ESAIM: M2AN 57 (2023) 1225–1255
https://doi.org/10.1051/m2an/2023008

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

TRUNCATION ERRORS AND MODIFIED EQUATIONS FOR THE LATTICE BOLTZMANN METHOD *VIA* THE CORRESPONDING FINITE DIFFERENCE SCHEMES

Thomas Bellotti*

Abstract. Lattice Boltzmann schemes are efficient numerical methods to solve a broad range of problems under the form of conservation laws. However, they suffer from a chronic lack of clear theoretical foundations. In particular, the consistency analysis and the derivation of the modified equations are still open issues. This has prevented, until today, to have an analogous of the Lax equivalence theorem for lattice Boltzmann schemes. We propose a rigorous consistency study and the derivation of the modified equations for any lattice Boltzmann scheme under acoustic and diffusive scalings. This is done by passing from a kinetic (lattice Boltzmann) to a macroscopic (Finite Difference) point of view at a fully discrete level in order to eliminate the non-conserved moments relaxing away from the equilibrium. We rewrite the lattice Boltzmann scheme as a multi-step Finite Difference scheme on the conserved variables, as introduced in our previous contribution. We then perform the usual analyses for Finite Difference by exploiting its precise characterization using matrices of Finite Difference operators. Though we present the derivation of the modified equations until second-order under acoustic scaling, we provide all the elements to extend it to higher orders, since the kinetic-macroscopic connection is conducted at the fully discrete level. Finally, we show that our strategy yields, in a more rigorous setting, the same results as previous works in the literature.

Mathematics Subject Classification. 65M75, 65M06, 65M12, 65M15.

Received May 4, 2022. Accepted January 23, 2023.

1. Introduction

Lattice Boltzmann methods form a vast category of numerical schemes to address the approximation of the solution of Partial Differential Equations (PDEs) under the form of conservation laws, called macroscopic equations or target PDEs. These numerical schemes act in a kinetic fashion by employing a certain number $q \in \mathbb{N}^\star$ of discrete velocities, larger than the number $N \in \mathbb{N}^\star$ of macroscopic equations to be solved. The scheme proceeds *via* a kinetic-like algorithm made up of two distinct steps. The first one is a local non-linear collision phase on each site of the mesh, followed by a lattice-constrained transport which is inherently linear. The local nature of the collision phase allows for massive parallelization of the method and the fact that the "particles" are constrained to dwell on the lattice allows to implement the stream phase as a pointer shift in memory. This results in a very efficient numerical method capable of reaching problems of important size in terms of

Keywords and phrases. Lattice Boltzmann, Finite Difference, truncation error, consistency, modified equations.

CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France.
*Corresponding author: thomas.bellotti@polytechnique.edu

computational and memory cost. The historical seminal papers from the end of the eighties are [26,39], while for a general modern presentation of the lattice Boltzmann schemes and their extremely broad fields of application, including hyperbolic systems of conservation laws, the quasi-incompressible Navier–Stokes equations, multiphase systems and porous media, the interested reader can consult [23,35,49]. The presentation of this plethora of interesting applications is however beyond the scope of our contribution.

To our understanding, the highest price to pay for this highly efficient implementation of the method is the lack of pure theoretical understanding on why the overall procedure works well at approximating the solution of the target PDEs. This is essentially due to the fact that – the standpoint of the lattice Boltzmann schemes being kinetic [46] – the number of discrete velocities is larger than the number of macroscopic equations. Therefore, the formal analyses for lattice Boltzmann schemes available in the literature try to bridge the gap between a kinetic and a macroscopic point of view relying essentially on the quasi-equilibrium of the non-conserved variables. In particular, as far as the consistency with the macroscopic equations and the modified equations are concerned in the limit of small discretization parameters, two main approaches are at our disposal. The first one is based on the Chapman–Enskog expansion [9, 28] from statistical mechanics, shaped to the context of lattice Boltzmann schemes, see for example [10,36,43]. The second approach features the so-called equivalent equations introduced by Dubois [15, 17], consisting in performing a Taylor expansion of the scheme both for the conserved and non-conserved moments and progressively re-inject the developments order-by-order. This approach has proved to yield information in accordance with the numerical simulations, see [2, 18, 19]. Despite their proved empirical reliability and the fact that they yield the same results at the dominant orders (see [16] for instance) these two strategies are both formal, especially for the computation of the truncation errors. Indeed, the Chapman–Enskog expansion relies on the introduction of two time variables with different scalings which are not present in the discrete lattice Boltzmann scheme. Moreover, in this approach and in the method of the equivalent equations, the values of the non-conserved variables are assumed to stem from the pointwise discretization of smooth functions, whose existence and smoothness cannot be guaranteed because they are absent from the target PDEs. Other approaches known in the literature are the asymptotic analysis under parabolic scaling deployed in [30–32] as well as the Maxwell iteration method [53,55], which shares strong bonds with the equivalent equations method presented before. The previous list of formal analysis techniques does not aim at being exhaustive (the interested reader can refer to [35]) and one should be aware that, despite efforts in this direction [7], there is no consensus on which is the right method to use [35].

A staple of all the previously mentioned approaches is that the expansion for the discretization parameters (time and space steps) tending to zero is performed on the kinetic numerical scheme, where both conserved and non-conserved variables are present. Eventually, the non-conserved variables are formally eliminated from the continuous formulation by scaling arguments, so to speak, using quasi-equilibrium. This corresponds to follow the diagonal path on Figure 1. In this contribution, we develop the other path, namely the top-down movement followed by the left-right one on Figure 1. In particular, in order to fill the hollow between lattice Boltzmann schemes and traditional approaches known to numerical analysts, such as Finite Difference schemes, we recently introduced [3] a formalism to recast any lattice Boltzmann scheme, regardless of its linearity, as a multi-step Finite Difference scheme solely on the conserved moments. It should be stressed that our standpoint, where lattice Boltzmann schemes are studied in terms of their Finite Difference counterpart, must not be seen as the right way of implementing them, because one would lose most of the previously mentioned computational efficiency coming from the kinetic vision. Conversely, our way of writing the scheme should be seen as a sort of one-way mathematical transform to pass from a kinetic standpoint to a macroscopic one in a purely discrete setting. The elimination of the non-conserved moments is carried exactly on the discrete formulation by algebraic devices, thus independently from the time-space scaling. The price to pay for the non-conserved moments relaxing away from the equilibria is the multi-step nature of the Finite Difference scheme. In our previous proposal [3], it has been crucial to be able to provide, thanks to a systematic mathematical approach, a precise description of the main ingredient needed to reduce the lattice Boltzmann scheme to a Finite Difference scheme, namely the characteristic polynomial of matrices of Finite Difference operators. We are therefore allowed to utilize this characteristic polynomial as a tool satisfying certain properties alone from the particular underlying

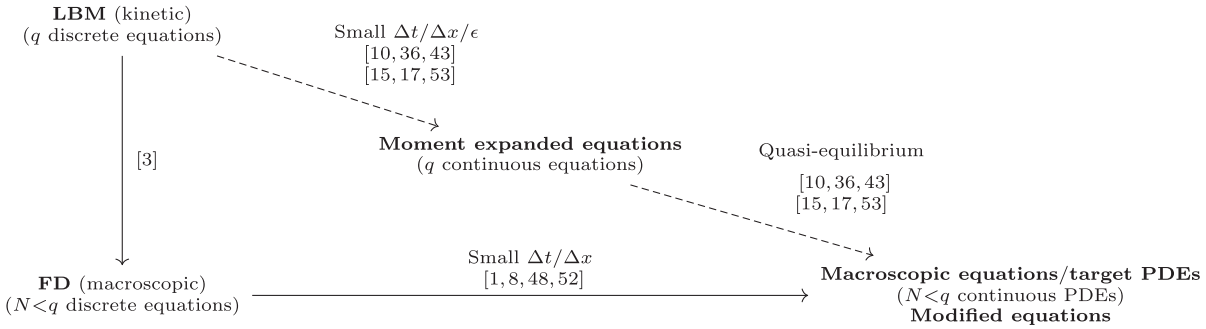

FIGURE 1. Different paths to recover the macroscopic equations and the modified equations. The formal approaches available in the literature [10, 15, 17, 36, 43, 53] rely on the path marked with dashed arrows. They perform Taylor expansions for small discretization parameters and then utilize the quasi-equilibrium of the non-conserved moments to get rid of them. Our way of proceeding is marked with full arrows: we eliminate exactly the non-conserved moments at the discrete level as in [3] and we perform the usual analyses for Finite Difference schemes as in [1, 8, 48, 52].

lattice Boltzmann scheme. Quite the opposite, using the algorithm proposed by Fučík and Straka [22], one is compelled to explicitly write down the corresponding Finite Difference scheme in order to perform the Taylor expansions to recover the target PDEs. In our case, the mathematical understanding that we *a priori* have on the corresponding (macroscopic) Finite Difference schemes, regardless of the (kinetic) lattice Boltzmann scheme they stand for, allows the following theoretical discussion. The theory of Finite Difference schemes features two important notions. One is the concept of truncation error ([24], Def. 5.1.3 or [1], Def. 2.2.4), which is rigorous and is the basic ingredient to prove the celebrated Lax equivalence theorem [38]. The computations of the truncation error are perfectly justified because of the existence and smoothness results on the target PDEs (*e.g.* transport equation with smooth initial datum, Burgers equation with smooth non-decreasing initial datum, *etc.*). The second one is the concept of modified equation [8, 52], which is formal. The modified equations are those which the numerical scheme is "more consistent" with, compared to the target PDEs, and thus they yield essential but formal information on the behavior of the scheme. The modified equations cannot be fully justified even for Finite Difference schemes because they assume that smooth functions which equal the discrete solution of the scheme at the grid points exist.

The main findings presented in the paper are the following.

- We propose a procedure to rigorously analyze the consistency (*i.e.* the truncation error) of any lattice Boltzmann scheme *via* its corresponding Finite Difference scheme.
- In the case of acoustic and diffusive scaling between time and space discretizations, we rigorously find the expression of the target PDEs approximated by any scheme and the truncation error. For the acoustic scaling, we also write the formal modified equations up to order two.
- Under acoustic scaling, these modified equations are the same than the ones obtained by Dubois [17] until second order.
- We rewrite the Maxwell iteration [53, 55] for general lattice Boltzmann schemes. This allows to show that both for the acoustic and diffusive scaling, the modified equations obtained through the corresponding Finite Difference scheme are the same as the ones from the Maxwell iteration at any order.

Our derivation of the truncation errors is rigorous – as the ones for Finite Difference schemes – and the formal modified equations rely on less unjustified assumptions than the existing approaches, for two main reasons. The first one is that Taylor expansions are applied to the conserved moments only, which also appear in the macroscopic equations. Therefore, one only postulates that the discrete conserved moments stem from

the point-wise evaluation of smooth functions. The second one is that we solely rely on the link between time and space steps as the lattices are refined and which must be specified for any time-space numerical method.

The paper is structured as follows. In Section 1, we set notations and assumptions concerning the lattice Boltzmann schemes we shall work with. Section 2 is devoted to recall the main results from our previous work [3] concerning the recast of any lattice Boltzmann scheme as a Finite Difference scheme. These results are then stated in a slightly different manner, facilitating the following analysis. The main results of the work are stated in Section 3 and come under the form of two theorems. The proof of the first is detailed in Section 4 under the assumption of dealing with one conservation law, for the sake of keeping the presentation and the notations as simple as possible. In Section 5, we indicate how the previous proof is easily extended to several conservation laws, whereas Section 6 is devoted to hint the links with some available approaches to find the modified equations of lattice Boltzmann schemes. The conclusions and perspectives of this work are drawn in Section 7.

1. Lattice Boltzmann schemes

To start our contribution, we present the classical framework of the multiple-relaxation-times (known as MRT) lattice Boltzmann schemes, as introduced by D'Humières [13]. For the sake of simplicity, we do not consider source terms which can be effortlessly introduced in the analysis. This fixes the perimeter of the schemes we shall be allowed to treat and study in the sequel.

1.1. Spatial and temporal discretization

We set the problem in spatial dimension $d = 1, 2, 3$ considering the whole space $\mathbb{R}^d$ since this work is not focused on the enforcement of boundary conditions. All the following material is valid only under the assumption of working on unbounded domains or sufficiently far from a boundary. The space is discretized by a d-dimensional lattice denoted $\mathfrak{L} := \Delta x\mathbb{Z}^d$ with constant step $\Delta x > 0$. The time is uniformly discretized with step $\Delta t > 0$, rendering a time lattice $\mathfrak{Z} := \Delta t\mathbb{N}$. The role of the initial conditions is not investigated and is a subject on its own, see [44, 51]. We introduce the so-called "lattice velocity" $\lambda > 0$ defined by $\lambda := \Delta x/\Delta t$. Observe that in the sequel, namely in Section 3, we shall introduce particular relations between space step Δx and Δt when $\Delta x \to 0$, in order to provide the main results of the work, as done in [15,17]. However, until the end of Section 2, the discussion remains valid for any choice of these parameters.

1.2. Discrete velocities

The discrete velocities are an essential ingredient of any lattice Boltzmann scheme. One has to choose $(\boldsymbol{e}_j)_{j=1}^{j=q} \subset \mathbb{R}^d$ with $q \in \mathbb{N}^\star$, discrete velocities, which are multiple of the lattice velocity λ, namely $\boldsymbol{e}_j = \lambda \boldsymbol{c}_j$ for any $j \in [1 \dots q]$[2] with $(\boldsymbol{c}_j)_{j=1}^{j=q} \subset \mathbb{Z}^d$. Thus, the virtual particles are stuck to the lattice $\mathfrak{L}$ at each time step of the method. We denote the distribution density of the virtual particles moving with velocity $\boldsymbol{e}_j$ by $f_j = f_j(t, \boldsymbol{x})$ for every $j \in [1 \dots q]$, depending on the time and space variables.

1.3. Lattice Boltzmann algorithm: collide and stream

As mentioned in the Introduction, any lattice Boltzmann scheme consists in a kinetic algorithm made up of two phases: a local collision phase performed on each site of the lattice $\mathfrak{L}$ and a stream phase where particles are exchanged between different sites of the lattice. Let us independently introduce them.

– *Collision phase.* We adopt the general point of view of the multiple-relaxation-times schemes, where the collision phase is written as a diagonal relaxation towards some equilibria in the moments basis, see [13]. We introduce a change of basis called moment matrix $\boldsymbol{M} \in \mathrm{GL}_q(\mathbb{R})$. Gathering the distributions into $\boldsymbol{f} = (f_1, \dots, f_q)^\intercal$, the moments are recovered by $\boldsymbol{m} = \boldsymbol{M}\boldsymbol{f}$ and *vice versa*. We also introduce

[2] This shall be a notation to indicate closed intervals of integers, namely for $a, b \in \mathbb{Z}$ with $a \leq b$, then $[a \dots b] := \{a, a+1, \dots, b\}$.

- the matrix $\boldsymbol{I} \in \mathrm{GL}_q(\mathbb{R})$, the identity matrix of size q;
- the matrix $\boldsymbol{S} \in \mathcal{M}_q(\mathbb{R})$, called relaxation matrix. This matrix is diagonal with $N \in [1 \dots q-1]$ being the number of conserved moments $\boldsymbol{S} = \mathrm{diag}(s_1, \dots, s_N, s_{N+1}, \dots, s_q)$, where $s_i \in \mathbb{R}$ for $i \in [1 \dots N]$ for the conserved moments and $s_i \in]0, 2]$ for $i \in [N+1 \dots q]$, see [15], for the non-conserved ones. Observe that the relaxation parameters corresponding to the conserved moments do not play any role in the lattice Boltzmann algorithm, therefore the matrix $\boldsymbol{S}$ can be singular, without any specific issue. In particular, we shall prove in Section 2.4 that the choice of relaxation parameter for the conserved variables does not have any influence on the outcomes presented in this work. For the sake of presentation, we start numbering the moments by the conserved ones;
- we employ the notation $\boldsymbol{m}^{\mathrm{eq}}(t, \boldsymbol{x}) = \boldsymbol{m}^{\mathrm{eq}}(m_1(t, \boldsymbol{x}), \dots, m_N(t, \boldsymbol{x}))$ for $t \in \mathfrak{Z}$ and $\boldsymbol{x} \in \mathfrak{L}$, where $\boldsymbol{m}^{\mathrm{eq}} : \mathbb{R}^N \to \mathbb{R}^q$ are possibly non-linear functions of the conserved moments. In order to guarantee that the first N moments are conserved through the collision process, irrespective of the values of $s_1, \dots, s_N$, the constraints
$$m_i^{\mathrm{eq}}(m_1, \dots, m_N) = m_i, \qquad \forall i \in [1 \dots N], \tag{1}$$
must hold [5].

Let $t \in \mathfrak{Z}$ and $\boldsymbol{x} \in \mathfrak{L}$, the collision phase reads, denoting by $\star$ any post-collision state

$$\boldsymbol{m}^{\star}(t, \boldsymbol{x}) = (\boldsymbol{I} - \boldsymbol{S})\boldsymbol{m}(t, \boldsymbol{x}) + \boldsymbol{S}\boldsymbol{m}^{\mathrm{eq}}(t, \boldsymbol{x}). \tag{2}$$

– *Stream phase.* The stream phase is diagonal in the space of the distributions and consist in an exact upwind advection of the particle distribution densities. It can be written, for $t \in \mathfrak{Z}$ and $\boldsymbol{x} \in \mathfrak{L}$, as

$$f_j(t + \Delta t, \boldsymbol{x}) = f_j^{\star}(t, \boldsymbol{x} - \boldsymbol{c}_j \Delta x), \tag{3}$$

for any $j \in [1 \dots q]$.

2. Finite difference formulation of a lattice Boltzmann scheme

Having defined the lattice Boltzmann schemes, we briefly introduce the setting allowing us to rewrite any lattice Boltzmann scheme (kinetic) as a multi-step Finite Difference scheme (macroscopic) on the N conserved moments of interest. The interested reader can refer to our previous contribution [3] for more details. Then, the formulation of the multi-step Finite Difference scheme is given using a more compact notation which is more suitable to the following discussion. We start with the assumptions needed in the sequel.

Assumption 2.1 (Finite Difference assumptions)**.** *The entries of $\boldsymbol{M}$ and $\boldsymbol{S}$ can depend on Δx and/or on Δt but cannot be a function of the time and space variables.*

2.1. Algebraic setting

Let us first introduce the necessary algebraic setting. In particular, we define the shift operators associated with each discrete velocity as well as the derived Finite Difference operators in space. In the following definition, the time variable does not play any role since kept frozen, thus it is not listed for the sake of readability.

