

the rank-adaptive augmented BUG integrator presented in [CKL22] allows for additional basis augmentations, ensuring that certain physical quantities are conserved in the bases over time. This approach is well-studied [EKS23] but not generally applicable as for example a specific choice of the temporal discretization is required.
Concerning the second part of this thesis, several future research projects on the application of DLRA to inverse problems are possible. A first natural extension of the results presented in Chapter 9 consists in considering numerical test examples in more than one spatial and angular variable since in higher dimensions the savings by the DLRA method are expected to be larger by orders of magnitude. Also, theoretical considerations concerning for example the stability of DLRA schemes applied to parameter identification inverse problems can provide valuable insights into the structure of such problems. In addition, various open questions arise when the structural order of the problem is changed, meaning that for example a "first low-rank, then optimize, then discretize" strategy is pursued. For instance, it is not clear how the adjoint equations can be derived from the low-rank components of the forward problem as the low-rank equations are highly nonlinear.
Altogether, this thesis underlines that each DLRA scheme has to be carefully constructed such that for instance stability estimates and conservation properties are ensured and that a direct transition of knowledge from one problem to another is only partially possible. However, various numerical experiments show that the DLRA method exhibits significant potential in reducing the computational costs, memory demands, and general complexities for the solution of kinetic equations, which especially for iterative optimization schemes can be extremely useful.

Glossary of abbreviations

1D one-dimensional. 15, 19, 27, 29, 30, 49, 50, 70, 75, 93, 94, 100, 117–119, 123, 126, 134, 141, 142, 155

2D two-dimensional. 19, 28, 30, 72–75, 94, 100, 117, 122–124, 126

3D three-dimensional. 5, 11, 30, 50, 72

BGK Bhatnagar-Gross-Krook. i, iii, 2–5, 10, 13, 44, 77, 78, 82, 96, 99, 100, 103, 105, 106, 125, 126, 157, 158

BUG basis update & Galerkin. 2, 3, 38–41, 43–45, 49, 51–53, 58, 68, 70, 86, 93, 96, 112, 126, 149, 155, 159

CFL Courant-Friedrichs-Lewy. 2, 3, 23, 24, 44, 60, 71, 74, 75, 77–79, 83, 94, 96, 101, 106, 111, 118, 122, 126, 152

DLRA dynamical low-rank approximation. i, iii, 1–4, 33, 36, 38, 39, 41, 43–45, 49, 51, 57, 62, 64, 68–72, 74, 75, 77–79, 82, 86–88, 90, 92–97, 99–101, 111, 112, 114, 116–120, 122–126, 141, 143, 149, 150, 152–155, 157–159

FD finite difference. 16, 17, 19–26, 54, 80, 82, 102, 106, 146, 147

IMEX implicit-explicit. 21, 58, 75

IVP initial value problem. 7, 22

ODE ordinary differential equation. 37, 39, 52, 54

PDE partial differential equation. i, iii, 1–3, 5–7, 15, 19, 20, 22, 23, 25, 33, 36, 37, 39, 43, 51, 52, 100, 129, 131–133, 141–144, 148, 155, 157

RTE radiative transfer equation. i, iii, 2–4, 41, 44, 45, 49–51, 70, 77, 141, 142, 155, 157, 158

SVD singular value decomposition. 2, 34, 39, 44, 46, 113, 130, 149, 152

Bibliography

[AAC16] F. Achleitner, A. Arnold, and E. A. Carlen. On linear hypocoercive BGK models. In P. Gonçalves and A. J. Soares, editors, *From Particle Systems to Partial Differential Equations III. Springer Proceedings in Mathematics & Statistics*, volume 162, pages 1–37, Cham, 2016. Springer. doi:10.1007/978-3-319-32144-8_1.

[Ale21] A. Alexanderian. Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: a review. *Inverse Problems*, 37(4):043001, 2021. doi:10.1088/1361-6420/abe10c.

[Alf42] H. Alfvén. Existence of electromagnetic-hydrodynamic waves. *Nature*, 150(3805):405–406, 1942. doi:10.1038/150405d0.

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, 2008. doi:10.1515/9781400830244.

[Arm66] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966. doi:10.2140/pjm.1966.16.1.

[AS72] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. United States Department of Commerce. National Bureau of Standards, Washington, D.C., tenth edition, 1972.

[AS01] I. K. Abu-Shumays. Angular quadratures for improved transport computations. *Transport Theory and Statistical Physics*, 30(2–3):169–204, 2001. doi:10.1081/TT-100105367.

[Atk89] K. Atkinson. *An Introduction to Numerical Analysis*. Wiley, New York, second edition, 1989.

[ATPP00] P. Andries, P. Le Tallec, J.-P. Perlat, and B. Perthame. The Gaussian-BGK model of Boltzmann equation with small Prandtl number. *European Journal of Mechanics – B/Fluids*, 19(6):813–830, 2000. doi:10.1016/S0997-7546(00)01103-1.

[Bal09] G. Bal. Inverse transport theory and applications. *Inverse problems*, 25(5):053001, 2009. doi:10.1088/0266-5611/25/5/053001.

[Bal19] G. Bal. Introduction to inverse problems. Lecture notes, University of Chicago, Chicago, 2019.

[BBM21] M. Bertero, P. Boccacci, and C. De Mol. *Introduction to Inverse Problems in Imaging*. CRC Press, Boca Raton, second edition, 2021. doi:10.1201/9781003032755.