Definition 2.2 (Shift and Finite Difference operators in space)**.** Let $\boldsymbol{z} \in \mathbb{Z}^d$, then the associated shift operator on the lattice $\mathfrak{L}$, denoted $\mathtt{t}_{\boldsymbol{z}}$, is defined in the following way. Take $f : \mathfrak{L} \to \mathbb{R}$ be any function defined on the lattice, then the action of $\mathtt{t}_{\boldsymbol{z}}$ is

$$(\mathtt{t}_{\boldsymbol{z}} f)(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{z}\Delta x), \qquad \forall \boldsymbol{x} \in \mathfrak{L}.$$

We also introduce $\mathsf{T} := \{\mathtt{t}_{\boldsymbol{z}} \text{ with } \boldsymbol{z} \in \mathbb{Z}^d\} \cong \mathbb{Z}^d$. The product $\circ : \mathsf{T} \times \mathsf{T} \to \mathsf{T}$ of two shift operators is defined by

$$\mathtt{t}_{\boldsymbol{z}} \circ \mathtt{t}_{\boldsymbol{w}} := \mathtt{t}_{\boldsymbol{z}+\boldsymbol{w}}, \qquad \forall \boldsymbol{z}, \boldsymbol{w} \in \mathbb{Z}^d.$$

The set of Finite Difference operators on the lattice $\mathfrak{L}$ is defined as

$$\mathsf{D} := \mathbb{R}\mathsf{T} = \left\{ \sum\nolimits_{\mathsf{t}\in\mathsf{T}} \alpha_{\mathsf{t}}\mathsf{t}, \quad \text{where} \quad \alpha_{\mathsf{t}} \in \mathbb{R} \quad \text{and} \quad \alpha_{\mathsf{t}} = 0 \quad \text{almost everywhere} \right\}, \tag{4}$$

the group ring (or group algebra) of T over $\mathbb{R}$. The sum $+ : \mathsf{D} \times \mathsf{D} \to \mathsf{D}$ and the product $\circ : \mathsf{D} \times \mathsf{D} \to \mathsf{D}$ of two elements are defined by

$$\left(\sum_{\mathsf{t}\in\mathsf{T}} \alpha_{\mathsf{t}}\mathsf{t}\right) + \left(\sum_{\mathsf{t}\in\mathsf{T}} \beta_{\mathsf{t}}\mathsf{t}\right) = \sum_{\mathsf{t}\in\mathsf{T}} (\alpha_{\mathsf{t}} + \beta_{\mathsf{t}})\mathsf{t}, \qquad \left(\sum_{\mathsf{t}\in\mathsf{T}} \alpha_{\mathsf{t}}\mathsf{t}\right) \circ \left(\sum_{\mathsf{h}\in\mathsf{T}} \beta_{\mathsf{h}}\mathsf{h}\right) = \sum_{\mathsf{t},\mathsf{h}\in\mathsf{T}} (\alpha_{\mathsf{t}}\beta_{\mathsf{h}})(\mathsf{t}\circ\mathsf{h}).$$

Furthermore, the product of $\sigma \in \mathbb{R}$ with elements of D is given by

$$\sigma\left(\sum_{\mathsf{t}\in\mathsf{T}} \alpha_{\mathsf{t}}\mathsf{t}\right) = \sum_{\mathsf{t}\in\mathsf{T}} (\sigma\alpha_{\mathsf{t}})\mathsf{t}.$$

In the sequel, the products $\circ$ are understood.

Remark 2.3. We could achieve exactly the same construction, following ([11], Chapt. 2), by considering functions on the lattice $\mathfrak{L}$ as sequences and the Finite Difference operators as sequences with compact support (whence the almost everywhere requirement in (4)). Then, the product $\circ$ can be seen as a convolution (Cauchy) product between compactly supported sequences and the action of a Finite Difference operator on a function as the convolution of a finitely supported sequence with a generic sequence.

Upon introducing the generating displacements along each axis $\mathsf{x}_k := \mathsf{t}_{\boldsymbol{e}_k}$ where $\boldsymbol{e}_k$ is the k-th vector of the canonical basis, for any $k \in [1 \,..\, d]$, we can isomorphically identify $\mathsf{D} \cong \mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$, the ring of multivariate Laurent polynomials. The real numbers $\mathbb{R}$ can be viewed as sub-ring of D, being the constant polynomials. This identification can be somehow interpreted as the historical starting point of umbral calculus [45], also known as calculus of Finite Differences [40]: allow to interchange indices in sequences (operators or functions) with exponents (in polynomials). The stream phase (3) can be recast under its non-diagonal form in the space of moments [17, 53] by introducing what we call the moments-stream matrix $\mathbf{T} := \boldsymbol{M}\mathrm{diag}(\mathsf{t}_{\boldsymbol{c}_1}, \ldots, \mathsf{t}_{\boldsymbol{c}_q})\boldsymbol{M}^{-1} \in \mathcal{M}_q(\mathsf{D})$ and merged with the collision phase (2) to obtain the scheme, for any $t \in \mathfrak{Z}$ and for any $\boldsymbol{x} \in \mathfrak{L}$

$$\boldsymbol{m}(t + \Delta t, \boldsymbol{x}) = \mathbf{A}\boldsymbol{m}(t, \boldsymbol{x}) + \mathbf{B}\boldsymbol{m}^{\mathrm{eq}}(t, \boldsymbol{x}), \tag{5}$$

where $\mathbf{A} := \mathbf{T}(\boldsymbol{I} - \boldsymbol{S}) \in \mathcal{M}_q(\mathsf{D})$ and $\mathbf{B} := \mathbf{T}\boldsymbol{S} \in \mathcal{M}_q(\mathsf{D})$.

2.2. Corresponding Finite Difference schemes

With this new compact algebraic form of any lattice Boltzmann scheme, namely (5), we are able to recall the main results proved in [3]. These results encompass the findings from [12, 22, 50]. The version for one conserved moment can be formulated as follows.

Proposition 2.4 (Corresponding Finite Difference scheme for $N = 1$, [3])**.** *Consider $N = 1$. Then the lattice Boltzmann scheme given by (5) corresponds to a multi-step explicit macroscopic Finite Difference scheme on the conserved moment m_1 under the form*

$$m_1(t + \Delta t, \boldsymbol{x}) = -\sum_{k=0}^{q-1} \mathsf{c}_k m_1(t + (1 - q + k)\Delta t, \boldsymbol{x}) + \left(\sum_{k=0}^{q-1} \left(\sum_{\ell=0}^{k} \mathsf{c}_{q+\ell-k} \mathbf{A}^{\ell} \right) \mathbf{B}\boldsymbol{m}^{\mathrm{eq}}(t - k\Delta t, \boldsymbol{x}) \right)_1, \tag{6}$$

for all $t \in \mathfrak{Z}$ and for all $\boldsymbol{x} \in \mathfrak{L}$, where $(\mathsf{c}_k)_{k=0}^{k=q} \subset \mathsf{D}$ are the coefficients of $\det(X\boldsymbol{I} - \mathbf{A}) = \sum_{k=0}^{k=q} \mathsf{c}_k X^k$, the characteristic polynomial of $\mathbf{A}$, with $\det(\cdot)$ indicating the determinant of a matrix.

The proof – given in [3] – relies on the fact that D is a commutative ring and that therefore the Cayley–Hamilton theorem [6], stipulating that any square matrix with entries in a commutative ring annihilates its characteristic polynomial, holds.

This result is easily generalized to the case of multiple conservation laws, namely $N > 1$. For this, let us introduce a new notation. For any square matrix $\boldsymbol{C} \in \mathcal{M}_q(\mathfrak{R})$ on a commutative ring $\mathfrak{R}$, consider $\boldsymbol{C}_I := (\sum_{i\in I} \boldsymbol{e}_i \otimes \boldsymbol{e}_i)\boldsymbol{C}(\sum_{i\in I} \boldsymbol{e}_i \otimes \boldsymbol{e}_i) \in \mathcal{M}_q(\mathfrak{R})$ for any $I \subset [1 \dots q]$, corresponding to the matrix where only the entries with row and column indices in I are kept and the remaining ones are set to zero. Then we have the following statement.

Proposition 2.5 (Corresponding Finite Difference schemes for $N \geq 1$, [3])**.** *Consider $N \geq 1$. Then the lattice Boltzmann scheme given by (5) corresponds to a family of multi-step explicit macroscopic Finite Difference schemes on the conserved moments $m_1, \dots, m_N$. This is, for any $i \in [1 \dots N]$*

$$m_i(t+\Delta t, \boldsymbol{x}) = -\sum_{k=0}^{q-N} \mathsf{c}_{i,k} m_i(t+(k-q+N)\Delta t, \boldsymbol{x}) + \left(\sum_{k=0}^{q-N}\left(\sum_{\ell=0}^{k} \mathsf{c}_{i,q+1-N+\ell-k}\mathbf{A}_i^{\ell}\right)\mathbf{A}_i^{\diamond}\boldsymbol{m}(t-k\Delta t, \boldsymbol{x})\right)_i + \left(\sum_{k=0}^{q-N}\left(\sum_{\ell=0}^{k} \mathsf{c}_{i,q+1-N+\ell-k}\mathbf{A}_i^{\ell}\right)\mathbf{B}\boldsymbol{m}^{\mathrm{eq}}(t-k\Delta t, \boldsymbol{x})\right)_i, \tag{7}$$

for all $t \in \mathfrak{Z}$ and $\boldsymbol{x} \in \mathfrak{L}$, where $\mathbf{A}_i := \mathbf{A}_{\{i\}\cup[N+1\dots q]}$ and $\mathbf{A}_i^{\diamond} := \mathbf{A} - \mathbf{A}_i$, with $(\mathsf{c}_{i,k})_{k=0}^{k=q+1-N} \subset \mathsf{D}$ being the coefficients of $\det(X\boldsymbol{I} - \mathbf{A}_i) = X^{N-1}\sum_{k=0}^{k=q+1-N} \mathsf{c}_{i,k}X^k$, the characteristic polynomial of $\mathbf{A}_i$.

This result is the natural generalization of Proposition 2.4 to the case $N > 1$, in the sense that each sub-problem for any $i \in [1 \dots N]$ deals with one conserved moment (the i-th) at each time, only trying to eliminate the non-conserved moments while keeping the conserved ones other than the i-th. This is achieved by using a tailored characteristic polynomial for each conserved moment in the problem.

Observe that each scheme in (7) is *a priori* a $(q-N)$-steps scheme (thus with $q-N+1$ stages), see Figure 2. When some non-conserved moment relaxes to its equilibrium, the schemes in (7) involve less time steps, see [3] for more details. However, this question is marginal in the present work since the results we will demonstrate hold whatever the number of steps in (7) and thus for any value of the relaxation parameters $s_{N+1}, \dots, s_q$ for the non-conserved moments. Further comments on Propositions 2.4 and 2.5 are postponed to the following Section.

2.3. A more compact form of corresponding Finite Difference schemes

Although the asymptotic analysis we shall develop in Section 4 can be carried on the formulations from Propositions 2.4 and 2.5 previously introduced in [3], we propose a different formalism based on shift operators in time. Having utilized both approaches, the advantage of this new standpoint – which shall be adopted in this paper – is to easily deal with the asymptotic analysis of the coefficients of the characteristic polynomial and of the powers of the matrix $\mathbf{A}$ on the right hand side of (6) or (7). In particular, this allows for the straightforward generalization of the procedure above second-order. Furthermore, the links with other asymptotic analysis of lattice Boltzmann schemes from the literature – which we shall develop in Section 6 – become noticeably more transparent. To this end, we introduce the following definition.

Definition 2.6 (Shift operator in time)**.** Let $f : \mathfrak{Z} \to \mathbb{R}$ be any function defined on the time lattice, then the time shift operator z acts as

$$(\mathsf{z}f)(t) = f(t+\Delta t), \qquad \forall t \in \mathfrak{Z}.$$

With this, the scheme (5) can be recast under the fully-operatorial form: for any $t \in \mathfrak{Z}$ and for any $\boldsymbol{x} \in \mathfrak{L}$

$$(\mathsf{z}\boldsymbol{I} - \mathbf{A})\boldsymbol{m}(t, \boldsymbol{x}) = \mathbf{B}\boldsymbol{m}^{\mathrm{eq}}(t, \boldsymbol{x}), \tag{8}$$

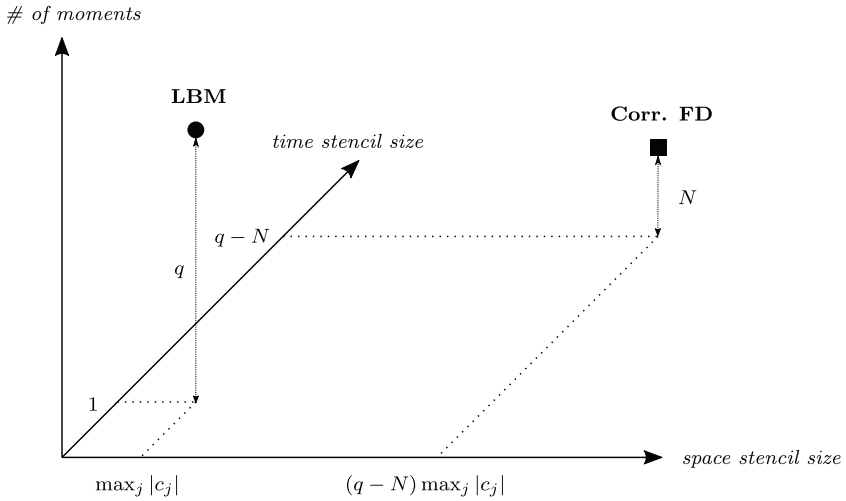

FIGURE 2. Comparison between lattice Boltzmann scheme (circle) and corresponding Finite Difference schemes (square) in terms of involved moments (respectively q and N), number of time steps (respectively 1 and $q-N$) and size of the maximal spatial stencil (respectively $\max_j |c_j|$ and $(q-N)\max_j |c_j|$).

which corresponds to taking the Z-transform [33] of the scheme in the variable z. Here, the inverse of the resolvent associated with $\mathbf{A}$, namely $\mathsf{z}\boldsymbol{I} - \mathbf{A} \in \mathcal{M}_q(\mathbb{R}[\mathsf{z}] \otimes_{\mathbb{R}} \mathsf{D})$, where $\mathbb{R}[\mathsf{z}] \otimes_{\mathbb{R}} \mathsf{D} \cong \mathbb{R}[\mathsf{z}, \mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$, with $\otimes_{\mathbb{R}}$ indicating the tensor product of $\mathbb{R}$-algebras (see [37], Chap. 16 or [34], Chap. 2), forms a commutative ring. In the sequel, we shall drop the time and the space variables when not strictly needed for the sake of readability, because the system given by (8) is intrinsically time and space invariant thanks to Assumption 2.1 and since we work on an unbounded domain, without considering the initial conditions.

Proposition 2.7 (Corresponding Finite Difference scheme for $N = 1$)**.** *Consider $N = 1$. Then the lattice Boltzmann scheme given by (5) or (8) corresponds to a multi-step explicit macroscopic Finite Difference scheme on the conserved moment m_1 under the form*

$$\det(\mathsf{z}\boldsymbol{I} - \mathbf{A}) m_1 = (\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A})\mathbf{B}\boldsymbol{m}^{\mathrm{eq}})_1, \tag{9}$$

where $\operatorname{adj}(\cdot)$ *indicates the adjugate matrix*[3]*, also known as classical adjoint, which is the transpose of the cofactor matrix [27].*

Up to a temporal shift of the whole scheme, the corresponding multi-step explicit Finite Difference scheme by (9) equals the one from (6).

Proof. The proof can be done starting from Proposition 2.4. Alternatively, using the fundamental relation between adjugate and determinant, see Chapter 0 of [27], which is a consequence of the Laplace formula, we have that for any $\boldsymbol{C} \in \mathcal{M}_q(\mathfrak{R})$ where $\mathfrak{R}$ is any commutative ring

$$\boldsymbol{C}\operatorname{adj}(\boldsymbol{C}) = \operatorname{adj}(\boldsymbol{C})\boldsymbol{C} = \det(\boldsymbol{C})\boldsymbol{I}. \tag{10}$$

Hence, multiplying (8) by $\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A})$ yields $\det(\mathsf{z}\boldsymbol{I} - \mathbf{A})\boldsymbol{m} = \operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A})\mathbf{B}\boldsymbol{m}^{\mathrm{eq}}$. Selecting the first row gives (9). □

[3]It is worthwhile observing that the determinant and the adjugate matrix are defined for any square matrix with elements in a commutative ring.

Remark 2.8 (From kinetic to macroscopic)**.** We observe the following facts:

- The procedure can be reversed – when keeping all the lines in $\det(\mathsf{z}\boldsymbol{I} - \mathbf{A})\boldsymbol{m} = \operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A})\mathbf{B}\boldsymbol{m}^{\mathrm{eq}}$ – using a multiplication by $\mathsf{z}\boldsymbol{I} - \mathbf{A}$ and then dividing by the polynomial $\det(\mathsf{z}\boldsymbol{I} - \mathbf{A})$. In this way, one comes back to the lattice Boltzmann scheme by (8). This can be done as long as one does not select and store only the first row as in (9). Contrarily, if this selection is performed, the irreversible passage from the kinetic to the macroscopic formulation is accomplished. The non-conserved moments $m_2, \dots, m_q$ are no longer defined and they cannot be recovered from (9). This fact has been observed by Dellacherie [12]: the same macroscopic Finite Difference scheme can correspond to distinct lattice Boltzmann schemes which can have different evolution equations for the non-conserved moments $m_2, \dots, m_q$. This is not surprising, since for a given monic polynomial, one can find an infinite number of matrices of which it is the characteristic polynomial.
- Though – as previously emphasized – the non-conserved moments are no longer present in the macroscopic Finite Difference scheme by (9), there is a residual shadow of their presence, namely the multi-step nature of the Finite Difference scheme, see Figure 2. In particular, each non-conserved moment m_i relaxing away from the equilibrium, namely with $s_i \neq 1$, for $i \in [2 \dots q]$, adds a time step to the corresponding Finite Difference scheme solely acting on the conserved moment m_1.

Remark 2.9 (Adjugate and characteristic polynomial)**.** A time shift and a change of variable in (6) allows to express $\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A})$ as a polynomial in z of degree $q-1$ computed from the characteristic polynomial. This relation is indeed classical and reads

$$\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}) = \sum_{k=0}^{q-1} \left(\sum_{\ell=0}^{q-1-k} \mathsf{c}_{k+\ell+1} \mathbf{A}^{\ell} \right) \mathsf{z}^k, \qquad \text{where} \qquad \det(\mathsf{z}\boldsymbol{I} - \mathbf{A}) = \sum_{k=0}^{q} \mathsf{c}_k \mathsf{z}^k.$$

In the same way, we can restate Proposition 2.5 using the new formalism.

Proposition 2.10 (Corresponding Finite Difference schemes for $N \geq 1$)**.** *Consider $N \geq 1$. Then the lattice Boltzmann scheme given by (5) or (8) corresponds to a family of multi-step explicit macroscopic Finite Difference schemes on the conserved moments $m_1, \dots, m_N$. This is, for any $i \in [1 \dots N]$*

$$\det(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i) m_i = \left(\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)\mathbf{A}_i^{\diamond}\boldsymbol{m}\right)_i + \left(\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)\mathbf{B}\boldsymbol{m}^{\mathrm{eq}}\right)_i. \tag{11}$$

Up to a temporal shift of the whole scheme, the corresponding multi-step explicit Finite Difference scheme by (11) equals the one from (7).

We could call the form of Finite Difference scheme from Propositions 2.5 and 2.10 "canonical" since we shall prove in Section 2.4 that it guarantees that the Finite Difference schemes do not depend on the choice of relaxation parameters for the conserved variables, which do not play any role in the original lattice Boltzmann scheme either, as previously discussed.

Remark 2.11 (Lack of scaling assumption)**.** The results in Propositions 2.4, 2.5, 2.7, 2.10 are fully discrete and do not make any assumption on the particular scaling between the time-step Δt and the space-step Δx.

The previous Remark 2.11 signifies that the corresponding Finite Difference schemes can be utilized to assess the consistency of the underlying lattice Boltzmann scheme with respect to the macroscopic equations for any particular scaling between time and space discretizations, as we will showcase in Section 3.

2.4. On the choice of relaxation parameters for the conserved moments

In Section 1, we have observed that the choice of relaxation parameters for the conserved moments, namely $s_1, \dots, s_N$, does not change the lattice Boltzmann scheme (2). However, it could be argued that different choices for $s_1, \dots, s_N$ can affect the formulations of the corresponding Finite Difference schemes resulting from Propositions 2.7 and 2.10. We now show that, as one could hope, this is not the case for the Finite Difference schemes given by Proposition 2.10.