[BCHR20] M. Bessemoulin-Chatard, M. Herda, and T. Rey. Hypocoercivity and diffusion limit of a finite volume scheme for linear kinetic equations. *Mathematics of Computation*, 89(323):1093–1133, 2020. doi:10.1090/mcom/3490.

[BEKK24a] L. Baumann, L. Einkemmer, C. Klingenberg, and J. Kusch. Energy stable and conservative dynamical low-rank approximation for the Su-Olson problem. *SIAM Journal on Scientific Computing*, 46(2):B137–B158, 2024. doi:10.1137/23M1586215.

[BEKK24b] L. Baumann, L. Einkemmer, C. Klingenberg, and J. Kusch. A stable multiplicative dynamical low-rank discretization for the linear Boltzmann-BGK equation, 2024. arXiv:2411.06844.

[BEKK25a] L. Baumann, L. Einkemmer, C. Klingenberg, and J. Kusch. An adaptive dynamical low-rank optimizer for solving kinetic parameter identification inverse problems, 2025. arXiv:2506.21405.

[BEKK25b] L. Baumann, L. Einkemmer, C. Klingenberg, and J. Kusch. An energy stable and conservative multiplicative dynamical low-rank discretization for the Su-Olson problem, 2025. arXiv:2502.03008.

[BG70] G. I. Bell and S. Glasstone. *Nuclear Reactor Theory*. Van Nostrand Reinhold, New York, 1970.

[BGK54] P. L. Bhatnagar, E. P. Gross, and M. Krook. A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. *Physical Review*, 94(3):511–525, 1954. doi:10.1103/PhysRev.94.511.

[BGL91] C. Bardos, F. Golse, and C. D. Levermore. Fluid dynamic limits of kinetic equations. I. Formal derivations. *Journal of Statistical Physics*, 63:323–344, 1991. doi:10.1007/BF01026608.

[BGL00] C. Bardos, F. Golse, and C. D. Levermore. The acoustic limit for the Boltzmann equation. *Archive for Rational Mechanics and Analysis*, 153:177–204, 2000. doi:10.1007/s002050000080.

[BJ09] G. Bal and A. Jollivet. Time-dependent angularly averaged inverse transport. *Inverse problems*, 25(7):075010, 2009. doi:10.1088/0266-5611/25/7/075010.

[BK89] H. T. Banks and K. Kunisch. *Estimation Techniques for Distributed Parameter Systems*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, 1989. doi:10.1007/978-1-4612-3700-6.

[BLA86] H. T. Banks, P. K. Lamm, and E. S. Armstrong. Spline-based distributed system identification with application to large space antennas. *Journal of Guidance, Control, and Dynamics*, 9(3):304–311, 1986. doi:10.2514/3.20107.

[Bol72] L. Boltzmann. Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. *Sitzungsberichte der Akademie der Wissenschaften Wien*, 66:275–370, 1872.

[Bou23] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, Cambridge, 2023. doi:10.1017/9781009166164.

[BS00] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2000. doi:10.1007/978-1-4612-1394-9.

[CC90] S. Chapman and T. G. Cowling. *The mathematical theory of non-uniform gases*. Cambridge University Press, Cambridge, third edition, 1990.

[CCEY20] J. A. Cañizo, C. Cao, J. Evans, and H. Yoldaş. Hypocoercivity of linear kinetic equations via Harris's Theorem. *Kinetic and Related Models*, 13(1):97–128, 2020. doi:10.3934/krm.2020004.

[CEKL24] G. Ceruti, L. Einkemmer, J. Kusch, and C. Lubich. A robust second-order low-rank BUG integrator based on the midpoint rule. *BIT Numerical Mathematics*, 64:30, 2024. doi:10.1007/s10543-024-01032-x.

[Cer88] C. Cercignani. *The Boltzmann Equation and Its Applications*, volume 67 of *Applied Mathematical Sciences*. Springer, New York, 1988. doi:10.1007/978-1-4612-1039-9.

[CFK22] G. Ceruti, M. Frank, and J. Kusch. Dynamical low-rank approximation for Marshak waves. CRC 1173 Preprint 2022/76, Karlsruhe Institute of Technology, Karlsruhe, 2022.

[CFKK19] T. Camminady, M. Frank, K. Küpper, and J. Kusch. Ray effect mitigation for the discrete ordinates method through quadrature rotation. *Journal of Computational Physics*, 382:105–123, 2019. doi:10.1016/j.jcp.2019.01.016.

[CFL28] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100:32–74, 1928. doi:10.1007/BF01448839.

[CFvN50] J. G. Charney, R. Fjörtoft, and J. von Neumann. Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254, 1950. doi:10.3402/tellusa.v2i4.8607.

[Cha16] S. Chapman. The kinetic theory of simple and composite monatomic gases: viscosity, thermal conduction, and diffusion. *Proceedings of the Royal Society of London. Section A. - Mathematical and Physical Sciences*, 93(646):1–20, 1916. doi:10.1098/rspa.1916.0046.

[Cha60] S. Chandrasekhar. *Radiative Transfer*. Dover Publications, New York, 1960.

[CIP94] C. Cercignani, R. Illner, and M. Pulvirenti. *The Mathematical Theory of Dilute Gases*, volume 106 of *Applied Mathematical Sciences*. Springer, New York, 1994. doi:10.1007/978-1-4419-8524-8.