Proposition 2.12. *The multi-step explicit macroscopic Finite Difference schemes given by (11) in Proposition 2.10 do not depend on the choice of $s_1, \dots, s_N$, the relaxation parameters of the conserved moments.*

Proof. Fix the indices of the conserved moment $i \in [1 \dots N]$. Let us decompose $\mathbf{B}$, the part of the lattice Boltzmann scheme dealing with the equilibria, as follows: $\mathbf{B} = \mathbf{b}_i \otimes \boldsymbol{e}_i + \mathbf{B}|_{s_i=0}$ where $\mathbf{b}_i = \mathbf{B}_{\cdot,i}$ is the i-th column of $\mathbf{B}$. On the one hand, the dependency of $\mathbf{B}$ on the choice of s_i is now fully contained in $\mathbf{b}_i$. On the other hand $\mathbf{B}|_{s_i=0}$ does not depend on it. The Finite Difference scheme from Proposition 2.10 can be therefore recast, upon rearranging and using well-known properties of the external product $\otimes$, as

$$(\det(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i) - \boldsymbol{e}_i^\intercal \mathrm{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)\mathbf{b}_i)m_i = (\mathrm{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)\mathbf{A}_i^\diamond \boldsymbol{m})_i + (\mathrm{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)\mathbf{B}|_{s_i=0}\boldsymbol{m}^{\mathrm{eq}})_i. \tag{12}$$

The left hand side does not depend on s_j for $j \in [1 \dots N] \setminus \{i\}$ by construction of $\mathbf{A}_i$ and $\mathbf{b}_i$. On the other hand, the right hand side does not depend on s_j for $j \in [1 \dots N] \setminus \{i\}$, because $(\mathbf{A}_i^\diamond)_{\cdot,j} + (\mathbf{B}|_{s_i=0})_{\cdot,j} = (\mathbf{A}_i^\diamond|_{s_j=0})_{\cdot,j}$, where we have used (2) and (1). We are left to discuss the possible dependency of (12) on s_i. For the left hand side, we need the following result concerning the determinant of matrices under rank-one updates, whose proof is analogous to that in [14].

Lemma 2.13 (Matrix determinant)**.** *Let $\mathfrak{R}$ be a commutative ring, $\boldsymbol{C} \in \mathcal{M}_q(\mathfrak{R})$ and $\boldsymbol{u}, \boldsymbol{v} \in \mathfrak{R}^q$, then* $\det(\boldsymbol{C} + \boldsymbol{u} \otimes \boldsymbol{v}) = \det(\boldsymbol{C}) + \boldsymbol{v}^\intercal \mathrm{adj}(\boldsymbol{C})\boldsymbol{u}$.

By this Lemma, we deduce that (12) now reads

$$\det(\mathsf{z}\boldsymbol{I} - (\mathbf{A}_i + \mathbf{b}_i \otimes \boldsymbol{e}_i))m_i = (\mathrm{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)\mathbf{A}_i^\diamond \boldsymbol{m})_i + (\mathrm{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)\mathbf{B}|_{s_i=0}\boldsymbol{m}^{\mathrm{eq}})_i. \tag{13}$$

Observe that $\mathbf{A}_i + \mathbf{b}_i \otimes \boldsymbol{e}_i = \mathbf{A}_i|_{s_i=0}$, thus the left hand side of (13) does not not depend on s_i. The right hand side of (13) is independent of s_i because $\mathbf{A}_i^\diamond$ does not depend on it and since the i-th row of $\mathrm{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A}_i)$ – the transpose of the cofactor matrix of $\mathsf{z}\boldsymbol{I} - \mathbf{A}_i$ – cannot depend on s_i, because only the i-th column of $\mathsf{z}\boldsymbol{I} - \mathbf{A}_i$ depends on s_i. This concludes the proof. □

We have thus shown that the Finite Difference schemes from Proposition 2.10 are the same regardless of the choice of relaxation parameters for the conserved moments and so that we are allowed to take them equal to zero or any other value of specific convenience without loss of generality. In particular, the choice of taking $s_i = 0$ for $i \in [1 \dots N]$ offers interesting simplifications in the computations to come in Section 4, in a way that shall be clearer by looking at the details. Moreover, this choice has the advantage of showing which moments are conserved at a glance.

3. Main results

Everything is in place to start the standard consistency analysis [1, 48] and computation of the modified equations [8, 52] of Finite Difference schemes. We stress the fact that we aim at studying these features for (9) and (11) without explicitly writing these schemes down. We start from the assumptions allowing us to identify each term once developing in formal power series of Δx, *i.e.* performing Taylor expansions. Observe that for any time-space numerical scheme at hand, the time step Δt and the space step Δx are linked (scaling) when the grids are refined. Therefore, we decide to take Δx as discretization parameter tending to zero. Specific bonds between these two parameters will be given in the following pages.

Assumption 3.1 (General assumptions)**.** *Assume that the change of basis $\boldsymbol{M}$ and the relaxation matrix $\boldsymbol{S}$ are fixed as $\Delta x \to 0$.*

We also introduce the spaces of differential operators which shall be obtained by taking the limit $\Delta x \to 0$ as well as other tightly associated concepts.

Definition 3.2 (Time-space differential operators)**.** We define.

- The commutative ring of time-space differential operators $\mathcal{D} := \mathbb{R}[\partial_t] \otimes_{\mathbb{R}} \mathbb{R}[\partial_{x_1}, \dots, \partial_{x_d}] \cong \mathbb{R}[\partial_t, \partial_{x_1}, \dots, \partial_{x_d}]$.
- We consider the commutative ring of formal power series [41,42] $\mathcal{S} := \mathcal{D}[\![\Delta x]\!]$.
- For any $\delta = \sum_{r=0}^{r=+\infty} \Delta x^r \delta^{(r)} \in \mathcal{S}$, we indicate $\delta = O(\Delta x^{r_\circ})$ for some $r_\circ \in \mathbb{N}$ if $\delta^{(r)} = 0$ for $r \in [0 \dots r_\circ - 1]$ and $\delta^{(r_\circ)} \neq 0$. The integer $r_\circ$ is called "order" of the formal power series δ, see Chapter 1 of [45].
- Finally, let $\mathsf{d} \in \mathbb{R}[\mathsf{z}] \otimes_{\mathbb{R}} \mathsf{D}$ and $\delta \in \mathcal{S}$, then we indicate $\mathsf{d} \asymp \delta$, called "asymptotic equivalence" of d and δ, if for any smooth function of the time and space variables $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$, we have

$$(\mathsf{d}f)(t, \boldsymbol{x}) = \sum_{r=0}^{+\infty} \Delta x^r (\delta^{(r)} f)(t, \boldsymbol{x}), \qquad \forall (t, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^d, \qquad \text{as} \qquad \Delta x \to 0.$$

The previous $O(\cdot)$ notation and the notion of asymptotic equivalence are effortlessly extended to vectors and matrices in an entry-wise fashion. It shall be common and harmless not to distinguish between $\mathcal{M}_q(\mathcal{S})$ and $(\mathcal{M}_q(\mathcal{D}))[\![\Delta x]\!]$.

The momentum-velocity operator matrix $\boldsymbol{\mathcal{G}} \in \mathcal{M}_q(\mathcal{D})$, introduced by Dubois [17] with slightly different notations, is defined as follows. It is indeed closely linked to the moment-stream matrix $\mathsf{T} \in \mathcal{M}_q(\mathsf{D})$ that we have previously introduced.

Definition 3.3 (Momentum-velocity operator matrix)**.** The momentum-velocity operator matrix made up of first-order differential operators in space is given by

$$\boldsymbol{\mathcal{G}} := \boldsymbol{M} \left(\sum_{|\boldsymbol{\nu}|=1} \operatorname{diag}\left(\boldsymbol{c}_{\boldsymbol{1}}^{\boldsymbol{\nu}}, \dots, \boldsymbol{c}_{\boldsymbol{q}}^{\boldsymbol{\nu}}\right) \partial^{\boldsymbol{\nu}} \right) \boldsymbol{M}^{-1} \in \mathcal{M}_q(\mathcal{D}),$$

where the multi-index notation is employed.

This momentum-velocity operator matrix can be partitioned in four blocks with different meanings according to the different nature (conserved or not) of the corresponding moments, as for Equation (8) in [17].

As previously announced, one needs to specify the used scaling between Δt and Δx in order to perform the consistency analysis and also to recover the modified equations. We start by the acoustic scaling, see for example [15,17,53], where $\Delta t \sim \Delta x$.

Assumption 3.4 (Acoustic scaling)**.** *The assumptions when considering schemes with the acoustic scaling are:*

(1) $\lambda > 0$ *is a fixed real number as* $\Delta x \to 0$.
(2) *The moments at equilibrium* $\boldsymbol{m}^{\mathrm{eq}}$ *are fixed as* $\Delta x \to 0$.

For the diffusive scaling, see [54,55], where $\Delta t \sim \Delta x^2$, we have:

Assumption 3.5 (Diffusive scaling)**.** *The assumptions when considering schemes with the diffusive scaling are:*

(1) $\lambda = \mu / \Delta x$ *where* $\mu > 0$ *is a fixed real number as* $\Delta x \to 0$.
(2) $\mathcal{G}_{ij} = 0$ *for* $i, j \in [1 \dots N]$.
(3) $m_i^{\mathrm{eq}} = \Delta x \hat{m}_i^{\mathrm{eq}}$ *where* $\hat{m}_i^{\mathrm{eq}}$ *are fixed, for* $i \in \Omega := \{ j \in [1 \dots q] \ : \ \boldsymbol{\mathcal{G}}_{\ell j} \neq 0 \text{ for some } \ell \in [1 \dots N] \}$, *as* $\Delta x \to 0$.
(4) m_i^{eq} *for* $i \notin \Omega$ *are fixed as* $\Delta x \to 0$.

Remark 3.6. These assumptions are needed to state the general results to come. However, there are examples in the literature [4] where they are violated, in particular because the relaxation parameters depend on Δx. This does not prevent from writing the corresponding Finite Difference schemes (9) or (11) for the lattice Boltzmann scheme at hand and then recover their modified equations, but introduces a difficulty to directly obtain the modified equations without explicitly write (9) or (11) down.

We are now ready to state and then prove the main results of the present contribution. The Taylor expansions are applied to the solution of the corresponding Finite Difference schemes given by Propositions 2.5 or 2.10, where non-conserved moments have been removed yielding purely macroscopic discrete equations.

Theorem 3.7 (Acoustic scaling)**.** *Under Assumptions 3.1, 3.4 and in the limit $\Delta x \to 0$, the corresponding macroscopic Finite Difference schemes given by Propositions 2.5 or 2.10 are consistent with the target PDEs*

$$\partial_t \tilde{m}_i + \lambda \sum_{j=1}^{N} \mathcal{G}_{ij} \tilde{m}_j + \lambda \sum_{j=N+1}^{q} \mathcal{G}_{ij} m_j^{\mathrm{eq}}(\tilde{m}_1, \dots, \tilde{m}_N) = 0, \tag{14}$$

for $i \in [1 \dots N]$. For smooth solutions $\tilde{m}_1, \dots, \tilde{m}_N : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$ of (14), the truncation error is given by

$$\tau_i = \lambda \Delta x \sum_{j=N+1}^{q} \left(\frac{1}{s_j} - \frac{1}{2} \right) \mathcal{G}_{ij} \left(\sum_{\ell=1}^{N} \mathcal{G}_{j\ell} \tilde{m}_\ell + \sum_{\ell=N+1}^{q} \mathcal{G}_{j\ell} m_\ell^{\mathrm{eq}}(\tilde{m}_1, \dots, \tilde{m}_N) \right.$$
$$\left. - \frac{1}{\lambda} \sum_{\ell=1}^{N} \frac{\mathrm{d} m_j^{\mathrm{eq}}}{\mathrm{d} m_\ell}(\tilde{m}_1, \dots, \tilde{m}_N) \gamma_{1,\ell}(\tilde{m}_1, \dots, \tilde{m}_N) \right) + O(\Delta x^2),$$

where $\gamma_{1,i}(\tilde{m}_1, \dots, \tilde{m}_N) := \lambda \sum_{j=1}^{j=N} \mathcal{G}_{ij} \tilde{m}_j + \lambda \sum_{j=N+1}^{j=q} \mathcal{G}_{ij} m_j^{\mathrm{eq}}(\tilde{m}_1, \dots, \tilde{m}_N)$. Therefore, the modified equations up to second order read

$$\partial_t m_i + \gamma_{1,i}(m_1, \dots, m_N) - \lambda \Delta x \sum_{j=N+1}^{q} \left(\frac{1}{s_j} - \frac{1}{2} \right) \mathcal{G}_{ij} \left(\sum_{\ell=1}^{N} \mathcal{G}_{j\ell} m_\ell + \sum_{\ell=N+1}^{q} \mathcal{G}_{j\ell} m_\ell^{\mathrm{eq}}(m_1, \dots, m_N) \right.$$
$$\left. - \frac{1}{\lambda} \sum_{\ell=1}^{N} \frac{\mathrm{d} m_j^{\mathrm{eq}}}{\mathrm{d} m_\ell}(m_1, \dots, m_N) \gamma_{1,\ell}(m_1, \dots, m_N) \right) = O(\Delta x^2).$$

The first term in $\gamma_{1,i}$ represents the derivatives of fluxes of the conserved variables, which are necessarily linear, while the second one represents the derivatives of the fluxes given by the equilibria of the non-conserved moments, which can be non-linear. In the numerical diffusion terms, the so-called Hénon's parameters [25] of type $1/s_j - 1/2$ appear. These terms are proportional to Δx. This is not surprising, since the only way of having a stable explicit Finite Difference scheme to simulate the heat equation under the acoustic scaling is to consider a diffusion coefficient proportional to Δx, in order to constrain the speed of propagation of information to remain finite in the limit $\Delta x \to 0$, see for instance Theorem 6.3.1 in [48].

Let us provide two examples for specific lattice Boltzmann schemes taken from the literature and employed with the acoustic scaling.

Example 3.8 ($\mathrm{D_1Q_3}$ with one conserved moment – acoustic scaling)**.** We consider the $\mathrm{D_1Q_3}$ scheme presented in [3, 20], for which $d = 1$, $q = 3$, $c_1 = 0$, $c_2 = 1$, $c_3 = -1$ and $N = 1$. The moment matrix is

$$\boldsymbol{M} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ -2 & 1 & 1 \end{pmatrix}, \qquad \text{hence} \qquad \boldsymbol{\mathcal{G}} = \begin{pmatrix} 0 & \partial_{x_1} & 0 \\ \frac{2}{3}\partial_{x_1} & 0 & \frac{1}{3}\partial_{x_1} \\ 0 & \partial_{x_1} & 0 \end{pmatrix}.$$

Theorem 3.7 immediately gives the modified equation for the acoustic scaling, which reads

$$\partial_t m_1 + \lambda \partial_{x_1} m_2^{\mathrm{eq}}(m_1) - \lambda \Delta x \left(\frac{1}{s_2} - \frac{1}{2} \right) \partial_{x_1} \left(\frac{2}{3} \partial_{x_1} m_1 + \frac{1}{3} \partial_{x_1} m_3^{\mathrm{eq}}(m_1) - \frac{\mathrm{d} m_2^{\mathrm{eq}}(m_1)}{\mathrm{d} m_1} \partial_{x_1} m_2^{\mathrm{eq}}(m_1) \right) = O(\Delta x^2).$$

Example 3.9 ($\mathrm{D_2Q_9}$ with three conserved moments – acoustic scaling)**.** We consider the $\mathrm{D_2Q_9}$ scheme presented in [36], for which $d = 2$, $q = 9$, with

$$\boldsymbol{c}_j = \begin{cases} (0,0)^\intercal, & \text{if} \quad j = 1, \\ (\cos((j-2)\pi/2), \sin((j-2)\pi/2))^\intercal, & \text{if} \quad j \in [2 \dots 5], \\ \sqrt{2}(\cos((2j-3)\pi/4), \sin((2j-3)\pi/4))^\intercal, & \text{if} \quad j \in [6 \dots 9], \end{cases}$$

and $N = 3$. The moment matrix is taken to be (we just permute rows compared to [36] to start with the conserved moments)

$$\boldsymbol{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & -1 & 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 1 & 0 & -1 & 1 & 1 & -1 & -1 \\ -4 & -1 & -1 & -1 & -1 & 2 & 2 & 2 & 2 \\ 4 & -2 & -2 & -2 & -2 & 1 & 1 & 1 & 1 \\ 0 & -2 & 0 & 2 & 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & -2 & 0 & 2 & 1 & 1 & -1 & -1 \\ 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 \end{pmatrix},$$

hence

$$\boldsymbol{\mathcal{G}} = \begin{pmatrix} 0 & \partial_{x_1} & \partial_{x_2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{2}{3}\partial_{x_1} & 0 & 0 & \frac{1}{6}\partial_{x_1} & 0 & 0 & 0 & \frac{1}{2}\partial_{x_1} & \partial_{x_2} \\ \frac{2}{3}\partial_{x_2} & 0 & 0 & \frac{1}{6}\partial_{x_2} & 0 & 0 & 0 & -\frac{1}{2}\partial_{x_2} & \partial_{x_1} \\ 0 & \partial_{x_1} & \partial_{x_2} & 0 & 0 & \partial_{x_1} & \partial_{x_2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \partial_{x_1} & \partial_{x_2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3}\partial_{x_1} & \frac{1}{3}\partial_{x_1} & 0 & 0 & -\partial_{x_1} & \partial_{x_2} \\ 0 & 0 & 0 & \frac{1}{3}\partial_{x_2} & \frac{1}{3}\partial_{x_2} & 0 & 0 & \partial_{x_2} & \partial_{x_1} \\ 0 & \frac{1}{3}\partial_{x_1} & -\frac{1}{3}\partial_{x_2} & 0 & 0 & -\frac{1}{3}\partial_{x_1} & \frac{1}{3}\partial_{x_2} & 0 & 0 \\ 0 & \frac{2}{3}\partial_{x_2} & \frac{2}{3}\partial_{x_1} & 0 & 0 & \frac{1}{3}\partial_{x_2} & \frac{1}{3}\partial_{x_1} & 0 & 0 \end{pmatrix}.$$

The equilibria defining the modified equations under acoustic scaling at second-order are taken as in [17, 36], that is

$$m_4^{\mathrm{eq}} = -2m_1 + 3\big(m_2^2 + m_3^2\big)/m_1, \qquad m_6^{\mathrm{eq}} = -m_2 + 3m_2\big(m_2^2 + m_3^2\big)/m_1^2, \qquad m_7^{\mathrm{eq}} = -m_3 + 3m_3\big(m_2^2 + m_3^2\big)/m_1^2,$$
$$m_8^{\mathrm{eq}} = \big(m_2^2 - m_3^2\big)/m_1, \qquad m_9^{\mathrm{eq}} = m_2 m_3/m_1.$$

It is well-known [17, 21] that the $O(\Delta x)$ terms for the second and third modified equations contain spurious third-order contributions in m_2, m_3. We shall neglect these terms considering that they are small (low-speed flow). Furthermore, we consider that m_1 varies slowly as far as the $O(\Delta x)$ term is concerned, thus we neglect its derivatives. Moreover, we take $s_9 = s_8$, see [17, 36]. Under these assumptions, the modified equations from Theorem 3.7 read

$$\partial_t m_1 + \partial_{x_1}\overline{m}_2 + \partial_{x_2}\overline{m}_3 = O(\Delta x^2),$$

$$\begin{aligned} &\partial_t\overline{m}_2 + \partial_{x_1}\left(\frac{\overline{m}_2^2}{m_1} + \frac{\lambda^2}{3}m_1\right) + \partial_{x_2}\left(\frac{\overline{m}_2\overline{m}_3}{m_1}\right) \\ &-\frac{\lambda}{3}\Delta x\left(\partial_{x_1}\left(2\left(\frac{1}{s_8}-\frac{1}{2}\right)\partial_{x_1}\overline{m}_2 + \left(\frac{1}{s_4}-\frac{1}{s_8}\right)(\partial_{x_1}\overline{m}_2 + \partial_{x_2}\overline{m}_3)\right) + \partial_{x_2}\left(\left(\frac{1}{s_8}-\frac{1}{2}\right)(\partial_{x_2}\overline{m}_2 + \partial_{x_1}\overline{m}_3)\right)\right) = O(\Delta x^2), \end{aligned}$$

$$\begin{aligned} &\partial_t\overline{m}_3 + \partial_{x_1}\left(\frac{\overline{m}_2\overline{m}_3}{m_1}\right) + \partial_{x_2}\left(\frac{\overline{m}_3^2}{m_1} + \frac{\lambda^2}{3}m_1\right) \\ &-\frac{\lambda}{3}\Delta x\left(\partial_{x_1}\left(\left(\frac{1}{s_8}-\frac{1}{2}\right)(\partial_{x_2}\overline{m}_2 + \partial_{x_1}\overline{m}_3)\right) + \partial_{x_2}\left(2\left(\frac{1}{s_8}-\frac{1}{2}\right)\partial_{x_2}\overline{m}_3 + \left(\frac{1}{s_4}-\frac{1}{s_8}\right)(\partial_{x_1}\overline{m}_2 + \partial_{x_2}\overline{m}_3)\right)\right) = O(\Delta x^2), \end{aligned}$$

where we have used $\overline{m}_2 := \lambda m_2$ and $\overline{m}_3 := \lambda m_3$. The first equation enforces the conservation of the density m_1 in the Euler system, discretized with a second-order scheme. The momentum along the first axis (respectively,

second) is $\overline{m}_2$ (respectively, $\overline{m}_3$). The second equation represents – at leading order – the conservation of momentum along the first axis in the Euler system. The pressure law is linear and prescribes that the pressure is equal to $\lambda^2 m_1/3$, hence the speed of the sound is $\lambda/\sqrt{3}$. The numerical diffusion at order $O(\Delta x)$ is what makes up the terms that are usually recognized (except for the previously described pressure) as the stress tensor from the Navier–Stokes system. Recalling that we have assumed slow variations of m_1 (weakly compressible flow), we have a first bulk viscosity (also known as shear or dynamic viscosity) which equals $\mu = \lambda\Delta x(1/s_8 - 1/2)m_1/3$ (not linked μ in Assumption 3.5) and a second bulk viscosity (also known as volume viscosity) given by $\kappa = \lambda\Delta x(3(1/s_4 - 1/2) - (1/s_8 - 1/2))m_1/9$. Hence, for this kind of system, the viscosity is modeled using numerical diffusion and is proportional to Δx, thus vanishing when going towards convergence. The same remarks hold for the last equation.

Theorem 3.10 (Diffusive scaling)**.** *Under Assumptions 3.1, 3.5 and in the limit $\Delta x \to 0$, the corresponding macroscopic Finite Difference schemes given by Propositions 2.5 or 2.10 are consistent with the target PDEs*

$$\partial_t \tilde{m}_i + \mu \sum_{\substack{j=N+1 \\ j\in\Omega}}^{q} \mathcal{G}_{ij}\hat{m}_j^{\mathrm{eq}}(\tilde{m}_1,\dots,\tilde{m}_N) - \mu \sum_{j=N+1}^{q}\left(\frac{1}{s_j}-\frac{1}{2}\right)\mathcal{G}_{ij}\left(\sum_{\ell=1}^{N}\mathcal{G}_{j\ell}\tilde{m}_\ell + \sum_{\substack{\ell=N+1 \\ \ell\notin\Omega}}^{q}\mathcal{G}_{j\ell}m_\ell^{\mathrm{eq}}(\tilde{m}_1,\dots,\tilde{m}_N)\right) = 0, \tag{15}$$

for $i \in [1 \dots N]$. For smooth solutions $\tilde{m}_1,\dots,\tilde{m}_N : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$ of (15), the truncation error is given by $\tau_i = O(\Delta x)$.

We can *a posteriori* explain the meaning of some Assumption 3.5 which were less clear before stating Theorem 3.10. The second assumption avoids to deal with terms which shall naturally appear at order $O(\Delta x)$ but which, since pertaining to the conserved moments, cannot be transformed into terms $O(\Delta x^2)$. Quite the opposite, the third assumption allows to rise to $O(\Delta x^2)$ those terms which contributed to the leading order in Theorem 3.7. This is achieved by a rescaling of the equilibria using $\hat{m}^{\mathrm{eq}}$. We therefore see that lattice Boltzmann schemes can be used to simulate non-linear transport/diffusion equations when using a diffusive scaling.

We also give two examples for specific lattice Boltzmann schemes considered under diffusive scaling.

Example 3.11 ($\mathrm{D_1Q_3}$ with one conserved moment – diffusive scaling)**.** We come back to the setting of Example 3.8 except that we consider a diffusive scaling. Thus we have to take $m_2^{\mathrm{eq}}(m_1) = \Delta x \hat{m}_2^{\mathrm{eq}}(m_1)$ to comply with Assumption 3.5. This yields the modified equation

$$\partial_t m_1 + \mu\partial_{x_1}\hat{m}_2^{\mathrm{eq}}(m_1) - \mu\left(\frac{1}{s_2}-\frac{1}{2}\right)\partial_{x_1}\left(\frac{2}{3}\partial_{x_1}m_1 + \frac{1}{3}\partial_{x_1}m_3^{\mathrm{eq}}(m_1)\right) = O(\Delta x).$$

The scheme allows to simulate non-linear transport phenomena using $\hat{m}_2^{\mathrm{eq}}$ as well as linear and non-linear diffusion using m_3^{eq}.

Example 3.12 ($\mathrm{D_2Q_9}$ with one conserved moment – diffusive scaling)**.** We consider the same scheme as Example 3.9 except for the fact that only one conserved moment $N = 1$ is present and that the equilibria are general, with $m_2^{\mathrm{eq}}(m_1) = \Delta x \hat{m}_2^{\mathrm{eq}}(m_1)$ and $m_3^{\mathrm{eq}}(m_1) = \Delta x \hat{m}_3^{\mathrm{eq}}(m_1)$, to fulfill Assumption 3.5. This is the setting introduced in [54]. The modified equation reads

$$\begin{aligned}\partial_t m_1 + \mu\partial_{x_1}\hat{m}_2^{\mathrm{eq}}(m_1) + \mu\partial_{x_2}\hat{m}_3^{\mathrm{eq}}(m_1) - \frac{2\mu}{3}\left(\frac{1}{s_2}-\frac{1}{2}\right)\partial_{x_1x_1}m_1 - \frac{2\mu}{3}\left(\frac{1}{s_3}-\frac{1}{2}\right)\partial_{x_2x_2}m_1 \\ - \frac{\mu}{6}\left(\left(\frac{1}{s_2}-\frac{1}{2}\right)\partial_{x_1x_1} + \left(\frac{1}{s_3}-\frac{1}{2}\right)\partial_{x_2x_2}\right)m_4^{\mathrm{eq}}(m_1) - \frac{\mu}{2}\left(\left(\frac{1}{s_2}-\frac{1}{2}\right)\partial_{x_1x_1} - \left(\frac{1}{s_3}-\frac{1}{2}\right)\partial_{x_2x_2}\right)m_8^{\mathrm{eq}}(m_1)\end{aligned}$$

$$-\mu\left(\frac{1}{s_2}+\frac{1}{s_3}-1\right)\partial_{x_1x_2}m_9^{\mathrm{eq}}(m_1)=O(\Delta x).$$

Therefore, the scheme allows to simulate non-linear transport phenomena using $\hat{m}_2^{\mathrm{eq}}$ and $\hat{m}_3^{\mathrm{eq}}$ as well as linear and non-linear diffusion with crossed terms *via* m_4^{eq}, m_8^{eq} and m_9^{eq}.

In this contribution, we have deliberately neglected the behavior of the schemes close to the initial time $t=0$. It is dictated by the choice of initial datum for the non-conserved moments, which is not unique for lattice Boltzmann schemes since $q>N$ but one only knows the N conserved moments at $t=0$, being the initial datum of the target PDEs to be solved. The interested reader can consult [44,51] for more information on this topic.

Let us sketch the main ideas of the proofs of Theorem 3.7 and 3.10:

- The result of Proposition 2.5 has allowed to eliminate the non-conserved moments from the discrete scheme, thus has completed the step represented by a vertical arrow in Figure 1. Contrarily to the existing approaches, we do not need (and we cannot, see Rem. 2.8) to estimate the Taylor expansions of the non-conserved moments.
- We benefit from the clever formulation from Proposition 2.10 instead of that of Proposition 2.5. Indeed, considering $\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}_i\asymp \mathsf{z}\boldsymbol{I}-\mathbf{A}_i$, we are allowed to write, for every $i\in[1\,\ldots\,N]$

$$\det(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}_i)m_i=\left(\mathrm{adj}(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}_i)\boldsymbol{\mathcal{A}}_i^{\diamond}\boldsymbol{m}\right)_i+\left(\mathrm{adj}(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}_i)\boldsymbol{\mathcal{B}}\boldsymbol{m}^{\mathrm{eq}}\right)_i,$$

  obtained applying the scheme to smooth functions $m_1,\ldots,m_N:\mathbb{R}^+\times\mathbb{R}\to\mathbb{R}$ and by replacing matrices with entries in the ring $\mathbb{R}[\mathsf{z}]\otimes_{\mathbb{R}}\mathsf{D}$ of discrete operators by their asymptotic equivalents in the ring $\mathcal{S}$. Here, for example, $\det(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}_i)\in\mathcal{S}$, and the expression perfectly makes sense because the determinant and the adjugate are well-defined polynomial functions of any square matrix on a commutative ring, like $\mathcal{S}$. Since the determinant and the adjugate are non-linear functions and thus mix different orders in the expansion $\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}_i$, if we want to recover a closed-form result at a given order of accuracy, we are compelled to utilize the Taylor expansions of the determinant and the adjugate. However, these expansions are well-known and can be computed at any order of accuracy.
  Quite the opposite, if we want to exploit the formulation of [3] stated in Proposition 2.5, we should characterize the asymptotic equivalents of any coefficient of the characteristic polynomial of $\mathbf{A}_i$ and then combine them with the asymptotic equivalents of the time shifts z alone and the terms on the right hand side of (7). Though this is actually feasible and we firstly did it, the computations are extremely involved[4] and very hard to generalize above second-order.
  This justifies the use of the formulation from Proposition 2.10 to achieve the step denoted by an horizontal arrow in Figure 1.

4. Detailed proofs

The vast majority of rest of this work is devoted to the detailed proof of Theorem 3.7 for the scalar case $N=1$. This choice has been adopted to keep the presentation and the involved notations as simple as possible. The idea behind the generalization to $N>1$ is eventually given in Section 5 and is straightforward except for the more involved notations. The proof of Theorem 3.10 follows exactly the same path of Theorem 3.7 and is therefore omitted.

Let us start by finding, for each shift operator from Definition 2.2, its asymptotically equivalent formal power series in Δx, see for instance [17,53]. This is formalized by the following Lemma.

[4]Probably, a deeper mastery of the elementary symmetric polynomials, the Newton's identities, the Bell polynomials and the Feddeev-Leverrier algorithm could simplify many reasonings.

Lemma 4.1 (Series expansion of a shift operator in space)**.** *Let $\boldsymbol{z} \in \mathbb{Z}^d$, then the associated shift operator in space $\mathsf{t}_{\boldsymbol{z}} \in \mathbb{R}[\mathsf{z}] \otimes_{\mathbb{R}} \mathsf{D}$ is asymptotically equivalent, in the limit of $\Delta x \to 0$, to the formal power series of differential operators of the form*

$$\mathsf{t}_{\boldsymbol{z}} \asymp \sum_{|\boldsymbol{\nu}|\geq 0} \frac{(-\Delta x)^{|\boldsymbol{\nu}|}\boldsymbol{z}^{\boldsymbol{\nu}}}{\boldsymbol{\nu}!}\partial^{\boldsymbol{\nu}} \in \mathcal{S}.$$

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function of the spatial variable. Then performing a Taylor expansion for $\Delta x \to 0$ yields

$$(\mathsf{t}_{\boldsymbol{z}} f)(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{z}\Delta x) = \sum_{|\boldsymbol{\nu}|\geq 0} \frac{(-\Delta x)^{|\boldsymbol{\nu}|}\boldsymbol{z}^{\boldsymbol{\nu}}}{\boldsymbol{\nu}!}\partial^{\boldsymbol{\nu}} f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

□

The extension of Lemma 4.1 to any Finite Difference operator in D according to Definition 2.2 is done by linearity. With this in mind, recalling the definition of $\mathbf{T} \in \mathcal{M}_q(\mathsf{D})$, the moments-stream matrix and using Assumption 3.1, we have that

$$\mathbf{T} := \boldsymbol{M}\mathrm{diag}(\mathsf{t}_{c_1}, \dots, \mathsf{t}_{c_q})\boldsymbol{M}^{-1} \asymp \boldsymbol{M}\left(\sum_{|\boldsymbol{\nu}|\geq 0} \frac{(-\Delta x)^{|\boldsymbol{\nu}|}}{\boldsymbol{\nu}!}\mathrm{diag}\big(\boldsymbol{c}_1^{\boldsymbol{\nu}}, \dots, \boldsymbol{c}_q^{\boldsymbol{\nu}}\big)\partial^{\boldsymbol{\nu}}\right)\boldsymbol{M}^{-1} =: \boldsymbol{\mathcal{T}} \in \mathcal{M}_q(\mathcal{S}). \tag{16}$$

Accordingly, we introduce $\boldsymbol{\mathcal{A}} := \boldsymbol{\mathcal{T}}(\boldsymbol{I} - \boldsymbol{S}) \in \mathcal{M}_q(\mathcal{S})$ and $\boldsymbol{\mathcal{B}} := \boldsymbol{\mathcal{T}}\boldsymbol{S} \in \mathcal{M}_q(\mathcal{S})$ such that $\boldsymbol{\mathcal{A}} \asymp \mathbf{A}$ and $\boldsymbol{\mathcal{B}} \asymp \mathbf{B}$. The tight bond between the momentum-velocity operator matrix $\boldsymbol{\mathcal{G}} \in \mathcal{M}_q(\mathcal{D})$ from [17] and our moments-stream matrix $\mathbf{T} \in \mathcal{M}_q(\mathsf{D})$ and its asymptotic equivalent matrix $\boldsymbol{\mathcal{T}} \in \mathcal{M}_q(\mathcal{S})$ is given by the following lemma.

Lemma 4.2 (Link between $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{T}}^{(r)}$)**.** *For any order $r \in \mathbb{N}$, the matrix $\boldsymbol{\mathcal{T}}^{(r)} \in \mathcal{M}_q(\mathcal{D})$ is linked to $\boldsymbol{\mathcal{G}} \in \mathcal{M}_q(\mathcal{D})$ by*

$$\boldsymbol{\mathcal{T}}^{(r)} = \frac{(-1)^r}{r!}\boldsymbol{\mathcal{G}}^r.$$

Moreover, using the Assumption 3.1, we also have

$$\boldsymbol{\mathcal{A}}^{(r)} = \frac{(-1)^r}{r!}\boldsymbol{\mathcal{G}}^r(\boldsymbol{I} - \boldsymbol{S}), \qquad \boldsymbol{\mathcal{B}}^{(r)} = \frac{(-1)^r}{r!}\boldsymbol{\mathcal{G}}^r\boldsymbol{S}.$$

Proof. By (21) in [17], we have that $\mathbf{T} \asymp \boldsymbol{\mathcal{T}} = \exp(-\Delta x\boldsymbol{\mathcal{G}})$. The expansion of the exponential function yields the result. Using Assumption 3.1, one obtains that $\boldsymbol{I} - \boldsymbol{S}$ and $\boldsymbol{S}$ do not perturb the orders of the expansion. □

As far as the time variable is concerned, we can complete by the development of the time shift operator z in order to provide the overall expansion of the inverse of the resolvent $\mathsf{z}\boldsymbol{I} - \mathbf{A} \in \mathcal{M}_q(\mathbb{R}[\mathsf{z}] \otimes_{\mathbb{R}} \mathsf{D})$.

Lemma 4.3 (Expansion of the inverse of the resolvent)**.** *Under Assumptions 3.1, 3.4 and in the limit of $\Delta x \to 0$, the inverse of the resolvent $\mathsf{z}\boldsymbol{I} - \mathbf{A} \in \mathcal{M}_q(\mathbb{R}[\mathsf{z}] \otimes_{\mathbb{R}} \mathsf{D})$ is asymptotically equivalent to $\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}} \in \mathcal{M}_q(\mathcal{S})$, where*

$$\begin{aligned}\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}} &= \sum_{r=0}^{+\infty} \frac{\Delta x^r}{r!}\left(\frac{1}{\lambda^r}\partial_t^r\boldsymbol{I} - (-1)^r\boldsymbol{\mathcal{G}}^r(\boldsymbol{I} - \boldsymbol{S})\right) \\ &= \boldsymbol{S} + \Delta x\left(\frac{1}{\lambda}\partial_t\boldsymbol{I} + \boldsymbol{\mathcal{G}}(\boldsymbol{I} - \boldsymbol{S})\right) + \frac{\Delta x^2}{2}\left(\frac{1}{\lambda^2}\partial_{tt}\boldsymbol{I} - \boldsymbol{\mathcal{G}}^2(\boldsymbol{I} - \boldsymbol{S})\right) + O(\Delta x^3).\end{aligned} \tag{17}$$

Proof. The standard Taylor expansion of z, using the assumption on the acoustic scaling, gives the claim. □

The consistency analysis of the Finite Difference schemes from Proposition 2.10 could be carried on infinite formal power series of differential operators $\mathcal{S}$ on the formulation

$$\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i) m_i = (\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)\boldsymbol{\mathcal{A}}_i^{\diamond} \boldsymbol{m})_i + (\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)\boldsymbol{\mathcal{B}} \boldsymbol{m}^{\mathrm{eq}})_i, \tag{18}$$

for each $i \in [1 \dots N]$, because the determinant and the adjugate perfectly make sense for any square matrix on a commutative ring, like $\mathcal{S}$. However, in order to prove Theorem 3.7, where formal power series are truncated at a certain order, we shall need (17) from Lemma 4.3 as well as the Taylor expansions of the determinant and the adjugate matrix around a given matrix. Indeed, these are non-linear functions and thus mix different orders in the expansions $\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}} \in \mathcal{M}_q(\mathcal{S})$. Since the product of the relaxation parameters for the non-conserved moments is a quantity which shall frequently appear in the computations to come, we fix a special notation for it, namely setting $\Pi := \prod_{i=2}^{i=q} s_i \neq 0$.

4.1. Determinant

We start by studying the expansion of the determinant up to second-order in the perturbation. For this, we need to characterize its derivatives. The expansion can be carried at higher order by employing the very same strategy.