[CKL22] G. Ceruti, J. Kusch, and C. Lubich. A rank-adaptive robust integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 62:1149–1174, 2022. doi:10.1007/s10543-021-00907-7.

[CKL24] G. Ceruti, J. Kusch, and C. Lubich. A parallel rank-adaptive integrator for dynamical low-rank approximation. *SIAM Journal on Scientific Computing*, 46(3):B205–B228, 2024. doi:10.1137/23M1565103.

[CL22] G. Ceruti and C. Lubich. An unconventional robust integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 62(1):23–44, 2022. doi:10.1007/s10543-021-00873-0.

[CLL18] K. Chen, Q. Li, and J.-G. Liu. Online learning in optical tomography: a stochastic approach. *Inverse Problems*, 34(7):075010, 2018. doi:10.1088/1361-6420/aac220.

[CLW18] K. Chen, Q. Li, and L. Wang. Stability of inverse transport equation in diffusion scaling and Fokker–Planck limit. *SIAM Journal on Applied Mathematics*, 78(5):2626–2647, 2018. doi:10.1137/17M1157969.

[CN47] J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, 43(1):50–67, 1947. doi:10.1017/S0305004100023197.

[Cox72] M. G. Cox. The numerical evaluation of B-splines. *IMA Journal of Applied Mathematics*, 10(2):134–149, 1972. doi:10.1093/imamat/10.2.134.

[CR16] R. E. Castillo and H. Rafeiro. *An Introductory Course in Lebesgue Spaces*. CMS Books in Mathematics. Springer Nature, Cham, 2016. doi:10.1007/978-3-319-30034-4.

[CZ67] K. M. Case and P. F. Zweifel. *Linear Transport Theory*. Addison-Wesley Series in Nuclear Engineering. Addison-Wesley, Reading, MA, 1967.

[d'A47] J.-B. d'Alembert. Récherches sur la courbe que forme une corde tenduë mise en vibration. *Histoire de l'académie royale des sciences et belles lettres de Berlin*, 3:214–219, 1747.

[Daf16] C. M. Dafermos. *Hyperbolic Conservation Laws in Continuum Physics*, volume 325 of *Grundlehren der mathematischen Wissenschaften. A Series of Comprehensive Studies in Mathematics*. Springer, Berlin, Heidelberg, fourth edition, 2016. doi:10.1007/978-3-662-49451-6.

[dB72] C. de Boor. On calculating with B-splines. *Journal of Approximation Theory*, 6(1):50–62, 1972. doi:10.1016/0021-9045(72)90080-9.

[dB78] C. de Boor. *A Practical Guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Springer, New York, 1978.

[Deg04] P. Degond. Macroscopic limits of the Boltzmann equation: a review. In P. Degond, L. Pareschi, and G. Russo, editors, *Modeling and Computational Methods for Kinetic Equations*, pages 3–57. Springer Science+Business Media, New York, 2004. doi:10.1007/978-0-8176-8200-2_1.

[DL21] Z. Ding and L. Einkemmer Q. Li. Dynamical low-rank integrator for the linear Boltzmann equation: error analysis in the diffusion limit. *SIAM Journal on Numerical Analysis*, 59(4):2254–2285, 2021. doi:10.1137/20M1380788.

[DP14] G. Dimarco and L. Pareschi. Numerical methods for kinetic equations. *Acta Numerica*, 23:369–520, 2014. doi:10.1017/S0962492914000063.

[DR84] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Computer Science and Applied Mathematics. Academic Press, New York, second edition, 1984. doi:10.1016/C2013-0-10566-1.

[EHK24] L. Einkemmer, J. Hu, and J. Kusch. Asymptotic-preserving and energy stable dynamical low-rank approximation. *SIAM Journal on Numerical Analysis*, 62(1):73–92, 2024. doi:10.1137/23M1547603.

[EHW21] L. Einkemmer, J. Hu, and Y. Wang. An asymptotic-preserving dynamical low-rank method for the multi-scale multi-dimensional linear transport equation. *Journal of Computational Physics*, 439:110353, 2021. doi:10.1016/j.jcp.2021.110353.

[EHY21] L. Einkemmer, J. Hu, and L. Ying. An efficient dynamical low-rank algorithm for the Boltzmann-BGK equation close to the compressible viscous flow regime. *SIAM Journal on Scientific Computing*, 43(5):B1057–B1080, 2021. doi:10.1137/21M1392772.

[EJ21] L. Einkemmer and I. Joseph. A mass, momentum, and energy conservative dynamical low-rank scheme for the Vlasov equation. *Journal of Computational Physics*, 443:110493, 2021. doi:10.1016/j.jcp.2021.110495.

[EKK+25] L. Einkemmer, K. Kormann, J. Kusch, R. G. McClarren, and J.-M. Qiu. A review of low-rank methods for time-dependent kinetic simulations. *Journal of Computational Physics*, 538:114191, 2025. doi:10.1016/j.jcp.2025.114191.

[EKS23] L. Einkemmer, J. Kusch, and S. Schotthöfer. Conservation properties of the augmented basis update & Galerkin integrator for kinetic problems. *Procedia Computer Science*, 2023. doi:http://dx.doi.org/10.2139/ssrn.4668132.

[EL18] L. Einkemmer and C. Lubich. A low-rank projector-splitting integrator for the Vlasov-Poisson equation. *SIAM Journal on Scientific Computing*, 40(5):B1330–B1360, 2018. doi:10.1137/18M116383X.

[EL19] L. Einkemmer and C. Lubich. A quasi-conservative dynamical low-rank algorithm for the Vlasov equation. *SIAM Journal on Scientific Computing*, 41(5):B1061–B1081, 2019. doi:10.1137/18M1218686.

[ELWY24] L. Einkemmer, Q. Li, L. Wang, and Y. Yang. Suppressing instability in a Vlasov–Poisson system by an external electric field through constrained optimization. *Journal of Computational Physics*, 498:112662, 2024. doi:10.1016/j.jcp.2023.112662.

[EMP24] L. Einkemmer, J. Mangott, and M. Prugger. A low-rank complexity reduction algorithm for the high-dimensional kinetic chemical master equation. *Journal of Computational Physics*, 503:112827, 2024. doi:10.1016/j.jcp.2024.112827.

[Ens17] D. Enskog. *Kinetische Theorie der Vorgänge in mässig verdünnten Gasen*. Almqvist & Wiksell, Uppsala, 1917.

[EOP20] L. Einkemmer, A. Ostermann, and C. Piazzola. A low-rank projector-splitting integrator for the Vlasov–Maxwell equations with divergence correction. *Journal of Computational Physics*, 403:109063, 2020. doi:10.1016/j.jcp.2019.109063.

[EOS23] L. Einkemmer, A. Ostermann, and C. Scalone. A robust and conservative dynamical low-rank algorithm. *Journal of Computational Physics*, 484:112060, 2023. doi:10.1016/j.jcp.2023.112060.

[EP04] R. Esposito and M. Pulvirenti. Chapter 1 – From Particles to Fluids. In S. Friedlander and D. Serre, editors, *Handbook of Mathematical Fluid Dynamics*, volume 3, pages 1–82. Elsevier, Amsterdam, 2004. doi:10.1016/S1874-5792(05)80004-7.

[Eul57] L. Euler. Principes généraux de l'état d'équilibre des fluides. *Mémoires de l'académie des sciences de Berlin*, 11:217–273, 1757.

[Eul68] L. Euler. Institutionum calculi integralis. *Petropoli impensis academiae imperialis scientiarum*, 1:1–542, 1768.

[Eva10] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, second edition, 2010. doi:10.1090/gsm/019.

[Eva21] J. Evans. Hypocoercivity in Phi-entropy for the linear relaxation Boltzmann equation on the torus. *SIAM Journal on Mathematical Analysis*, 53(2):1357–1378, 2021. doi:10.1137/19M1277631.

[FHJ12] F. Filbet, J. Hu, and S. Jin. A numerical scheme for the quantum Boltzmann equation with stiff collision terms. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(2):443–463, 2012. doi:10.1051/m2an/2011051.

[FKCH20] M. Frank, J. Kusch, T. Camminady, and C. D. Hauck. Ray effect mitigation for the discrete ordinates method using artificial scattering. *Nuclear Science and Engineering*, 194(11):971–988, 2020. doi:10.1080/00295639.2020.1730665.

[FKP25] M. Frank, J. Kusch, and C. Patwardhan. Asymptotic-preserving and energy stable dynamical low-rank approximation for thermal radiative transfer equations. *Multiscale Modeling & Simulation*, 23(1):278–312, 2025. doi:10.1137/24M1646303.

[Fou08] J. B. J. Fourier. Mémoire sur la propagation de la chaleur dans les corps solides. *Nouveau Bulletin des Sciences par la Société Philomathique*, 1(6):112–116, 1808.

[Fou22] J. B. J. Fourier. *Théorie analytique de la chaleur*. Didot, Paris, 1822.

[Fri54] K. O. Friedrichs. Symmetric hyperbolic linear differential equations. *Communications on Pure and Applied Mathematics*, 7(2):345–392, 1954. doi:10.1002/cpa.3160070206.

[Gan08] B. D. Ganapol. *Analytical Benchmarks for Nuclear Engineering Applications. Case Studies in Neutron Transport Theory*. Nuclear Energy Agency, Organisation for Economic Co-operation and Development, 2008.

[GBD+01] B. D. Ganapol, R. S. Baker, J. A. Dahl, R. E. Alcouffe, and Los Alamos National Laboratory. Homogeneous infinite media time-dependent analytical benchmarks. In *International Meeting on Mathematical Methods for Nuclear Applications*, Salt Lake City, 2001.

[GK99] C. Geiger and C. Kanzow. *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer-Lehrbuch. Springer, Berlin, Heidelberg, 1999. doi:10.1007/978-3-642-58582-1.

[GKO13] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time-Dependent Problems and Difference Methods*, volume 123 of *Pure and Applied Mathematics*. Wiley, Hoboken, second edition, 2013. doi:10.1002/9781118548448.

[Gol05] F. Golse. Chapter 3 – The Boltzmann Equation and Its Hydrodynamic Limits. In C. M. Dafermos and E. Feireisl, editors, *Handbook of Differential Equations: Evolutionary Equations*, volume 2, pages 159–301. Elsevier, Amsterdam, 2005. doi:10.1016/S1874-5717(06)80006-X.

[GQ24] W. Guo and J.-M. Qiu. A conservative low rank tensor method for the Vlasov dynamics. *SIAM Journal on Scientific Computing*, 46(1):A232–A263, 2024. doi:10.1137/22M1473960.