Lemma 4.4 (Derivatives and expansion of the determinant function)**.** *Let $\boldsymbol{C} \in \mathrm{GL}_q(\mathfrak{R})$ and $\boldsymbol{D}, \boldsymbol{E} \in \mathcal{M}_q(\mathfrak{R})$, where $\mathfrak{R}$ is a commutative ring. Then the determinant function*

$$\begin{aligned} \det \colon \mathcal{M}_q(\mathfrak{R}) &\to \mathfrak{R} \\ \boldsymbol{C} &\mapsto \det(\boldsymbol{C}), \end{aligned}$$

has the following derivatives.

$$\mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D}) = \det(\boldsymbol{C})\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D}), \tag{19}$$

$$\mathrm{D}_{\boldsymbol{C}\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{E}) = \det(\boldsymbol{C})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E})\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D}) - \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E}\boldsymbol{C}^{-1}\boldsymbol{D})\big), \tag{20}$$

where $\mathrm{tr}(\cdot)$ *indicates the trace,* i.e. *the sum of the diagonal entries.* (19) *is known as Jacobi formula. Moreover, the second-order Taylor expansion of the determinant function reads*

$$\det(\boldsymbol{C} + \boldsymbol{D}) = \det(\boldsymbol{C}) + \mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D}) + \frac{1}{2}\mathrm{D}_{\boldsymbol{C}\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{D}) + O(\|\boldsymbol{D}\|^3),$$

where the derivatives are given by (19) *and* (20)*.*

Proof. The Jacobi formula (19) is a standard result, see Chapter 0 of [27] or Chapter 5 of [56]. Let us prove (20).

$$\begin{aligned} \mathrm{D}_{\boldsymbol{C}\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{E}) &:= \mathrm{D}_{\boldsymbol{C}}(\mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D}))(\boldsymbol{E}) = \mathrm{D}_{\boldsymbol{C}}\big(\det(\boldsymbol{C})\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{D}\big)\big)(\boldsymbol{E}), \\ &= \mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{E})\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{D}\big) + \det(\boldsymbol{C})\mathrm{D}_{\boldsymbol{C}}\big(\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{D}\big)\big)(\boldsymbol{E}), \\ &= \det(\boldsymbol{C})\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{E}\big)\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{D}\big) + \det(\boldsymbol{C})\mathrm{tr}\big(\mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\boldsymbol{D}\big)(\boldsymbol{E})\big), \\ &= \det(\boldsymbol{C})\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{E}\big)\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{D}\big) - \det(\boldsymbol{C})\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{E}\boldsymbol{C}^{-1}\boldsymbol{D}\big), \end{aligned}$$

where we have used, in this order, the product rule for derivatives, the Jacobi formula (19), the linearity of the trace and the fact that $\mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\big)(\boldsymbol{D}) = -\boldsymbol{C}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1}$, see Chapter 5 of [56]. □

Remark 4.5 (On the invertibility assumption)**.** There exists a form of the Jacobi formula (19) for general $\boldsymbol{C} \in \mathcal{M}_q(\mathfrak{R})$ without assuming invertibility, under the form $\mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D}) = \mathrm{tr}(\mathrm{adj}(\boldsymbol{C})\boldsymbol{D})$. This is equivalent to (19), since (10) holds. Nevertheless, we decided to state Lemma 4.4 using the invertibility assumption. This is done, as we shall see, without loss of generality by taking advantage of some invertible approximation of real matrices and allows to easily find the formulæ for higher order derivatives and expansions *via* basic differential calculus, as illustrated in the previous proof.

In the sequel, we shall take $\mathfrak{R} = \mathcal{S}$ and $\boldsymbol{C} = \boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R}) \subset \mathrm{GL}_q(\mathcal{S})$ and $\boldsymbol{D} = O(\Delta x) \in \mathcal{M}_q(\mathcal{S})$. To simplify the computations and relying on the findings of Section 2.4, we can consider $\boldsymbol{S}$ singular by having $s_1 = 0$. To avoid the difficulties linked with singular matrices, in the spirit of Remark 4.5, we take advantage of the fact that the derivatives of the determinant (and the determinant itself) around $\boldsymbol{C}$ are smooth (indeed, polynomial) functions of $\boldsymbol{C}$. Thus, we introduce the non-singular approximation $\boldsymbol{S}$ where $s_1 \neq 0$, which is such that $\boldsymbol{S} \to \boldsymbol{S}|_{s_1=0}$ as $s_1 \to 0$ for any matricial topology.

We are now ready to use the expansion given by Lemma 4.3 into the terms stemming from Lemma 4.4 to find the leading order terms of the left hand side of (9), namely of $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}) \in \mathcal{S}$. This is nothing but computing the Taylor series of composite functions (see the Faà di Bruno's formulæ [29]) or the composition of formal series

$$\begin{aligned}\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}) = \det(\boldsymbol{S}) + \Delta x \mathrm{D}_{\boldsymbol{S}}(\det(\boldsymbol{S}))((\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(1)}) \\ + \Delta x^2 \big(\mathrm{D}_{\boldsymbol{S}}(\det(\boldsymbol{S}))((\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(2)}) + \frac{1}{2}\mathrm{D}_{\boldsymbol{SS}}(\det(\boldsymbol{S}))((\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(1)})((\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(1)})\big) + O(\Delta x^3).\end{aligned}$$

– One clearly has $\det(\boldsymbol{S}) = s_1 \Pi$, because the matrix $\boldsymbol{S}$ is diagonal. Thus, the Taylor expansion of $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$ does not contain zero-order terms if $s_1 = 0$.

– Let $\boldsymbol{C} = \boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R}) \subset \mathrm{GL}_q(\mathcal{S})$ and $\boldsymbol{D} = \Delta x\big(\frac{1}{\lambda}\partial_t \boldsymbol{I} + \boldsymbol{\mathcal{G}}(\boldsymbol{I} - \boldsymbol{S})\big) + \frac{\Delta x^2}{2}\big(\frac{1}{\lambda^2}\partial_{tt}\boldsymbol{I} - \boldsymbol{\mathcal{G}}^2(\boldsymbol{I} - \boldsymbol{S})\big) + O(\Delta x^3) \in \mathcal{M}_q(\mathcal{S})$ from Lemma 4.3. Using (19) from Lemma 4.4 and performing elementary computations, we have

$$\begin{aligned}\mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D}) = \Delta x \Pi\left(\frac{1}{\lambda}\partial_t + (1 - s_1)\mathcal{G}_{11} + s_1 \sum_{i=2}^{q} \frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t + (1 - s_i)\mathcal{G}_{ii}\right)\right) \\ + \frac{\Delta x^2}{2}\Pi\left(\frac{1}{\lambda^2}\partial_{tt} - (1 - s_1)\mathcal{G}_{11}\mathcal{G}_{11} - (1 - s_1)\sum_{\ell=2}^{q}\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1}\right. \\ \left. + s_1 \sum_{i=2}^{q}\frac{1}{s_i}\left(\frac{1}{\lambda^2}\partial_{tt} - (1 - s_i)\sum_{\ell=1}^{q}\mathcal{G}_{i\ell}\mathcal{G}_{\ell i}\right)\right) + O(\Delta x^3). \qquad (21)\end{aligned}$$

We keep this expression without taking the limit in s_1, for future use. Taking the limit for $s_1 \to 0$ yields the derivative around the singular matrix $\boldsymbol{S}|_{s_1=0}$ instead of $\boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R})$ for $s_1 \neq 0$.

$$\lim_{s_1 \to 0} \mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D}) = \Delta x \Pi\Big(\frac{1}{\lambda}\partial_t + \mathcal{G}_{11}\Big) + \frac{\Delta x^2}{2}\Pi\left(\frac{1}{\lambda^2}\partial_{tt} - \mathcal{G}_{11}\mathcal{G}_{11} - \sum_{\ell=2}^{q}\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1}\right) + O(\Delta x^3). \qquad (22)$$

This gives all the first-order term and part of the second-order term in the series $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$.

– Let $\boldsymbol{C} = \boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R}) \subset \mathrm{GL}_q(\mathcal{S})$ and $\boldsymbol{D} = \Delta x\big(\frac{1}{\lambda}\partial_t \boldsymbol{I} + \boldsymbol{\mathcal{G}}(\boldsymbol{I} - \boldsymbol{S})\big) + O(\Delta x^2) \in \mathcal{M}_q(\mathcal{S})$ from Lemma 4.3. Using (20) from Lemma 4.4, we have, after some algebra

$$\begin{aligned}\mathrm{D}_{\boldsymbol{CC}}(\det(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{D}) = \Delta x^2 \Pi\left(2\left(\frac{1}{\lambda}\partial_t + (1 - s_1)\mathcal{G}_{11}\right)\sum_{i=2}^{q}\frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t + (1 - s_i)\mathcal{G}_{ii}\right)\right. \\ + s_1\left(\sum_{i=2}^{q}\frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t + (1 - s_i)\mathcal{G}_{ii}\right)\right)^2 \\ - 2(1 - s_1)\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell} - 1\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1} - s_1\sum_{i=2}^{q}\frac{1}{s_i^2}\left(\frac{1}{\lambda}\partial_t + (1 - s_i)\mathcal{G}_{ii}\right)^2 \\ \left. - s_1\sum_{i=2}^{q}\sum_{\substack{\ell=2 \\ \ell \neq i}}^{q}\left(\frac{1}{s_i} - 1\right)\left(\frac{1}{s_\ell} - 1\right)\mathcal{G}_{i\ell}\mathcal{G}_{\ell i}\right) + O(\Delta x^3). \qquad (23)\end{aligned}$$

Once more, we take the limit for $s_1 \to 0$ in order to find the desired result on the remaining second-order terms in the development $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$

$$\lim_{s_1\to 0} \mathrm{D}_{\boldsymbol{CC}}(\det(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{D}) = 2\Delta x^2 \Pi \Bigg(\frac{1}{\lambda^2}\partial_{tt} \sum_{\ell=2}^{q} \frac{1}{s_\ell} + \frac{1}{\lambda}\mathcal{G}_{11}\partial_t \sum_{\ell=2}^{q} \frac{1}{s_\ell} + \frac{1}{\lambda}\partial_t \sum_{i=2}^{q} \left(\frac{1}{s_i} - 1\right)\mathcal{G}_{ii} + \mathcal{G}_{11} \sum_{i=2}^{q} \left(\frac{1}{s_i} - 1\right)\mathcal{G}_{ii} - \sum_{\ell=2}^{q} \left(\frac{1}{s_\ell} - 1\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1} \Bigg) + O(\Delta x^3). \tag{24}$$

Putting (22) and (24) together in Lemma 4.4, with expansion around $\boldsymbol{S}$, allows to write $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$ up to third order. This is

$$\lim_{s_1\to 0} \det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}) = \Delta x \Pi \left(\frac{1}{\lambda}\partial_t + \mathcal{G}_{11} \right) + \Delta x^2 \Pi \Bigg(\frac{1}{\lambda^2} \left(\frac{1}{2} + \sum_{\ell=2}^{q} \frac{1}{s_\ell} \right) \partial_{tt} + \frac{1}{\lambda}\mathcal{G}_{11}\partial_t \sum_{\ell=2}^{q} \frac{1}{s_\ell} + \frac{1}{\lambda}\partial_t \sum_{i=2}^{q} \left(\frac{1}{s_i} - 1\right)\mathcal{G}_{ii} - \frac{1}{2}\mathcal{G}_{11}\mathcal{G}_{11} - \sum_{\ell=2}^{q} \left(\frac{1}{s_\ell} - \frac{1}{2}\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1} + \mathcal{G}_{11} \sum_{i=2}^{q} \left(\frac{1}{s_i} - 1\right)\mathcal{G}_{ii} \Bigg) + O(\Delta x^3). \tag{25}$$

4.2. Adjugate

We now switch to the formal power series of the adjugate function of the inverse of the resolvent, in order to deal with the right hand side of the corresponding Finite Difference scheme given by (9). Let us start by characterizing its derivatives.

Lemma 4.6 (Derivatives and expansion of the adjugate function)**.** *Let* $\boldsymbol{C} \in \mathrm{GL}_q(\mathfrak{R})$ *and* $\boldsymbol{D}, \boldsymbol{E} \in \mathcal{M}_q(\mathfrak{R})$*, where* $\mathfrak{R}$ *is a commutative ring. Then the adjugate function*

$$\begin{aligned} \mathrm{adj}\colon \mathcal{M}_q(\mathfrak{R}) &\to \mathcal{M}_q(\mathfrak{R}) \\ \boldsymbol{C} &\mapsto \mathrm{adj}(\boldsymbol{C}), \end{aligned}$$

has the following derivatives.

$$\mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}) = \det(\boldsymbol{C})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1}, \tag{26}$$

$$\begin{aligned} \mathrm{D}_{\boldsymbol{CC}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{E}) = \det(\boldsymbol{C})\Big(&\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E})\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D}) - \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E}\boldsymbol{C}^{-1}\boldsymbol{D})\big)\boldsymbol{C}^{-1} \\ &+ \boldsymbol{C}^{-1}\big(\boldsymbol{E}\boldsymbol{C}^{-1}\boldsymbol{D} + \boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{E} - \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E})\boldsymbol{D} - \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{E}\big)\boldsymbol{C}^{-1}\Big). \end{aligned} \tag{27}$$

Moreover, the second-order Taylor expansion of the adjugate function reads

$$\mathrm{adj}(\boldsymbol{C} + \boldsymbol{D}) = \mathrm{adj}(\boldsymbol{C}) + \mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}) + \frac{1}{2}\mathrm{D}_{\boldsymbol{CC}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{D}) + O(\|\boldsymbol{D}\|^3),$$

where the derivatives are given by (26) *and* (27)*.*

Proof. Since (10) holds and $\boldsymbol{C}$ is invertible, we have that $\mathrm{adj}(\boldsymbol{C}) = \det(\boldsymbol{C})\boldsymbol{C}^{-1}$. Therefore

$$\begin{aligned} \mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}) &= \mathrm{D}_{\boldsymbol{C}}\big(\det(\boldsymbol{C})\boldsymbol{C}^{-1}\big)(\boldsymbol{D}) = \mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D})\boldsymbol{C}^{-1} + \det(\boldsymbol{C})\mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\big)(\boldsymbol{D}), \\ &= \det(\boldsymbol{C})\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{C}^{-1} - \det(\boldsymbol{C})\boldsymbol{C}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1}, \end{aligned}$$

where we have used the rule for the derivative of a product, the Jacobi formula (19) and the identity $\mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\big)(\boldsymbol{D}) = -\boldsymbol{C}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1}$. For the second derivative, we have

$$\begin{aligned} \mathrm{D}_{\boldsymbol{CC}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{E}) &:= \mathrm{D}_{\boldsymbol{C}}(\mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}))(\boldsymbol{E}) = \mathrm{D}_{\boldsymbol{C}}\big(\det(\boldsymbol{C})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1}\big)(\boldsymbol{E}), \\ &= \mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{E})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1} + \det(\boldsymbol{C})\mathrm{D}_{\boldsymbol{C}}\big(\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1}\big)(\boldsymbol{E}), \end{aligned}$$

$$
\begin{aligned}
&= \det(\boldsymbol{C})\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1} + \det(\boldsymbol{C})\mathrm{D}_{\boldsymbol{C}}\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)(\boldsymbol{E})\boldsymbol{C}^{-1}\\
&\quad + \det(\boldsymbol{C})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\big)(\boldsymbol{E}),\\
&= \det(\boldsymbol{C})\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1} + \det(\boldsymbol{C})\big(\mathrm{tr}\big(\mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\big)(\boldsymbol{E})\boldsymbol{D}\big)\boldsymbol{I} - \mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\big)(\boldsymbol{E})\boldsymbol{D}\big)\boldsymbol{C}^{-1}\\
&\quad - \det(\boldsymbol{C})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1}\boldsymbol{E}\boldsymbol{C}^{-1},\\
&= \det(\boldsymbol{C})\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1} - \det(\boldsymbol{C})\big(\mathrm{tr}\big(\boldsymbol{C}^{-1}\boldsymbol{E}\boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{E}\boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1}\\
&\quad - \det(\boldsymbol{C})\big(\mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{D}\big)\boldsymbol{C}^{-1}\boldsymbol{E}\boldsymbol{C}^{-1},
\end{aligned}
$$

where we have used the rule for the derivative of a product, the Jacobi formula (19), the linearity of the derivative and the trace and the identity $\mathrm{D}_{\boldsymbol{C}}\big(\boldsymbol{C}^{-1}\big)(\boldsymbol{D}) = -\boldsymbol{C}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1}$. Upon rearrangement, this yields the result. □

Remark 4.7. We observe that, looking at (26) and (27) compared to (19) and (20), we have that

$$
\begin{aligned}
\mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}) &= \mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D})\boldsymbol{C}^{-1} - \det(\boldsymbol{C})\boldsymbol{C}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1},\\
\mathrm{D}_{\boldsymbol{C}\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{E}) &= \mathrm{D}_{\boldsymbol{C}\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{E})\boldsymbol{C}^{-1} + \det(\boldsymbol{C})\boldsymbol{C}^{-1}(\boldsymbol{E}\boldsymbol{C}^{-1}\boldsymbol{D} + \boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{E}\\
&\quad - \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{E})\boldsymbol{D} - \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{D})\boldsymbol{E})\boldsymbol{C}^{-1}.
\end{aligned}
$$

This implies that we can reuse the computations we did for the determinant in the current treatment of the adjugate, as far as the first terms on the right hand sides are concerned. However, one must be careful that now they are multiplied by $\boldsymbol{C}^{-1}$.

If we had stopped the developments at first order, we could have used the first-order perturbation theory of the adjugate matrix as provided by Theorem 2.1 from [47]. However, to the best of our knowledge, no second-order perturbation theory for this matrix is available in the literature, thus we have been compelled to independently develop it using differential calculus. Lemma 4.6 is thus a generalization of the results from [47] and can therefore be used – beyond the application presented in this contribution – by researchers needing a second-order perturbation theory for the adjugate matrix.

Since we are ultimately interested, as one can notice from (9), in multiplying the formal power series $\mathrm{adj}(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}) \in \mathcal{M}_q(\mathcal{S})$ by $\boldsymbol{\mathcal{B}} \in \mathcal{M}_q(\mathcal{S})$ in a Cauchy-like fashion (the standard product of formal power series) and select the first row, see Proposition 2.7, we perform the computations only for the first row of $\mathrm{adj}(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})$.

– Using the definition of the adjugate matrix in combination with the Laplace formula or using the explicit formula for the adjugate of an upper triangular matrix, see [27], we have

$$
\mathrm{adj}(\boldsymbol{S}) = \Pi\mathrm{diag}\left(1, \frac{s_1}{s_2}, \dots, \frac{s_1}{s_q}\right), \qquad \text{thus} \qquad \lim_{s_1 \to 0} \mathrm{adj}(\boldsymbol{S}) = \Pi\boldsymbol{e}_1 \otimes \boldsymbol{e}_1.
$$

Hence, contrarily to the determinant, the zero-order term in $\mathrm{adj}(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})$ is not zero for $s_1 = 0$ but a singular one-rank diagonal matrix.