[Gra49] H. Grad. On the kinetic theory of rarefied gases. *Communications on Pure and Applied Mathematics*, 2(4):331–407, 1949. doi:10.1002/cpa.3160020403.

[Gro93] C. W. Groetsch. *Inverse Problems in the Mathematical Sciences*. Vieweg+Teubner Verlag, Wiesbaden, 1993. doi:10.1007/978-3-322-99202-4.

[Had02] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13(4):49–52, 1902.

[Had23] J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Yale University Press, New Haven, 1923.

[Hel25] K. Hellmuth. *On qualitative experimental design for PDE parameter identification inverse problems*. PhD thesis, University of Würzburg, Würzburg, 2025. doi:10.25972/OPUS-40309.

[HH13] K. Höllig and J. Hörner. *Approximation and Modeling with B-Splines*. SIAM, Philadelphia, 2013. doi:10.1137/1.9781611972955.

[HHK+21] J. Haack, C. Hauck, C. Klingenberg, M. Pirner, and S. Warnecke. A consistent BGK model with velocity-dependent collision frequency for gas mixtures. *Journal of Statistical Physics*, 184(31), 2021. doi:10.1007/s10955-021-02821-2.

[Hil12] D. Hilbert. Begründung der kinetischen Gastheorie. *Mathematische Annalen*, 72:562–577, 1912. doi:10.1007/BF01456676.

[HJM24] X. Huan, J. Jagalur, and Y. Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840, 2024. doi:10.1017/S0962492924000023.

[HKLT25] K. Hellmuth, C. Klingenberg, Q. Li, and M. Tang. Reconstructing the kinetic chemotaxis kernel using macroscopic data: Well-posedness and ill-posedness. *SIAM Journal on Applied Mathematics*, 85(2):613–635, 2025. doi:10.1137/24M165507X.

[HLW06] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, second edition, 2006. doi:10.1007/3-540-30666-8.

[HMA81] K. R. Hogstrom, M. D. Mills, and P. R. Almond. Electron beam dose calculations. *Physics in Medicine & Biology*, 26(3):445–459, 1981. doi:10.1088/0031-9155/26/3/008.

[HMDS20] J. R. Howell, M. P. Mengüc, K. Daun, and R. Siegel. *Thermal Radiation Heat Transfer*. CRC Press, Boca Raton, seventh edition, 2020. doi:10.1201/9780429327308.

[Hol66] L. H. Holway. New statistical models for kinetic theory: Methods of construction. *Physics of Fluids*, 9(9):1658–1673, 1966. doi:10.1063/1.1761920.

[HPUU08] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, Dordrecht, 2008. doi:10.1007/978-1-4020-8839-1.

[HW96] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, second edition, 1996. doi:10.1007/978-3-642-05221-7.

[HW22] J. Hu and Y. Wang. An adaptive dynamical low rank method for the nonlinear Boltzmann equation. *Journal of Scientific Computing*, 92:75, 2022. doi:10.1007/s10915-022-01934-4.

[IK66] E. Isaacson and H. B. Keller. *Analysis of Numerical Methods*. Wiley, New York, 1966.

[Jäc05] P. Jäckel. A note on multivariate Gauss-Hermite quadrature. *ABN AMRO*, 2005.

[KEC23] J. Kusch, L. Einkemmer, and G. Ceruti. On the stability of robust dynamical low-rank approximations for hyperbolic problems. *SIAM Journal on Scientific Computing*, 45(1):A1–A24, 2023. doi:10.1137/21M1446289.

[KH02] A. D. Klose and A. H. Hielscher. Optical tomography using the time-independent equation of radiative transfer – Part 2: inverse model. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 72(5):715–732, 2002. doi:10.1016/S0022-4073(01)00151-0.

[Kir21] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Applied Mathematical Sciences. Springer, Cham, third edition, 2021. doi:10.1007/978-3-030-63343-1.

[KL07] O. Koch and C. Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):434–454, 2007. doi:10.1137/050639703.

[KLW16] E. Kieri, C. Lubich, and H. Walach. Discretized dynamical low-rank approximation in the presence of small singular values. *SIAM Journal on Numerical Analysis*, 54(2):1020–1038, 2016. doi:10.1137/15M1026791.

[KNBH02] A. D. Klose, U. Netz, J. Beuthan, and A. H. Hielscher. Optical tomography using the time-independent equation of radiative transfer – Part 1: forward model. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 72(5):691–713, 2002. doi:10.1016/S0022-4073(01)00150-9.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016*, volume 9851 of *Lecture Notes in Artificial Intelligence*, pages 795–811. Springer, Cham, 2016. doi:10.1007/978-3-319-46128-1_50.

[Kre63] H.-O. Kreiss. Über implizite Differenzmethoden für partielle Differentialgleichungen. *Numerische Mathematik*, 5:24–47, 1963. doi:10.1007/BF01385876.

[KS70] E. F. Keller and L. A. Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26(3):399–415, 1970. doi:10.1016/0022-5193(70)90092-5.

[KS16] K. Kormann and E. Sonnendrücker. Sparse grids for the Vlasov–Poisson equation. In J. Garcke and D. Pflüger, editors, *Sparse Grids and Applications – Stuttgart 2014. Lecture Notes in Computational Science and Engineering*, volume 109, pages 163–190, Cham, 2016. Springer. doi:10.1007/978-3-319-28262-6_7.

[KS23] J. Kusch and P. Stammer. A robust collision source method for rank adaptive dynamical low-rank approximation in radiation therapy. *ESAIM: M2AN*, 57(2):865–891, 2023. doi:10.1051/m2an/2022090.

[Kus20] J. Kusch. *Realizability-preserving discretization strategies for hyperbolic and kinetic equations with uncertainty*. PhD thesis, Karlsruhe Institute of Technology, Karlsruhe, 2020. doi:10.5445/IR/1000121168.

[Kus25] J. Kusch. Second-order robust parallel integrators for dynamical low-rank approximation. *BIT Numerical Mathematics*, 65:31, 2025. doi:10.1007/s10543-025-01073-w.

[Lat68] K. D. Lathrop. Ray effects in discrete ordinates equations. *Nuclear Science and Engineering*, 32(3):357–369, 1968. doi:10.13182/NSE68-4.

[Lat71] K. D. Lathrop. Remedies for ray effects. *Nuclear Science and Engineering*, 45(3):255–268, 1971. doi:10.13182/NSE45-03-255.

[Lax61] P. D. Lax. On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients. *Communications on Pure and Applied Mathematics*, 14(3):497–520, 1961. doi:10.1002/cpa.3160140324.

[Lee60] M. Lees. Energy inequalities for the solution of differential equations. *Transactions of the American Mathematical Society*, 94(1):58–73, 1960. doi:10.2307/1993277.

[LeV92] R. J. LeVeque. *Numerical Methods for Conservation Laws*. Lectures in Mathematics. ETH Zürich. Birkhäuser, Basel, second edition, 1992. doi:10.1007/978-3-0348-8629-1.

[LeV02] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002. doi:10.1017/CBO9780511791253.

[LeV07] R. J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. SIAM, Philadelphia, 2007. doi:10.1137/1.9780898717839.

[LO84] D. R. Lynch and C. B. Officer. Nonlinear parameter estimation for sediment cores. *Chemical Geology*, 44(1):203–225, 1984. doi:10.1016/0009-2541(84)90073-1.

[LO14] C. Lubich and I. V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT Numerical Mathematics*, 54(1):171–188, 2014. doi:10.1007/s10543-013-0454-0.

[Lou89] A. K. Louis. *Inverse und schlecht gestellte Problem*. Teubner Studienbücher Mathematik. Vieweg+Teubner Verlag, Wiesbaden, 1989. doi:10.1007/978-3-322-84808-6.

[LR56] P. D. Lax and R. D. Richtmeyr. Survey of the stability of linear finite difference equations. *Communications on Pure and Applied Mathematics*, 9(2):267–293, 1956. doi:10.1002/cpa.3160090206.

[LWY23] Q. Li, L. Wang, and Y. Yang. Monte Carlo gradient in optimization constrained by radiative transport equation. *SIAM Journal on Numerical Analysis*, 61(6):2744–2774, 2023. doi:10.1137/22M1524515.

[Mar58] R. E. Marshak. Effect of radiation on shock wave behavior. *Physics of Fluids*, 1(1):24–29, 1958. doi:10.1063/1.1724332.

[Mar21] S. Markfelder. *Convex Integration Applied to the Multi-Dimensional Compressible Euler Equations*, volume 2294 of *Lecture Notes in Mathematics*. Springer, Cham, 2021. doi:10.1007/978-3-030-83785-3.

[Mat99] K. A. Mathews. On the propagation of rays in discrete ordinates. *Nuclear Science and Engineering*, 132(2):155–180, 1999. doi:10.13182/NSE99-A2057.

[Max67] J. C. Maxwell. On the dynamical theory of gases. *Philosophical Transactions of the Royal Society of London*, 157:49–88, 1867. doi:10.1098/rstl.1867.0004.

[MELD08] R. G. McClarren, T. M. Evans, R. B. Lowrie, and J. D. Densmore. Semi-implicit time integration for P_n thermal radiative transfer. *Journal of Computational Physics*, 227(16):7561–7586, 2008. doi:10.1016/j.jcp.2008.04.029.

[MHB08a] R. G. McClarren, J. P. Holloway, and T. A. Brunner. Analytic P_1, solutions for time-dependant, thermal radiative transfer in several geometries. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 109(3):389–403, 2008. doi:10.1016/j.jqsrt.2007.08.006.

[MHB08b] R. G. McClarren, J. P. Holloway, and T. A. Brunner. On solutions to the P_n equations for thermal radiative transfer. *Journal of Computational Physics*, 227(5):2864–2885, 2008. doi:10.1016/j.jcp.2007.11.027.

[MM05] K. W. Morton and D. F. Mayers. *Numerical Solution of Partial Differential Equations*. Cambridge University Press, Cambridge, second edition, 2005. doi:10.1017/CBO9780511812248.

[MWLP03] J. E. Morel, T. A. Wareing, R. B. Lowrie, and D. K. Parsons. Analysis of ray-effect mitigation techniques. *Nuclear Science and Engineering*, 144(1):1–22, 2003. doi:10.13182/NSE01-48.

[Nat86] F. Natterer. *The Mathematics of Computerized Tomography*. Vieweg+Teubner Verlag, Wiesbaden, 1986. doi:10.1007/978-3-663-01409-6.