– Let $\boldsymbol{C} = \boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R}) \subset \mathrm{GL}_q(\mathcal{S})$ and $\boldsymbol{D} = \Delta x\big(\frac{1}{\lambda}\partial_t\boldsymbol{I} + \boldsymbol{\mathcal{G}}(\boldsymbol{I} - \boldsymbol{S})\big) + \frac{\Delta x^2}{2}\big(\frac{1}{\lambda^2}\partial_{tt}\boldsymbol{I} - \boldsymbol{\mathcal{G}}^2(\boldsymbol{I} - \boldsymbol{S})\big) + O(\Delta x^3) \in \mathcal{M}_q(\mathcal{S})$ from Lemma 4.3. We utilize the previous computations from (21), as suggested in Remark 4.7, into (26).

$$
\begin{aligned}
\mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}) = &\left(\Delta x\Pi\left(\frac{1}{\lambda}\partial_t + (1 - s_1)\mathcal{G}_{11} + s_1\sum_{i=2}^{q}\frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t + (1 - s_i)\mathcal{G}_{ii}\right)\right)\right.\\
&+ \frac{\Delta x^2}{2}\Pi\left(\frac{1}{\lambda^2}\partial_{tt} - (1 - s_1)\mathcal{G}_{11}\mathcal{G}_{11} - (1 - s_1)\sum_{\ell=2}^{q}\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1}\right.\\
&\left.\left.+ s_1\sum_{i=2}^{q}\frac{1}{s_i}\left(\frac{1}{\lambda^2}\partial_{tt} - (1 - s_i)\sum_{\ell=1}^{q}\mathcal{G}_{i\ell}\mathcal{G}_{\ell i}\right)\right)\right)\mathrm{diag}\left(\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_q}\right)\\
&\left. - \mathrm{diag}\left(\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_q}\right)\boldsymbol{D}\mathrm{diag}\left(\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_q}\right)\right) + O(\Delta x^3).
\end{aligned}
$$

In this case, we do not even have to take the limit for $s_1 \to 0$, since all the terms in s_1 cancel. Therefore, for the very first component, we get

$$\begin{aligned}(\mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}))_{11} =& \Delta x \Pi \left(\frac{1}{\lambda} \partial_t \sum_{\ell=2}^{q} \frac{1}{s_\ell} + \sum_{i=2}^{q} \left(\frac{1}{s_i} - 1 \right) \mathcal{G}_{ii} \right) \\ &+ \frac{\Delta x}{2} \Pi \left(\frac{1}{\lambda^2} \partial_{tt} \sum_{\ell=2}^{q} \frac{1}{s_\ell} - \sum_{i=2}^{q} \left(\frac{1}{s_i} - 1 \right) \sum_{\ell=1}^{q} \mathcal{G}_{i\ell} \mathcal{G}_{\ell i} \right) + O(\Delta x^3). \end{aligned} \tag{28}$$

Now consider $j \in [2 \dots q]$, then

$$(\mathrm{D}_{\boldsymbol{C}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D}))_{1j} = -\Delta x \Pi \left(\frac{1}{s_j} - 1 \right) \mathcal{G}_{1j} + \frac{\Delta x^2}{2} \Pi \left(\frac{1}{s_j} - 1 \right) \left(\mathcal{G}_{11} \mathcal{G}_{1j} + \sum_{\ell=2}^{q} \mathcal{G}_{1\ell} \mathcal{G}_{\ell j} \right) + O(\Delta x^3). \tag{29}$$

This gives all the first-order terms on the first row of $\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$ and part of the second-order terms.

– Let $\boldsymbol{C} = \boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R}) \subset \mathrm{GL}_q(\mathcal{S})$ and $\boldsymbol{D} = \Delta x \big(\frac{1}{\lambda} \partial_t \boldsymbol{I} + \boldsymbol{\mathcal{G}}(\boldsymbol{I} - \boldsymbol{S}) \big) + O(\Delta x^2) \in \mathcal{M}_q(\mathcal{S})$ from Lemma 4.3. We reuse computations from (23) as well as (27).

$$\begin{aligned}\mathrm{D}_{\boldsymbol{CC}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{D}) = & \left(\Delta x^2 \Pi \left(2 \Big(\frac{1}{\lambda} \partial_t + (1 - s_1) \mathcal{G}_{11} \Big) \sum_{i=2}^{q} \frac{1}{s_i} \Big(\frac{1}{\lambda} \partial_t + (1 - s_i) \mathcal{G}_{ii} \Big) \right. \right. \\ &+ s_1 \left(\sum_{i=2}^{q} \frac{1}{s_i} \Big(\frac{1}{\lambda} \partial_t + (1 - s_i) \mathcal{G}_{ii} \Big) \right)^2 - 2(1 - s_1) \sum_{\ell=2}^{q} \left(\frac{1}{s_\ell} - 1 \right) \mathcal{G}_{1\ell} \mathcal{G}_{\ell 1} \\ &\left. \left. - s_1 \sum_{i=2}^{q} \frac{1}{s_i^2} \Big(\frac{1}{\lambda} \partial_t + (1 - s_i) \mathcal{G}_{ii} \Big)^2 - s_1 \sum_{i=2}^{q} \sum_{\substack{\ell=2 \\ \ell \neq i}}^{q} \left(\frac{1}{s_i} - 1 \right) \left(\frac{1}{s_\ell} - 1 \right) \mathcal{G}_{i\ell} \mathcal{G}_{\ell i} \right) \right) \\ &\times \mathrm{diag} \left(\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_q} \right) + 2 s_1 \Pi \mathrm{diag} \left(\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_q} \right) \left(\boldsymbol{D} \boldsymbol{S}^{-1} \boldsymbol{D} - \mathrm{tr}(\boldsymbol{S}^{-1} \boldsymbol{D}) \boldsymbol{D} \right) \\ &\times \mathrm{diag} \left(\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_q} \right) + O(\Delta x^3). \end{aligned}$$

Then we have, for the first matrix entry

$$\begin{aligned}(\mathrm{D}_{\boldsymbol{CC}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{D}))_{11} = \Delta x^2 \Pi \left(\left(\sum_{i=2}^{q} \frac{1}{s_i} \Big(\frac{1}{\lambda} \partial_t + (1 - s_i) \mathcal{G}_{ii} \Big) \right)^2 - \sum_{i=2}^{q} \frac{1}{s_i^2} \Big(\frac{1}{\lambda} \partial_t + (1 - s_i) \mathcal{G}_{ii} \Big)^2 \right. \\ \left. - \sum_{i=2}^{q} \left(\frac{1}{s_i} - 1 \right) \sum_{\substack{\ell=2 \\ \ell \neq i}}^{q} \left(\frac{1}{s_\ell} - 1 \right) \mathcal{G}_{i\ell} \mathcal{G}_{\ell i} \right) + O(\Delta x^3), \end{aligned} \tag{30}$$

independent from s_1. For $j \in [2 \dots q]$

$$\begin{aligned}(\mathrm{D}_{\boldsymbol{CC}}(\mathrm{adj}(\boldsymbol{C}))(\boldsymbol{D})(\boldsymbol{D}))_{1j} = 2 \Delta x^2 \Pi \left(\frac{1}{s_j} - 1 \right) \left(\frac{1}{s_j} \mathcal{G}_{1j} \Big(\frac{1}{\lambda} \partial_t + (1 - s_j) \mathcal{G}_{jj} \Big) + \sum_{\substack{\ell=2 \\ \ell \neq j}}^{q} \left(\frac{1}{s_\ell} - 1 \right) \mathcal{G}_{1\ell} \mathcal{G}_{\ell j} \right. \\ \left. - \mathcal{G}_{1j} \sum_{i=2}^{q} \frac{1}{s_i} \Big(\frac{1}{\lambda} \partial_t + (1 - s_i) \mathcal{G}_{ii} \Big) \right) + O(\Delta x^3). \end{aligned} \tag{31}$$

Using (28) and (30), we have that the first entry on the first row of $\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$ is

$$\begin{aligned}\lim_{s_1\to 0}(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}))_{11} = \Pi + \Delta x \Pi\left(\frac{1}{\lambda}\partial_t \sum_{\ell=2}^{q}\frac{1}{s_\ell} + \sum_{i=2}^{q}\left(\frac{1}{s_i}-1\right)\mathcal{G}_{ii}\right) + \frac{\Delta x^2}{2}\Pi\left(\frac{1}{\lambda^2}\partial_{tt}\sum_{\ell=2}^{q}\frac{1}{s_\ell} - \sum_{i=2}^{q}\left(\frac{1}{s_i}-1\right)\right.\\ \times \sum_{\ell=1}^{q}\mathcal{G}_{i\ell}\mathcal{G}_{\ell i} + \left(\sum_{i=2}^{q}\frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t + (1-s_i)\mathcal{G}_{ii}\right)\right)^2 - \sum_{i=2}^{q}\frac{1}{s_i^2}\left(\frac{1}{\lambda}\partial_t + (1-s_i)\mathcal{G}_{ii}\right)^2\\ \left. - \sum_{i=2}^{q}\left(\frac{1}{s_i}-1\right)\sum_{\substack{\ell=2\\ \ell\neq i}}^{q}\left(\frac{1}{s_\ell}-1\right)\mathcal{G}_{i\ell}\mathcal{G}_{\ell i}\right) + O(\Delta x^3). \qquad (32)\end{aligned}$$

Using (29) and (31), for any $j \in [2 \dots q]$, we write

$$\begin{aligned}\lim_{s_1\to 0}(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}))_{1j} = -\Delta x \Pi\left(\frac{1}{s_j}-1\right)\mathcal{G}_{1j} + \frac{\Delta x^2}{2}\Pi\left(\frac{1}{s_j}-1\right)\left(\mathcal{G}_{11}\mathcal{G}_{1j} + \sum_{\ell=2}^{q}\mathcal{G}_{1\ell}\mathcal{G}_{\ell j} + \frac{2}{s_j}\mathcal{G}_{1j}\right.\\ \times \left(\frac{1}{\lambda}\partial_t + (1-s_j)\mathcal{G}_{jj}\right) + 2\sum_{\substack{\ell=2\\ \ell\neq j}}^{q}\left(\frac{1}{s_\ell}-1\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell j} - 2\mathcal{G}_{1j}\sum_{i=2}^{q}\frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t + (1-s_i)\mathcal{G}_{ii}\right)\Bigg)\\ + O(\Delta x^3). \qquad (33)\end{aligned}$$

In general, we have written, for the first row, the leading terms in $\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$. We shall take its product with $\boldsymbol{\mathcal{B}}$. Thus, one has

$$\begin{aligned}\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}} = \mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(0)}\boldsymbol{\mathcal{B}}^{(0)} + \Delta x\Big(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(0)}\boldsymbol{\mathcal{B}}^{(1)} + \mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(1)}\boldsymbol{\mathcal{B}}^{(0)}\Big),\\ + \Delta x^2\Big(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(0)}\boldsymbol{\mathcal{B}}^{(2)} + \mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(1)}\boldsymbol{\mathcal{B}}^{(1)} + \mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(2)}\boldsymbol{\mathcal{B}}^{(0)}\Big) + O(\Delta x^3), \qquad (34)\end{aligned}$$

generating products of terms in the fashion of the Cauchy product. This completes the preliminary results needed to prove Theorem 3.7.

4.3. Overall computation

We now put all the previous calculations together to prove Theorem 3.7. As previously pointed out, we can assume, without loss of generality, that $s_1 = 0$, passing to the limit. This allows to deal with simpler expressions with less terms.

4.3.1. First-order equations

To find the target PDE, it is sufficient to truncate all the formal power series at $O(\Delta x^2)$. In particular, using the fact that the first column of $\boldsymbol{\mathcal{B}}$ is zero for $s_1 = 0$, we have that $\lim_{s_1\to 0}(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}})_{11} = 0$. Observe that if the relaxation parameter corresponding to the conserved moment were not equal to zero, we would have $(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}})_{11} = O(1)$. Still the matrix $\boldsymbol{S}$ would not be singular, thus we would have some non vanishing zero-order term in $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})$ to compensate the one from the adjugate.

For any $j \in [2 \dots q]$, using (33), Lemma 4.2 and (34), entails

$$\begin{aligned}\lim_{s_1\to 0}(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}})_{1j} &= \lim_{s_1\to 0}\Delta x\left(\left(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(0)}\boldsymbol{\mathcal{B}}^{(1)}\right)_{1j} + \left(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(1)}\boldsymbol{\mathcal{B}}^{(0)}\right)_{1j}\right) + O(\Delta x^2)\\ &= -\Delta x \Pi \mathcal{G}_{1j} + O(\Delta x^2).\end{aligned}$$

Equation (25) directly yields

$$\lim_{s_1\to 0}\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}) = \Delta x \Pi\left(\frac{1}{\lambda}\partial_t + \mathcal{G}_{11}\right) + O(\Delta x^2),$$

thus we obtain the modified equation (whatever the choice of $s_1 \in \mathbb{R}$)

$$\Delta x \frac{\Pi}{\lambda}\left(\partial_t m_1 + \lambda \mathcal{G}_{11} m_1 + \lambda \sum_{j=2}^{q} \mathcal{G}_{1j} m_j^{\mathrm{eq}}\right) = O(\Delta x^2),$$

giving the desired result for $N = 1$ upon dividing by the constant Π. We explicitly see the target PDE. Observe that the term Π is never present in the computations by Dubois [17] because they are done on the original lattice Boltzmann scheme (5) or (8) which has only one time step. For instance, in [17], the multi-step nature of the problem, generated by the non-conserved moments relaxing away from the equilibrium, is damped at the very beginning of the procedure by performing the Taylor expansions of the scheme on the non-conserved variables and then plugging them into the expansions for the conserved moments.

Before clarifying the terms at the next order in the modified equation (for any m_1) or equivalently, finding the precise expression of the truncation error (for $m_1 \equiv \tilde{m}_1$ solution of the target PDE), let us utilize the previous equation to get rid of the time derivatives in the second order terms. This can be rigorously done if $m_1 \equiv \tilde{m}_1$, where $\tilde{m}_1$ is the smooth solution of the target PDE and yields the truncation error. For any m_1, this is formal because one assumes that differentiation preserves the asymptotic relations from the symbol $O(\cdot)$. This process constitutes the policy by Dubois [15,17] and is common to all the approaches (Chapman–Enskog, equivalent equation, Maxwell iteration, *etc.*) in order to find the value of the diffusion coefficients from the second-order terms. Moreover, this is classical for Finite Difference schemes, see [8,52]. Notice that in this case, where $N = 1$, $\gamma_{1,\cdot}$ is a scalar, here denoted γ_1 for brevity.

$$\partial_t m_1 = -\lambda \mathcal{G}_{11} m_1 - \lambda \sum_{j=2}^{q} \mathcal{G}_{1j} m_j^{\mathrm{eq}} + O(\Delta x) = -\gamma_1 + O(\Delta x), \tag{35}$$

$$\partial_t \boldsymbol{m}^{\mathrm{eq}} = \frac{\mathrm{d}\boldsymbol{m}^{\mathrm{eq}}}{\mathrm{d}m_1} \partial_t m_1 = -\frac{\mathrm{d}\boldsymbol{m}^{\mathrm{eq}}}{\mathrm{d}m_1}\left(\lambda \mathcal{G}_{11} m_1 + \lambda \sum_{j=2}^{q} \mathcal{G}_{1j} m_j^{\mathrm{eq}}\right) + O(\Delta x) = -\frac{\mathrm{d}\boldsymbol{m}^{\mathrm{eq}}}{\mathrm{d}m_1}\gamma_1 + O(\Delta x), \tag{36}$$

$$\partial_{tt} m_1 = -\partial_t\left(\lambda \mathcal{G}_{11} m_1 + \lambda \sum_{j=2}^{q} \mathcal{G}_{1j} m_j^{\mathrm{eq}}\right) + O(\Delta x) = -\lambda \mathcal{G}_{11} \partial_t m_1 - \lambda \sum_{j=2}^{q} \mathcal{G}_{1j} \partial_t m_j^{\mathrm{eq}} + O(\Delta x), \tag{37}$$

$$= \lambda \mathcal{G}_{11} \gamma_1 + \lambda \sum_{j=2}^{q} \mathcal{G}_{1j} \frac{\mathrm{d}m_j^{\mathrm{eq}}}{\mathrm{d}m_1} \gamma_1 + O(\Delta x) = \lambda^2 \mathcal{G}_{11} \mathcal{G}_{11} m_1 + \lambda^2 \mathcal{G}_{11} \sum_{j=2}^{q} \mathcal{G}_{1j} m_j^{\mathrm{eq}} + \lambda \sum_{j=2}^{q} \mathcal{G}_{1j} \frac{\mathrm{d}m_j^{\mathrm{eq}}}{\mathrm{d}m_1} \gamma_1 + O(\Delta x). \tag{38}$$

These formal equalities are obtained by taking advantage either of the chain rule, since the moments at equilibrium are functions of the conserved moments, or of the re-injection of (35) by assuming that the differentiation preserves the asymptotic relations from the symbol $O(\cdot)$. These equalities become rigorous and lack of the $O(\Delta x)$ term if $m_1 \equiv \tilde{m}_1$, the smooth solution of the target PDE.

4.3.2. Second-order equations

We can now go to the computation of the truncation error in Theorem 3.7, which is more involved due to the presence of more terms to estimate. To make the link with the findings of [17], the increased complexity comes from the more intricate and entangled block structure of $\boldsymbol{\mathcal{G}}^2$. We have to treat the second-order term in (34), made up of three products. For any $j \in [2 \dots q]$ (once again, the first component vanishes for $s_1 = 0$)

– Using Lemma 4.2 and the zero-order expansion of the adjugate gives

$$\lim_{s_1 \to 0}\left(\mathrm{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}})^{(0)} \boldsymbol{\mathcal{B}}^{(2)}\right)_{1j} = \frac{s_j \Pi}{2}\left(\mathcal{G}_{11}\mathcal{G}_{1j} + \sum_{\ell=2}^{q} \mathcal{G}_{1\ell}\mathcal{G}_{\ell j}\right).$$

– Using Lemma 4.2 with (32) and (33)

$$\lim_{s_1\to 0}\Big(\mathrm{adj}(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}})^{(1)}\boldsymbol{\mathcal{B}}^{(1)}\Big)_{1j}=-s_j\Pi\left(\frac{1}{\lambda}\mathcal{G}_{1j}\partial_t\sum_{\ell=2}^{q}\frac{1}{s_\ell}+\boldsymbol{\mathcal{G}}_{1j}\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-1\right)\mathcal{G}_{\ell\ell}-\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-1\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell j}\right).$$

– Using Lemma 4.2 and (33)

$$\begin{aligned}\lim_{s_1\to 0}\Big(\mathrm{adj}(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}})^{(2)}\boldsymbol{\mathcal{B}}^{(0)}\Big)_{1j}=\Pi(1-s_j)\Bigg(&\frac{1}{2}\mathcal{G}_{11}\mathcal{G}_{1j}+\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-\frac{1}{2}\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell j}-\left(\frac{1}{s_j}-1\right)\mathcal{G}_{1j}\mathcal{G}_{jj}\\&+\frac{1}{s_j}\mathcal{G}_{1j}\Big(\frac{1}{\lambda}\partial_t+(1-s_j)\mathcal{G}_{jj}\Big)-\mathcal{G}_{1j}\sum_{i=2}^{q}\frac{1}{s_i}\Big(\frac{1}{\lambda}\partial_t+(1-s_i)\mathcal{G}_{ii}\Big)\Bigg).\end{aligned}$$

Summing these three contributions and after some straightforward but tedious computations, the second-order term in (34) is given by

$$\lim_{s_1\to 0}\Big((\mathrm{adj}(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}})^{(2)}\Big)_{1j}=\Pi\left(\frac{1}{2}\mathcal{G}_{11}\mathcal{G}_{1j}+\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-\frac{1}{2}\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell j}-\frac{1}{\lambda}\left(1+\sum_{\substack{\ell=2\\\ell\neq j}}^{q}\frac{1}{s_\ell}\right)\mathcal{G}_{1j}\partial_t-\mathcal{G}_{1j}\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-1\right)\mathcal{G}_{\ell\ell}\right).$$

Hence, using (36) to get rid of the time derivative of the equilibria, we have

$$\begin{aligned}\lim_{s_1\to 0}\sum_{j=2}^{q}\Big((\mathrm{adj}(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}})^{(2)}\Big)_{1j}m_j^{\mathrm{eq}}=\Pi\sum_{j=2}^{q}\Bigg(&\frac{1}{2}\mathcal{G}_{11}\mathcal{G}_{1j}m_j^{\mathrm{eq}}+\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-\frac{1}{2}\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell j}m_j^{\mathrm{eq}}\\&+\frac{1}{\lambda}\left(1+\sum_{\substack{\ell=2\\\ell\neq j}}^{q}\frac{1}{s_\ell}\right)\mathcal{G}_{1j}\frac{\mathrm{d}m_j^{\mathrm{eq}}}{\mathrm{d}m_1}\gamma_1-\mathcal{G}_{1j}\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-1\right)\mathcal{G}_{\ell\ell}m_j^{\mathrm{eq}}\Bigg)+O(\Delta x).\end{aligned}$$