[Nav22] C. L. M. H. Navier. Sur les lois des mouvements des fluides, en ayant égard à l'adhésion des molecules. *Annales de Chimie et de Physique*, 19:244–260, 1822.

[Nav27] C. L. M. H. Navier. Memoire sur les lois du mouvement des fluides. *Mémoires de l'Académie royale des sciences de l'Institut de France*, 6:389–440, 1827.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization. A Basic Course*. Applied Optimization. Springer, New York, 2004. doi:10.1007/978-1-4419-8853-9.

[New87] I. Newton. *Philosophiae naturalis principia mathematica*. Societatis Regiae, 1687.

[Nol87] G. Nolet. Seismic wave propagation and seismic tomography. In G. Nolet, editor, *Seismic Tomography. With Applications in Global Seismology and Exploration Geophysics*, volume 5 of *Modern Approaches in Geophysics*, pages 1–23. Springer, Dordrecht, 1987. doi:10.1007/978-94-009-3899-1_1.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006. doi:10.1007/978-0-387-40065-5.

[OAH00] G. L. Olson, L. H. Auer, and L. Hall. Diffusion, P_1, and other approximate forms of radiation transport. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 62(6):619–634, 2000. doi:10.1016/S0022-4073(99)00150-8.

[PEL23] M. Prugger, L. Einkemmer, and C. F. Lopez. A dynamical low-rank approach to solve the chemical master equation for biological reaction networks. *Journal of Computational Physics*, 489:112250, 2023. doi:10.1016/j.jcp.2023.112250.

[Per89] B. Perthame. Global existence to the BGK model of Boltzmann equation. *Journal of Differential Equations*, 82(1):191–205, 1989. doi:10.1016/0022-0396(89)90173-3.

[Per90] B. Perthame. Boltzmann type schemes for gas dynamics and the entropy property. *SIAM Journal on Numerical Analysis*, 27(6):1405–1421, 1990. doi:10.1137/0727081.

[Pia19] C. Piazzola. *Dynamical low-rank approaches for differential equations*. PhD thesis, University of Innsbruck, Innsbruck, 2019.

[Pir18] M. Pirner. *Kinetic modelling of gas mixtures*. Würzburg University Press, Würzburg, 2018. doi:10.25972/WUP-978-3-95826-081-8.

[PK25] C. Patwardhan and J. Kusch. A parallel, energy-stable low-rank integrator for nonlinear multi-scale thermal radiative transfer, 2025. arXiv:2502.20883.

[Ple06] R.-E. Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, 2006. doi:10.1111/j.1365-246X.2006.02978.x.

[PM21] Z. Peng and R. G. McClarren. A high-order/low-order (HOLO) algorithm for preserving conservation in time-dependent low-rank transport calculations. *Journal of Computational Physics*, 447:110672, 2021. doi:10.1016/j.jcp.2021.110672.

[PM23] Z. Peng and R. G. McClarren. A sweep-based low-rank method for the discrete ordinate transport equation. *Journal of Computational Physics*, 473:111748, 2023. doi:10.1016/j.jcp.2022.111748.

[PMF20] Z. Peng, R. G. McClarren, and M. Frank. A low-rank method for two-dimensional time-dependent radiation transport calculations. *Journal of Computational Physics*, 421:109735, 2020. doi:10.1016/j.jcp.2020.109735.

[Pol63] B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. doi:10.1016/0041-5553(63)90382-3.

[Pom73] G. C. Pomraning. *The Equations of Radiation Hydrodynamics*, volume 54 of *International series of monographs in natural philosophy*. Pergamon Press, Oxford, 1973.

[Pom79] G. C. Pomraning. The non-equilibrium marshak wave problem. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 21(3):249–261, 1979. doi:10.1016/0022-4073(79)90016-5.

[PP93] B. Perthame and M. Pulvirenti. Weighted L^∞ bounds and uniqueness for the Boltzmann BGK model. *Archive for Rational Mechanics and Analysis*, 125:289–295, 1993. doi:10.1007/BF00383223.

[QSS02] A. Quarteroni, R. Sacco, and F. Saleri. *Numerische Mathematik 2*. Springer-Lehrbuch. Springer, Berlin, Heidelberg, 2002. doi:10.1007/978-3-642-56191-7.

[RBH07] K. Ren, G. Bal, and A. H. Hielscher. Transport- and diffusion-based optical tomography in small domains: a comparative study. *Applied Optics*, 46(27):6669–6679, 2007. doi:10.1364/AO.46.006669.

[Ren10] K. Ren. Recent developments in numerical techniques for transport-based medical imaging methods. *Communications in Computational Physics*, 8(1):1–50, 2010.

[RM67] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial-Value Problems*, volume 4 of *Interscience Tracts in Pure and Applied Mathematics*. Wiley, New York, second edition, 1967.

[SB02] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*, volume 12 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2002. doi:10.1007/978-0-387-21738-3.

[Sch07] L. L. Schumaker. *Spline Functions: Basic Theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, third edition, 2007. doi:10.1017/CBO9780511618994.

[Sch15] L. L. Schumaker. *Spline Functions: Computational Methods*. SIAM, Phildelphia, 2015. doi:10.1137/1.9781611973907.

[Sch22] S. Schrammer. *On dynamical low-rank integrators for matrix differential equations*. PhD thesis, Karlsruhe Institute of Technology, Karlsruhe, 2022. doi:10.5445/IR/1000148853.