Notice that in this result, a reminder of order $O(\Delta x)$ appears. Once again, if $m_1\equiv\tilde{m}_1$, this reminder is not present and we would find part of the truncation error. Once more, using (35) and (38) to eliminate the time derivatives in the second-order terms from (25) gives

$$\begin{aligned}\lim_{s_1\to 0}(\det(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}))^{(2)}m_1=\Pi\Bigg(&\frac{1}{\lambda^2}\left(\frac{1}{2}+\sum_{\ell=2}^{q}\frac{1}{s_\ell}\right)\partial_{tt}m_1+\frac{1}{\lambda}\mathcal{G}_{11}\partial_t\sum_{\ell=2}^{q}\frac{1}{s_\ell}+\frac{1}{\lambda}\sum_{i=2}^{q}\left(\frac{1}{s_i}-1\right)\mathcal{G}_{ii}\partial_t m_1\\&-\frac{1}{2}\mathcal{G}_{11}\mathcal{G}_{11}m_1-\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-\frac{1}{2}\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1}m_1+\mathcal{G}_{11}\sum_{i=2}^{q}\left(\frac{1}{s_i}-1\right)\mathcal{G}_{ii}m_1\Bigg)\\=\Pi\Bigg(&\left(\frac{1}{2}+\sum_{\ell=2}^{q}\frac{1}{s_\ell}\right)\left(\mathcal{G}_{11}\mathcal{G}_{11}m_1+\mathcal{G}_{11}\sum_{j=2}^{q}\mathcal{G}_{1j}m_j^{\mathrm{eq}}+\frac{1}{\lambda}\sum_{j=2}^{q}\mathcal{G}_{1j}\frac{\mathrm{d}m_j^{\mathrm{eq}}}{\mathrm{d}m_1}\gamma_1\right)\\&-\mathcal{G}_{11}\left(\mathcal{G}_{11}m_1+\sum_{j=2}^{q}\mathcal{G}_{1j}m_j^{\mathrm{eq}}\right)\sum_{\ell=2}^{q}\frac{1}{s_\ell}-\left(\sum_{i=2}^{q}\left(\frac{1}{s_i}-1\right)\mathcal{G}_{ii}\right)\left(\mathcal{G}_{11}m_1+\sum_{j=2}^{q}\mathcal{G}_{1j}m_j^{\mathrm{eq}}\right)\\&-\frac{1}{2}\mathcal{G}_{11}\mathcal{G}_{11}m_1-\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-\frac{1}{2}\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell 1}m_1+\mathcal{G}_{11}\sum_{i=2}^{q}\left(\frac{1}{s_i}-1\right)\mathcal{G}_{ii}m_1\Bigg)+O(\Delta x).\end{aligned}$$

With this, after simplifications, we obtain the remaining term to master the second-order contributions in the modified equation of the Finite Difference scheme (9).

$$(\det(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}}))^{(2)}m_1-\sum_{j=2}^{q}\Big((\mathrm{adj}(\zeta\boldsymbol{I}-\boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}})^{(2)}\Big)_{1j}m_j^{\mathrm{eq}}$$

$$= -\Pi\left(\sum_{j=2}^{q}\left(\frac{1}{s_j}-\frac{1}{2}\right)\mathcal{G}_{1j}\mathcal{G}_{j1}m_1 + \sum_{j=2}^{q}\sum_{\ell=2}^{q}\left(\frac{1}{s_\ell}-\frac{1}{2}\right)\mathcal{G}_{1\ell}\mathcal{G}_{\ell j}m_j^{\mathrm{eq}} - \frac{1}{\lambda}\sum_{j=2}^{q}\left(\frac{1}{s_j}-\frac{1}{2}\right)\mathcal{G}_{1j}\frac{\mathrm{d}m_j^{\mathrm{eq}}}{\mathrm{d}m_1}\gamma_1\right) + O(\Delta x).$$

To wrap up, these computations yield, together with the ones from Section 4.3.1, the expected result for $N = 1$, which reads

$$\Delta x\frac{\Pi}{\lambda}\left(\partial_t m_1 + \gamma_1 - \lambda\Delta x\sum_{j=2}^{q}\left(\frac{1}{s_j}-\frac{1}{2}\right)\mathcal{G}_{1j}\left(\mathcal{G}_{j1}m_1 + \sum_{\ell=2}^{q}\mathcal{G}_{j\ell}m_\ell^{\mathrm{eq}} - \frac{1}{\lambda}\frac{\mathrm{d}m_j^{\mathrm{eq}}}{\mathrm{d}m_1}\gamma_1\right)\right) = O(\Delta x^3),$$

and thus proves Theorem 3.7.

5. Extension of the proofs to several conserved moments: key ideas

In this Section, we sketch the demonstration of Theorem 3.7 for any $N \geq 1$. For the sake of providing a quick and effective presentation of this matter, we limit ourselves to first-order in Δx. Select a conserved moment, which shall be indexed by $i \in [1 \dots N]$.

Remark 5.1. The operation selecting rows and columns to yield $\mathbf{A}_i$ and $\mathbf{A}_i^\diamond$ from Proposition 2.10 does not change the orders of the expansions. This is, let $\mathbf{C} \in \mathcal{M}_q(\mathbb{R}[\mathsf{z}] \otimes_{\mathbb{R}} \mathsf{D})$ and $\boldsymbol{\mathcal{C}} = \sum_{r=0}^{r=+\infty}\Delta x^r\boldsymbol{\mathcal{C}}^{(r)} \in \mathcal{M}_q(\mathcal{S})$ such that $\mathbf{C} \asymp \boldsymbol{\mathcal{C}}$ and $I \subset [1 \dots q]$ a set of indices, then

$$\mathbf{C}_I \asymp \left(\sum_{r=0}^{+\infty}\Delta x^r\boldsymbol{\mathcal{C}}^{(r)}\right)_I = \sum_{r=0}^{+\infty}\Delta x^r\left(\boldsymbol{\mathcal{C}}^{(r)}\right)_I.$$

Thus we have the analogous of Lemma 4.3, where $\mathsf{z}\boldsymbol{I} - \mathbf{A}_i \asymp \zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}_i$, with

$$\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}_i = \sum_{r=0}^{+\infty}\frac{\Delta x^r}{r!}\left(\frac{1}{\lambda^r}\partial_t^r\boldsymbol{I} - (-1)^r(\boldsymbol{\mathcal{G}}^r(\boldsymbol{I}-\boldsymbol{S}))_{\{i\}\cup[N+1\dots q]}\right).$$

The first two term in the expansion of the inverse of the resolvent are

$$(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)^{(0)} = \mathrm{diag}(1,\dots,1,s_i,1,\dots,1,s_{N+1},\dots,s_q).$$

In the spirit of Remark 4.5 and the computations developed in Section 4, for the case $s_i = 0$, we introduce a regularization with $s_i \neq 0$ and then we pass to the limit. Moreover

$$(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)^{(1)} = \frac{1}{\lambda}\partial_t\boldsymbol{I} + \left(\begin{array}{ccc|c|ccc} 0\cdots0 & 0 & 0\cdots0 & 0 & \cdots & 0 \\ \vdots\ddots\vdots & \vdots & \vdots\ddots\vdots & \vdots & \ddots & \vdots \\ 0\cdots0 & 0 & 0\cdots0 & 0 & \cdots & 0 \\ \hline 0\cdots0 & (1-s_i)\mathcal{G}_{ii} & 0\cdots0 & (1-s_{N+1})\mathcal{G}_{i(N+1)} & \cdots & (1-s_q)\mathcal{G}_{iq} \\ \hline 0\cdots0 & 0 & 0\cdots0 & 0 & \cdots & 0 \\ \vdots\ddots\vdots & \vdots & \vdots\ddots\vdots & \vdots & \ddots & \vdots \\ 0\cdots0 & 0 & 0\cdots0 & 0 & \cdots & 0 \\ \hline 0\cdots0 & (1-s_i)\mathcal{G}_{(N+1)i} & 0\cdots0 & (1-s_{N+1})\mathcal{G}_{(N+1)(N+1)} & \cdots & (1-s_q)\mathcal{G}_{(N+1)q} \\ \vdots\ddots\vdots & \vdots & \vdots\ddots\vdots & \vdots & \ddots & \vdots \\ 0\cdots0 & (1-s_i)\mathcal{G}_{qi} & 0\cdots0 & (1-s_{N+1})\mathcal{G}_{q(N+1)} & \cdots & (1-s_q)\mathcal{G}_{qq} \end{array}\right).$$

We thus have

– As for the case $N = 1$ treated in detail, we have that $\lim_{s_i\to 0}\det((\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)^{(0)}) = 0$. Using the formula for the adjugate of an upper triangular matrix, see [27], we have $\lim_{s_i\to 0}\operatorname{adj}((\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)^{(0)}) = \Pi \boldsymbol{e}_i \otimes \boldsymbol{e}_i$, where in this Section $\Pi := \prod_{\ell=N+1}^{\ell=q} s_\ell$.

– Taking $\boldsymbol{C} = (\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)^{(0)} \in \mathrm{GL}_q(\mathbb{R}) \subset \mathrm{GL}_q(\mathcal{S})$ and $\boldsymbol{D} = \Delta x(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)^{(1)} + O(\Delta x^2) \in \mathcal{M}_q(\mathcal{S})$ in the Jacobi formula (19)

$$\begin{aligned}\lim_{s_i\to 0} \mathrm{D}_{\boldsymbol{C}}(\det(\boldsymbol{C}))(\boldsymbol{D}) &= \lim_{s_i\to 0} \Delta x \Pi\left(\frac{s_i(N-1)}{\lambda}\partial_t + \frac{1}{\lambda}\partial_t + (1-s_i)\mathcal{G}_{ii} + \sum_{\ell=N+1}^{q}\frac{1}{s_\ell}\Big(\frac{1}{\lambda}\partial_t + (1-s_\ell)\mathcal{G}_{\ell\ell}\Big)\right) + O(\Delta x^2)\\ &= \Delta x \Pi\Big(\frac{1}{\lambda}\partial_t + \mathcal{G}_{ii}\Big) + O(\Delta x^2).\end{aligned}$$

To handle the term with the adjugate, observe that the first-order term is made up of the terms

$$(\operatorname{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)\boldsymbol{\mathcal{A}}_i^{\diamond})^{(1)} = (\operatorname{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i))^{(0)}(\boldsymbol{\mathcal{A}}_i^{\diamond})^{(1)} + (\operatorname{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i))^{(1)}(\boldsymbol{\mathcal{A}}_i^{\diamond})^{(0)}, \tag{39}$$

and in particular, we are interested in the i-th line of this matrix. Because of the fact that $(\boldsymbol{\mathcal{A}}_i^{\diamond})^{(0)} = \operatorname{diag}(1-s_1,\dots,1-s_{i-1},0,1-s_{i+1},\dots,1-s_N,0,\dots,0)$, the i-th line of the second term on the right hand side of (39) is zero, thus we do not have do study it. For the remaining term, it can be easily seen that

$$(\boldsymbol{\mathcal{A}}_i^{\diamond})^{(1)} = -(\boldsymbol{I} - \boldsymbol{S})\left(\begin{array}{ccc|c|ccc||ccc}
\mathcal{G}_{11} & \cdots & \mathcal{G}_{1(i-1)} & \mathcal{G}_{1i} & \mathcal{G}_{1(i+1)} & \cdots & \mathcal{G}_{1N} & \mathcal{G}_{1(N+1)} & \cdots & \mathcal{G}_{1q}\\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\
\mathcal{G}_{(i-1)1} & \cdots & \mathcal{G}_{(i-1)(i-1)} & \mathcal{G}_{(i-1)i} & \mathcal{G}_{(i-1)(i+1)} & \cdots & \mathcal{G}_{(i-1)N} & \mathcal{G}_{(i-1)(N+1)} & \cdots & \mathcal{G}_{(i-1)q}\\
\hline
\mathcal{G}_{i1} & \cdots & \mathcal{G}_{i(i-1)} & 0 & \mathcal{G}_{i(i+1)} & \cdots & \mathcal{G}_{iN} & 0 & \cdots & 0\\
\hline
\mathcal{G}_{(i+1)1} & \cdots & \mathcal{G}_{(i+1)(i-1)} & \mathcal{G}_{(i+1)i} & \mathcal{G}_{(i+1)(i+1)} & \cdots & \mathcal{G}_{(i+1)N} & \mathcal{G}_{(i+1)(N+1)} & \cdots & \mathcal{G}_{(i+1)q}\\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\
\mathcal{G}_{N1} & \cdots & \mathcal{G}_{N(i-1)} & \mathcal{G}_{Ni} & \mathcal{G}_{N(i+1)} & \cdots & \mathcal{G}_{NN} & \mathcal{G}_{N(N+1)} & \cdots & \mathcal{G}_{Nq}\\
\hline\hline
\mathcal{G}_{(N+1)1} & \cdots & \mathcal{G}_{(N+1)(i-1)} & 0 & \mathcal{G}_{(N+1)(i+1)} & \cdots & \mathcal{G}_{(N+1)N} & 0 & \cdots & 0\\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\
\mathcal{G}_{q1} & \cdots & \mathcal{G}_{q(i-1)} & 0 & \mathcal{G}_{q(i+1)} & \cdots & \mathcal{G}_{qN} & 0 & \cdots & 0
\end{array}\right),$$

thus we deduce that

$$\begin{aligned}\Big((\operatorname{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)\boldsymbol{\mathcal{A}}_i^{\diamond})^{(1)}\Big)_{i,\cdot} = -\Pi\Big(&(1-s_1)\mathcal{G}_{i1},\dots,(1-s_{i-1})\mathcal{G}_{i(i-1)},0,(1-s_{i+1})\mathcal{G}_{i(i+1)},\\ &\dots,(1-s_N)\mathcal{G}_{iN},0,\dots,0\Big).\end{aligned}$$

Dealing with the zero and first order term in $\operatorname{adj}(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}_i)\boldsymbol{\mathcal{B}}$ works the same than $N = 1$, thus we do not repeat it. Moreover, these terms allow for the compensation of the dependence on the choice of the relaxation parameter of the other conserved moments $s_1,\cdots,s_{i-1},s_{i+1},\dots,s_N$ in the previous equation, as claimed in Section 2.4, thanks to (1).

Putting all the previously discussed facts together into the truncated (18) yields

$$\Delta x\frac{\Pi}{\lambda}\left(\partial_t m_i + \lambda\mathcal{G}_{ii}m_i + \lambda\sum_{\substack{j=1\\ j\neq i}}^{N}\mathcal{G}_{ij}m_j + \lambda\sum_{j=N+1}^{q}\mathcal{G}_{ij}m_j^{\mathrm{eq}}\right) = O(\Delta x^2),$$

which is the result from Theorem 3.7 for $N \geq 1$ at dominant order. The next order is demonstrated in the same way.

6. Link with the existing approaches

To finish our contribution, we briefly sketch the links with previous works on the target PDEs and modified equations like [15,17,53]. A more complete study shall be the object of future investigations.

6.1. Equivalent equations

Our result Theorem 3.7 coincides with the analogous result in [17] up to second order. The substantial difference is that we apply the Taylor expansions to the solution of the corresponding Finite Difference scheme given either by Proposition 2.7 or Proposition 2.10, where non-conserved moments have been removed. We therefore reasonably conjecture that the obtained macroscopic equations coincide at any order. The mathematical justification of this conjecture shall be the object of future investigations.

The quasi-equilibrium, which is extensively used in [17] can be somehow recovered in our previous discussion. Let $N = 1$ to fix ideas. In the proof of Proposition 2.7, nothing prevents us from selecting, instead of the first row, the $i \in [2 \dots q]$ row, corresponding to a non-conserved moment. This is

$$\det(\mathsf{z}\boldsymbol{I} - \mathbf{A})m_i = (\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A})\mathbf{B}\boldsymbol{m}^{\mathrm{eq}})_i. \tag{40}$$

Let us stress that even if this could seem to be a viable Finite Difference scheme for the non-conserved variable m_i, it is not independent from the conserved moment m_1 the equilibria depend on and furthermore, this formulation certainly depends on the choice of s_1, the relaxation parameter of the conserved moment. This is somehow unwanted since s_1 is *in fine* not present in the original lattice Boltzmann scheme. From the computations of Section 4, we see that

$$\det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}) = s_1\Pi + O(\Delta x), \qquad \operatorname{adj}(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}) = \Pi\operatorname{diag}\left(1, \frac{s_1}{s_2}, \dots, \frac{s_1}{s_q}\right) + O(\Delta x), \qquad \boldsymbol{\mathcal{B}} = \boldsymbol{S} + O(\Delta x).$$

Using the asymptotic equivalents truncated at leading order in (40) thus provides

$$s_1\Pi m_i + O(\Delta x) = s_1\Pi m_i^{\mathrm{eq}} + O(\Delta x), \qquad \text{hence also} \qquad m_i = m_i^{\mathrm{eq}} + O(\Delta x),$$

provided that $s_1 \neq 0$. This is the quasi-equilibrium of the non-conserved moments, which is re-injected in the lattice Boltzmann schemes to eliminate them in the procedure by Dubois [17]. The previous procedure is formal because there is no guarantee that the discrete non-conserved moments m_i for $i \in [2 \dots q]$ in the scheme originate from the point-wise values of a smooth function.

6.2. Maxwell iteration

In [53], the computations have been carried only for the $\mathrm{D_2Q_9}$ scheme by Lallemand and Luo [36] with $N = 3$, which we have presented in Example 3.9. In this part of our work, we are first going to develop the computations until third-order for any lattice Boltzmann scheme under acoustic scaling, *i.e.* Assumption 3.4. Then, we are going to demonstrate that the modified equations obtained by the Maxwell iteration [53] and the one from the corresponding Finite Difference schemes are the same at any order, regardless of the time-space scaling. In this Section, it is crucial to assume that $\boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R})$. Observe that this assumption ensures that $\det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})$ is a unit (invertible) in the ring $\mathcal{S}$ or equivalently that $\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}}$ belongs to $\mathrm{GL}_q(\mathcal{S})$. The Maxwell iteration [53] at step $k \in \mathbb{N}$ reads, after simple computations

$$\boldsymbol{m}^{[k]} = \left(\sum_{r=0}^{k}(-\boldsymbol{S}^{-1}(\zeta\overline{\boldsymbol{\mathcal{T}}} - \boldsymbol{I}))^r\right)\boldsymbol{m}^{\mathrm{eq}}, \tag{41}$$

where the quasi-equilibrium is encoded in the choice $\boldsymbol{m}^{[0]} = \boldsymbol{m}^{\mathrm{eq}}$ and where we have taken, as for (16)

$$\overline{\mathbf{T}} := \boldsymbol{M}\operatorname{diag}(\mathsf{t}_{-c_1}, \dots, \mathsf{t}_{-c_q})\boldsymbol{M}^{-1} \asymp \boldsymbol{M}\left(\sum_{|\boldsymbol{\nu}|\geq 0}\frac{\Delta x^{|\boldsymbol{\nu}|}}{\boldsymbol{\nu}!}\operatorname{diag}\big(\boldsymbol{c}_{\boldsymbol{1}}^{\boldsymbol{\nu}}, \dots, \boldsymbol{c}_{\boldsymbol{q}}^{\boldsymbol{\nu}}\big)\partial^{\boldsymbol{\nu}}\right)\boldsymbol{M}^{-1} =: \overline{\boldsymbol{\mathcal{T}}} \in \mathcal{M}_q(\mathcal{S}).$$

It is easy to see that $\boldsymbol{\mathcal{T}}\overline{\boldsymbol{\mathcal{T}}} = \overline{\boldsymbol{\mathcal{T}}}\boldsymbol{\mathcal{T}} = \boldsymbol{I}$ and moreover, in analogy with Lemma 4.3

$$\zeta\overline{\boldsymbol{\mathcal{T}}} - \boldsymbol{I} = \Delta x\Big(\frac{1}{\lambda}\partial_t\boldsymbol{I} + \boldsymbol{\mathcal{G}}\Big) + \frac{\Delta x^2}{2}\Big(\frac{1}{\lambda^2}\partial_{tt}\boldsymbol{I} + \frac{2}{\lambda}\boldsymbol{\mathcal{G}}\partial_t + \boldsymbol{\mathcal{G}}^2\Big) + O(\Delta x^3). \tag{42}$$

The Maxwell iteration works by assuming that $\boldsymbol{m} = \boldsymbol{m}^{[k]} + O(\Delta x^{k+1})$. Taking $k = 1$ in (41) and using (42), we have

$$\boldsymbol{m} = \boldsymbol{m}^{\mathrm{eq}} - \boldsymbol{S}^{-1}\Delta x\Big(\frac{1}{\lambda}\partial_t\boldsymbol{I} + \boldsymbol{\mathcal{G}}\Big)\boldsymbol{m}^{\mathrm{eq}} + O(\Delta x^2).$$

Let $i \in [1 \dots N]$, then taking advantage of (1)

$$m_i = m_i - \Delta x\frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t m_i + \sum_{j=1}^{N}\mathcal{G}_{ij}m_j + \sum_{j=N+1}^{q}\mathcal{G}_{1j}m_j^{\mathrm{eq}}\right) + O(\Delta x^2),$$

which upon division, is the same result than Theorem 3.7. Going up to order two considering $k = 2$, we have

$$\begin{aligned}\boldsymbol{m} = \boldsymbol{m}^{\mathrm{eq}} &- \boldsymbol{S}^{-1}\Delta x\Big(\frac{1}{\lambda}\partial_t\boldsymbol{I} + \boldsymbol{\mathcal{G}}\Big)\boldsymbol{m}^{\mathrm{eq}} \\ &+ \frac{\Delta x^2}{2}\boldsymbol{S}^{-1}\left(\frac{1}{\lambda^2}(2\boldsymbol{S}^{-1} - \boldsymbol{I})\partial_{tt} + \frac{2}{\lambda}(\boldsymbol{S}^{-1}\boldsymbol{\mathcal{G}} + \boldsymbol{\mathcal{G}}\boldsymbol{S}^{-1} - \boldsymbol{\mathcal{G}})\partial_t + \boldsymbol{\mathcal{G}}(2\boldsymbol{S}^{-1} - \boldsymbol{I})\boldsymbol{\mathcal{G}}\right)\boldsymbol{m}^{\mathrm{eq}} + O(\Delta x^3).\end{aligned}$$

Once more, selecting the i-th row provides

$$\begin{aligned}m_i = m_i - \Delta x\frac{1}{s_i}\left(\frac{1}{\lambda}\partial_t m_i + \frac{1}{\lambda}\gamma_{1,i}\right) + \Delta x^2\frac{1}{s_i}\Bigg(\frac{1}{\lambda^2}\left(\frac{1}{s_i} - \frac{1}{2}\right)\partial_{tt}m_i + \frac{1}{\lambda}\sum_{j=1}^{q}\left(\frac{1}{s_i} + \frac{1}{s_j} - 1\right)\mathcal{G}_{ij}\partial_t m_j^{\mathrm{eq}} \\ + \sum_{j=1}^{q}\sum_{\ell=1}^{q}\left(\frac{1}{s_\ell} - \frac{1}{2}\right)\mathcal{G}_{i\ell}\mathcal{G}_{\ell j}m_j^{\mathrm{eq}}\Bigg)\Bigg) + O(\Delta x^3).\end{aligned}$$

Using relations analogous to (35), (36) and (38) for $N \geq 1$, formally obtained by differentiating the result at the previous order , we finally obtain, after tedious but elementary computations

$$\begin{aligned}m_i = m_i - \frac{\Delta x}{\lambda s_i}\Bigg(\partial_t m_i + \gamma_{1,i} - \lambda\Delta x\sum_{j=N+1}^{q}\left(\frac{1}{s_j} - \frac{1}{2}\right)\mathcal{G}_{ij}\left(\sum_{\ell=1}^{N}\mathcal{G}_{j\ell}m_\ell + \sum_{\ell=N+1}^{q}\mathcal{G}_{j\ell}m_\ell^{\mathrm{eq}} - \frac{1}{\lambda}\sum_{\ell=1}^{N}\frac{\mathrm{d}m_j^{\mathrm{eq}}}{\mathrm{d}m_\ell}\gamma_{1,\ell}\right)\Bigg) \\ + O(\Delta x^3),\end{aligned}$$

which coincides with the result from Theorem 3.7. Therefore, up to order two, our approach yields results consistent with those from the procedure by Yong *et al.* [53].

To demonstrate that we recover the same result at any order for any scaling between time and space discretizations, let us assume $N = 1$. Then we have, using that $\boldsymbol{S} \in \mathrm{GL}_q(\mathbb{R})$, $\boldsymbol{\mathcal{T}}\overline{\boldsymbol{\mathcal{T}}} = \overline{\boldsymbol{\mathcal{T}}}\boldsymbol{\mathcal{T}} = \boldsymbol{I}$, the rule for the inverse of a product of matrix and the identity relative to geometric series in the context of formal power series, that

$$\begin{aligned}\boldsymbol{0} &= \det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})\boldsymbol{m} - \mathrm{adj}(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})\boldsymbol{\mathcal{B}}\boldsymbol{m}^{\mathrm{eq}} = \det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})\big(\boldsymbol{m} - (\zeta\boldsymbol{I} - \boldsymbol{\mathcal{T}}(\boldsymbol{I} - \boldsymbol{S}))^{-1}\boldsymbol{\mathcal{T}}\boldsymbol{S}\boldsymbol{m}^{\mathrm{eq}}\big) \\ &= \det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})\big(\boldsymbol{m} - (\boldsymbol{S}^{-1}\overline{\boldsymbol{\mathcal{T}}}(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{T}}(\boldsymbol{I} - \boldsymbol{S})))^{-1}\boldsymbol{m}^{\mathrm{eq}}\big) = \det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})\big(\boldsymbol{m} - (\boldsymbol{I} + \boldsymbol{S}^{-1}(\zeta\overline{\boldsymbol{\mathcal{T}}} - \boldsymbol{I}))^{-1}\boldsymbol{m}^{\mathrm{eq}}\big) \\ &= \det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})\left(\boldsymbol{m} - \left(\sum_{r=0}^{+\infty}(-\boldsymbol{S}^{-1}(\zeta\overline{\boldsymbol{\mathcal{T}}} - \boldsymbol{I}))^r\right)\boldsymbol{m}^{\mathrm{eq}}\right) = \det(\zeta\boldsymbol{I} - \boldsymbol{\mathcal{A}})\left(\boldsymbol{m} - \lim_{k\to+\infty}\boldsymbol{m}^{[k]}\right).\end{aligned}$$

Therefore the expansion of the Finite Difference scheme from Proposition 2.7 and the non-truncated Maxwell iteration method on the lattice Boltzmann scheme coincide up to a multiplication by a formal power series of time-space differential operators, *i.e.* $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}) \in \mathcal{S}$. *A priori*, the resulting modified equations are not the same, but since $\det(\zeta \boldsymbol{I} - \boldsymbol{\mathcal{A}}) = \det(\boldsymbol{S}) + O(\Delta x) = s_1 \Pi + O(\Delta x)$, thus we "pay" only a constant factor we can divide by at dominant order, the modified equations at leading order are the same. Then, at each order, the result must be the same because we re-inject, in a recursive fashion, the solution truncated at the previous order to eliminate the higher-order time derivatives, see for instance (36) and (38). The fact that the modified equations recovered by the Maxwell iteration are the same than the ones from the corresponding Finite Difference scheme at any order provides an *a posteriori* justification of the Maxwell iteration. We also emphasize that using the Maxwell iteration to compute these equations is generally less involved in terms of computations than doing the same on the corresponding Finite Difference schemes.

7. Conclusions and perspectives

In this paper, we have rigorously derived the target PDEs for any lattice Boltzmann scheme under acoustic and diffusive scalings by restating it as a multi-step macroscopic Finite Difference scheme on the conserved moments [3]. Moreover, the modified equations – which the schemes are "more consistent" with – have been found up to second order. These findings allow to utilize – upon studying the stability [3] of the lattice Boltzmann scheme at hand – the Lax equivalence theorem [38] to conclude on its convergence and order of convergence towards the solution of the target PDEs. Since the passage from the kinetic to the macroscopic standpoint is fully discrete, our analysis can handle any type of time-space scaling and be pushed forward to reach higher orders in the discretization parameters. Contrarily to the existing techniques, the quasi-equilibrium of the non-conserved moments in the limit of small discretization parameters or the introduction of several time scales in the problem are not the keys to eliminate the non-conserved variables from the macroscopic equations. The obtained results confirm, going beyond empirical evidence, that the formal Taylor expansion by Dubois [15,17] and the Maxwell iteration by Yong *et al.* [53] are well-grounded from the perspective of numerical analysts and traditional numerical methods for PDEs, such as Finite Difference. In particular, we have extended the Maxwell iteration [53] to any lattice Boltzmann scheme and shown that the modified equations found by this procedure are the same than the ones from the corresponding Finite Difference schemes, at any order. The general results that we have presented allow to immediately recover the modified equations without need for computing the corresponding Finite Difference schemes, which would be time consuming. This allows – for example – to easily consider families of schemes depending on some parameters and investigate the dependence of the modified equations on these factors.

An improvement of the present work could be the establishment of the equivalence between different analyses [10,15,17,30–32,36,43] for higher orders and ideally for any order. Even if more involved from the standpoint of computations, the extension can be easily done by considering derivatives of higher order for the determinant and adjugate functions, in the spirit of Lemmas 4.4 and 4.6. In this work, all the computations have been done by hand but one could envision to seek some help from symbolic computations. This is a current path of investigation which final aim is to provide the computation – inside the package `pylbm`[5] – of the modified equations of any lattice Boltzmann scheme either by the corresponding Finite Difference scheme or using the Maxwell iteration.

Acknowledgements. The author deeply thanks his PhD advisors, M. Massot and B. Graille, for the fruitful discussions and advice on the subject, his brother P. Bellotti for the useful tips to improve the style of manuscript and S. Simonis for having read and commented the preprint of this paper. The author also thanks the two anonymous referees for the valuable questions and suggestions. The author is supported by a PhD funding (year 2019) from the Ecole polytechnique.

[5] https://pylbm.readthedocs.io.

References

[1] G. Allaire, Numerical Analysis and Optimization: An Introduction to Mathematical Modelling and Numerical Simulation. Oxford University Press (2007).

[2] T. Bellotti, L. Gouarin, B. Graille and M. Massot, High accuracy analysis of adaptive multiresolution-based lattice Boltzmann schemes via the equivalent equations. *SMAI J. Comput. Math.* **8** (2022) 161–199.

[3] T. Bellotti, B. Graille and M. Massot, Finite difference formulation of any lattice Boltzmann scheme. *Numer. Math.* **152** (2022) 1–40.

[4] B. Boghosian, F. Dubois, B. Graille, P. Lallemand and M.-M. Tekitek, Curious convergence properties of lattice boltzmann schemes for diffusion with acoustic scaling. *Commun. Comput. Phys.* **23** (2018) 1263–1278.

[5] F. Bouchut, F.R. Guarguaglini and R. Natalini, Diffusive BGK approximations for nonlinear multidimensional parabolic equations. *Indiana Univ. Math. J.* **49** (2000) 723–749.

[6] J.W. Brewer, J.W. Bunce and F.S. Van Vleck, Linear Systems Over Commutative Rings. CRC Press (1986).

[7] A. Caiazzo, M. Junk and M. Rheinländer, Comparison of analysis techniques for the lattice Boltzmann method. *Comput. Math. Appl.* **58** (2009) 883–897.

[8] R. Carpentier, A. de La Bourdonnaye and B. Larrouturou, On the derivation of the modified equation for the analysis of linear numerical methods. *ESAIM: Math. Modell. Numer. Anal.* **31** (1997) 459–470.

[9] S. Chapman and T.G. Cowling, The Mathematical Theory of Non-Uniform Gases: An Account of the Kinetic Theory of Viscosity, Thermal Conduction and Diffusion in Gases. Cambridge university press (1990).

[10] S. Chen and G.D. Doolen, Lattice Boltzmann method for fluid flows. *Ann. Rev. Fluid Mech.* **30** (1998) 329–364.

[11] S.S. Cheng, Partial Difference Equations. Vol. 3. CRC Press (2003).

[12] S. Dellacherie, Construction and analysis of lattice Boltzmann methods applied to a 1D convection-diffusion equation. *Acta Appl. Math.* **131** (2014) 69–140.

[13] D. D'Humières, Generalized Lattice-Boltzmann Equations. American Institute of Aeronautics and Astronautics, Inc. (1992) 450–458.

[14] J. Ding and A. Zhou, Eigenvalues of rank-one updated matrices with some applications. *Appl. Math. Lett.* **20** (2007) 1223–1226.

[15] F. Dubois, Equivalent partial differential equations of a lattice Boltzmann scheme. *Comput. Math. Appl.* **55** (2008) 1441–1449.

[16] F. Dubois, General third order Chapman-Enskog expansion of lattice Boltzmann schemes, in 16th International Conference for Mesoscopic Methods in Engineering and Science, Edinburgh, 22–26 July 2019. Edimburgh, United Kingdom (July 2019).

[17] F. Dubois, Nonlinear fourth order Taylor expansion of lattice Boltzmann schemes. *Asymptotic Anal.* **127** (2021) 297–337.

[18] F. Dubois and P. Lallemand, Towards higher order lattice Boltzmann schemes. *J. Stat. Mech.: Theory Exp. 2009* **6** (2009) P06006.

[19] F. Dubois and P. Lallemand, Quartic parameters for acoustic applications of lattice Boltzmann scheme. *Comput. Math. Appl.* **61** (2011) 3404–3416.

[20] F. Dubois, B. Graille and S.R. Rao, A notion of non-negativity preserving relaxation for a mono-dimensional three velocities scheme with relative velocity. *J. Comput. Sci.* **47** (2020) 101181.

[21] T. Février, *Extension et analyse des schémas de Boltzmann sur réseau: les schémas à vitesse relative*. Ph.D. thesis, Université Paris Sud-Paris XI (2014).

[22] R. Fučík and R. Straka, Equivalent finite difference and partial differential equations for the lattice Boltzmann method. *Comput. Math. Appl.* **90** (2021) 96–103.

[23] Z. Guo and C. Shu, Lattice Boltzmann Method and its Application in Engineering. Vol. 3. World Scientific (2013).

[24] B. Gustafsson, H.-O. Kreiss and J. Oliger, Time Dependent Problems and Difference Methods. Vol. 24. John Wiley & Sons (1995).

[25] M. Hénon, Viscosity of a lattice gas, in Lattice Gas Methods for Partial Differential Equations. CRC Press (1987) 179–207.

[26] F.J. Higuera and J. Jiménez, Boltzmann approach to lattice gas simulations. *EPL (Europhys. Lett.)* **9** (1989) 663.

[27] R.A. Horn and C.R. Johnson, Matrix Analysis. Cambridge University Press (2012).

[28] K. Huang, Statistical Mechanics, 2 edition. John Wiley & Sons (1987).

[29] W.P. Johnson, The curious history of Faà di Bruno's formula. *Am. Math. Monthly* **109** (2002) 217–234.

[30] M. Junk and Z. Yang, Convergence of lattice Boltzmann methods for Navier–Stokes flows in periodic and bounded domains. *Numer. Math.* **112** (2009) 65–87.

[31] M. Junk and W.-A. Yong, Rigorous Navier–Stokes limit of the lattice Boltzmann equation. *Asymptotic Anal.* **35** (2003) 165–185.

[32] M. Junk, A. Klar and L.-S. Luo, Asymptotic analysis of the lattice Boltzmann equation. *J. Comput. Phys.* **210** (2005) 676–704.

[33] E.I. Jury, Theory and Application of the z-Transform Method. Krieger Publishing Co. (1964).

[34] C. Kassel, Quantum Groups, 1 edition. *Graduate Texts in Mathematics*. Springer-Verlag New York (1995).

[35] T. Krüger, H. Kusumaatmaja, A. Kuzmin, O. Shardt, G. Silva and E.M. Viggen, The lattice Boltzmann method. *Springer Int. Publ.* **10** (2017) 1–15.

[36] P. Lallemand and L.-S. Luo, Theory of the lattice Boltzmann method: dispersion, dissipation, isotropy, Galilean invariance, and stability. *Phys. Rev. E* **61** (2000) 6546.

[37] S. Lang, Algebra, 3 edition. *Graduate Texts in Mathematics*. Springer-Verlag New York (2002).

[38] P.D. Lax and R.D. Richtmyer, Survey of the stability of linear finite difference equations. *Commun. Pure Appl. Math.* **9** (1956) 267–293.
[39] G.R. McNamara and G. Zanetti, Use of the Boltzmann equation to simulate lattice-gas automata. *Phys. Rev. Lett.* **61** (1988) 2332.
[40] K.S. Miller, An Introduction to the Calculus of Finite Differences and Difference Equations. Dover Publications (1960).
[41] A.A. Monforte and M. Kauers, Formal Laurent series in several variables. *Expositiones Math.* **31** (2013) 350–367.
[42] I. Niven, Formal power series. *Am. Math. Monthly* **76** (1969) 871–889.
[43] Y.-H. Qian and Y. Zhou, Higher-order dynamics in lattice-based models using the Chapman-Enskog method. *Phys. Rev. E* **61** (2000) 2103.
[44] M.K. Rheinländer, *Analysis of lattice-Boltzmann methods: asymptotic and numeric investigation of a singularly perturbed system*. Ph.D. thesis (2007).
[45] S. Roman, The Umbral Calculus. Dover Publications (2005).
[46] S. Simonis, M. Frank and M.J. Krause, On relaxation systems and their relation to discrete velocity Boltzmann models for scalar advection–diffusion equations. *Philos. Trans. R. Soc. A* **378** (2020) 20190400.
[47] G. Stewart, On the adjugate matrix. *Linear Algebra App.* **283** (1998) 151–164.
[48] J.C. Strikwerda, Finite Difference Schemes and Partial Differential Equations. SIAM (2004).
[49] S. Succi, The Lattice Boltzmann Equation: For Fluid Dynamics and Beyond. Oxford University Press (2001).
[50] S. Suga, An accurate multi-level finite difference scheme for 1D diffusion equations derived from the lattice Boltzmann method. *J. Stat. Phys.* **140** (2010) 494–503.
[51] P. Van Leemput, M. Rheinländer and M. Junk, Smooth initialization of lattice Boltzmann schemes. *Comput. Math. App.* **58** (2009) 867–882.
[52] R.F. Warming and B. Hyett, The modified equation approach to the stability and accuracy analysis of finite-difference methods. *J. Comput. Phys.* **14** (1974) 159–179.
[53] W.-A. Yong, W. Zhao and L.-S. Luo, Theory of the lattice Boltzmann method: derivation of macroscopic equations via the Maxwell iteration. *Phys. Rev. E* **93** (2016) 033310.
[54] M. Zhang, W. Zhao and P. Lin, Lattice Boltzmann method for general convection-diffusion equations: MRT model and boundary schemes. *J. Comput. Phys.* **389** (2019) 147–163.
[55] W. Zhao and W.-A. Yong, Maxwell iteration for the lattice Boltzmann method with diffusive scaling. *Phys. Rev. E* **95** (2017) 033311.
[56] D. Zwillinger, CRC Standard Mathematical Tables and Formulas. Chapman and Hall – CRC (2018).