[SEKM25] C. Scalone, L. Einkemmer, J. Kusch, and R. G. McClarren. A multi-fidelity adaptive dynamical low-rank based optimization algorithm for fission criticality problems. *Journal of Scientific Computing*, 104:27, 2025. doi:10.1007/s10915-025-02907-z.

[SN14] M. Svärd and J. Nordström. Review of summation-by-parts schemes for initial–boundary-value problems. *Journal of Computational Physics*, 268:17–38, 2014. doi:10.1016/j.jcp.2014.02.031.

[SO97] B. Su and G. L. Olson. An analytical benchmark for non-equilibrium radiative transfer in an isotropically scattering medium. *Annals of Nuclear Energy*, 24(13):1035–1055, 1997. doi:10.1016/S0306-4549(96)00100-4.

[Son19] E. Sonnendrücker. *Numerical Methods for the Vlasov-Maxwell equations*. Book manuscript received by the author himself, 2019.

[Ste03] P. Stefanov. Inverse problems in transport theory. In G. Uhlmann, editor, *Inside Out: Inverse Problems and Applications*, volume 47 of *Mathematical Sciences Research Institute Publications*, pages 111–131. Cambridge University Press, Cambridge, 2003. doi:10.1017/9781009701310.005.

[Sto45] G. G. Stokes. On the theories of the internal friction of fluids in motion and of the equilibrium and motion of elastic solids. *Transactions of the Cambridge Philosophical Society*, 8:287–305, 1845.

[Str71] A. H. Stroud. *Approximate Calculation of Multiple Integrals*. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Englewood Cliffs, 1971.

[Str97] H. Struchtrup. The BGK-model with velocity-dependent collision frequency. *Continuum Mechanics and Thermodynamics*, 9:23–31, 1997. doi:10.1007/s001610050053.

[Str04] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. SIAM, Phildadelphia, second edition, 2004. doi:10.1137/1.9780898717938.

[Str05] H. Struchtrup. *Macroscopic Transport Equations for Rarefied Gas Flows. Approximation Methods in Kinetic Theory*. Springer, Berlin, Heidelberg, 2005. doi:10.1007/3-540-32386-4.

[Tar05] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Other Titles in Applied Mathematics. SIAM, Philadelphia, 2005. doi:10.1137/1.9780898717921.

[Ten16] J. Tencer. Ray effect mitigation through reference frame rotation. *Journal of Heat Transfer*, 138(11):112701, 2016. doi:10.1115/1.4033699.

[Tho95] J. W. Thomas. *Numerical Partial Differential Equations: Finite Difference Methods*, volume 22 of *Texts in Applied Mathematics*. Springer, New York, 1995. doi:10.1007/978-1-4899-7278-1.

[Tit48] E. C. Titchmarsh. *Introduction To The Theory Of Fourier Integrals*. Clarendon Press, Oxford, second edition, 1948.

[Trö10] F. Tröltzsch. *Optimal Control of Partial Differential Equations. Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2010. doi:10.1090/gsm/112.

[UA82] S. Ukai and K. Asano. On the Cauchy problem of the Boltzmann equation with a soft potential. *Publications of the Research Institute for Mathematical Sciences*, 18(2):57–99, 1982. doi:10.2977/prims/1195183569.

[UU12] M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung*. Mathematik Kompakt. Birkhäuser, Basel, 2012. doi:10.1007/978-3-0346-0654-7.

[Vil02] C. Villani. Chapter 2 – A Review of Mathematical Topics in Collisional Kinetic Theory. In S. Friedlander and D. Serre, editors, *Handbook of Mathematical Fluid Dynamics*, volume 1, pages 71–305. Elsevier, Amsterdam, 2002. doi:10.1016/S1874-5792(02)80004-0.

[Vog02] C. R. Vogel. *Computational Methods for Inverse Problems*. Frontiers in Applied Mathematics. SIAM, Phildelphia, 2002. doi:10.1137/1.9780898717570.

[War22] S. Warnecke. *Numerical schemes for multi-species BGK equations based on a variational procedure. Applied to multi-species BGK equations with velocity-dependent collision frequency and to quantum multi-species BGK equations*. Würzburg University Press, Würzburg, 2022. doi:10.25972/WUP-978-3-95826-193-8.

[WR22] S. J. Wright and B. Recht. *Optimization for Data Analysis*. Applied Optimization. Cambridge University Press, Cambridge, 2022. doi:10.1017/9781009004282.

[WSA07] S. Wright, M. Schweiger, and S. R. Arridge. Reconstruction in optical tomography using the P_N approximations. *Measurement Science and Technology*, 18(1):79–86, 2007. doi:10.1088/0957-0233/18/1/010.

[YEHS24] P. Yin, E. Endeve, C. D. Hauck, and S. R. Schnake. Towards dynamical low-rank approximation for neutrino kinetic equations. Part I: Analysis of an idealized relaxation model. *Mathematics of Computation*, 94:1199–1233, 2024. doi:10.1090/mcom/3997.

[Yos95] K. Yosida. *Functional Analysis*. Classics in Mathematics. Springer, Berlin, Heidelberg, sixth edition, 1995. doi:10.1007/978-3-642-61859-8.

[ZTZ13] O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu. *The Finite Element Method: Its Basis and Fundamentals*. Elsevier Butterworth-Heinemann, Amsterdam, seventh edition, 2013. doi:10.1016/C2009-0-24909-9.