Fourth-order entropy-stable lattice Boltzmann schemes for hyperbolic systems

Thomas Bellotti, Philippe Helluy, Laurent Navoret

Fourth-order entropy-stable lattice Boltzmann schemes for hyperbolic systems

Thomas Bellotti* Philippe Helluy* Laurent Navoret*

September 6, 2024

Abstract

We present a novel framework for the development of fourth-order lattice Boltzmann schemes to tackle multidimensional nonlinear systems of conservation laws. As for other numerical schemes for hyperbolic problems, high-order accuracy applies only to smooth solutions. Our numerical schemes preserve two fundamental characteristics inherent in classical lattice Boltzmann methods: a local relaxation phase and a transport phase composed of elementary shifts on a Cartesian grid. Achieving fourth-order accuracy is accomplished through the composition of second-order time-symmetric basic schemes utilizing rational weights. This enables the representation of the transport phase in terms of elementary shifts. Introducing local variations in the relaxation parameter during each stage of relaxation ensures entropy stability of the schemes. This not only enhances stability in the long-time limit but also maintains fourth-order accuracy. To validate our approach, we conduct comprehensive testing on scalar equations and systems in both one and two spatial dimensions.

Keywords— lattice Boltzmann, fourth-order, hyperbolic systems of conservation laws, entropy
MSC— 76M28, 65M99, 65M12

Introduction

Lattice Boltzmann schemes [30] have gained acclaim for their computational efficiency and ease of use on modern computer architectures (*e.g.* GPUs), owing to their distinctive structure, comprising a local collision/relaxation phase and a linear transport phase. The latter is constructed through shifts of data on a regular Cartesian grid. Despite their recent application in simulating non-linear systems of conservation laws [23, 27, 21, 9, 8, 4, 28], these methods exhibit lower accuracy for such problems compared to more conventional approaches like Finite Volume and Discontinuous Galerkin methods.

While the attainment of second-order accuracy in lattice Boltzmann schemes is well-understood, achieved by setting relaxation parameters to two [20, 24, 5], obtaining third and fourth-order accuracy proves to be a more intricate challenge [24]. The ability to increase the order is not guaranteed *a priori*—especially for non-linear equations—and depends on the specific lattice Boltzmann scheme in use. When possible, higher accuracy is attained by delicately tuning equilibria that do not contribute to consistency at the leading order, a process that can be complex. Moreover, it is challenging to ascertain the stability of the scheme under such modifications. Consequently, the only fourth-order schemes identified so far only address linear scalar equations in 1D [15, 7], with minimal practical significance, the linear diffusion equation in 2D [16], and a specific kind of coupled Burgers' equations [17]. Finally, a third-order time-accurate / sixth-order space-accurate scheme [40] exists to tackle 1D linear diffusion equations.

This contribution aims at establishing a comprehensive framework for constructing fourth-order kinetic schemes for non-linear systems of multi-dimensional conservation laws. To achieve this objective, a departure from standard lattice Boltzmann schemes is necessary. Nevertheless, the numerical schemes to be developed maintain the two keys to success of any standard lattice Boltzmann scheme, namely the locality of the collision phase and a transport phase made up of simple shifts, thus retain their notorious efficiency, compared to standard kinetic schemes [36, 1]. The essential idea to increase the order of the schemes is to allow both forward and backward steps in time.

In the whole paper, we deal mostly with smooth solutions: the development of limiting procedures in the context of lattice Boltzmann schemes—a recently emerging topic [34]—remains an open and fundamental question worth thorough discussions that we do not address in the present contribution.

*IRMA, Université de Strasbourg, 67000 Strasbourg, France.

Matching the discussion concerning limiters for Finite Volume schemes, limiters can be divided into *a priori* limiters, such as slope limiters, see [39, Chapter 6] for a general overview, and *a posteriori* ones, *e.g.* [18]. The latter correct the solution after a tentative first guess with pathologies has been computed. Ignoring limiters, the present contribution aims at being a proof of concept of a new way to construct high-order lattice Boltzmann schemes for general hyperbolic problems.

The paper is organized as follows. In Section 1, we introduce the system of conservation laws addressed in this paper, along with its relaxation approximation, facilitating the handling of non-linearity. Section 2 outlines our numerical strategy, beginning with a conventional lattice Boltzmann scheme, followed by a time-symmetrization step [19], and eventual composition to achieve fourth-order accuracy [41]. We finish on brief considerations about stability, specifically in terms of the L^2 norm in a basic scalar linear setting. A first batch of numerical experiments, presented in Section 3, empirically confirms the theoretical predictions. In Section 4, we introduce a method for adjusting the relaxation parameter to ensure entropy stability for the numerical scheme. The need for this procedure and the fact that it does not alter the order of the scheme are studied by means of new numerical experiments in Section 5. Finally, in Section 6, we propose, study, and test several variations on our fourth-order scheme. We eventually conclude in Section 7.

1 Target system of conservation laws and relaxation approximation

1.1 Target system of conservation laws

We aim at approximating the solution of the system on $\boldsymbol{u} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^M$:

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \boldsymbol{\varphi}^j(\boldsymbol{u}) = 0, \tag{1}$$

where $\boldsymbol{\varphi}^j : \mathbb{R}^M \to \mathbb{R}^M$ for $j \in [\![1, d]\!]$ are smooth and possibly non-linear fluxes, see [26]. To give a simple example, taking $d = 1$, $M = 1$, and $\varphi(u) = \frac{1}{2}u^2$ yields the inviscid Burgers' equation. We assume that (1) admits a Lax entropy–entropy fluxes pair $(S, G^1, \dots, G^d)$, with $S : \mathbb{R}^M \to \mathbb{R}$ and $G^j : \mathbb{R}^M \to \mathbb{R}$ for $j \in [\![1, d]\!]$, such that $\nabla_{\boldsymbol{u}} \boldsymbol{\varphi}^j \nabla_{\boldsymbol{u}} S = \nabla_{\boldsymbol{u}} G^j$ with S convex. Further properties on this construction can be found in [10, 11].

1.2 Relaxation systems

In order to isolate the non-linearity of the fluxes appearing in (1) into a local relaxation term which is easily tractable, we consider the following discrete-velocity BGK relaxation systems [10, 2, 11] on the distribution functions $\boldsymbol{f}_1, \dots, \boldsymbol{f}_q : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^M$, under the form

$$\partial_t \boldsymbol{f}_k + \sum_{j=1}^{d} V_k^j \frac{\partial \boldsymbol{f}_k}{\partial x_j} = -\frac{1}{\epsilon}(\boldsymbol{f}_k - \boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u})), \qquad k \in [\![1, q]\!]. \tag{2}$$

Here, $q \geq 2$ is the number of discrete velocities, which are $\boldsymbol{V}_k \in \mathbb{R}^d$, $\boldsymbol{u} = \sum_{k=1}^{k=q} \boldsymbol{f}_k$, and we indicate the relaxation time by $\epsilon > 0$. Moreover, the equilibria $\boldsymbol{f}_k^{\mathrm{eq}}$ are non-linear functions of $\boldsymbol{u}$,and fulfill the compatibility relations

$$\boldsymbol{u} = \sum_{k=1}^{q} \boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}), \qquad \boldsymbol{\varphi}^j(\boldsymbol{u}) = \sum_{k=1}^{q} V_k^j \boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}), \quad j \in [\![1, d]\!], \tag{3}$$

under which, the formal limit $\epsilon \to 0^+$ gives that $\boldsymbol{f}_k \approx \boldsymbol{f}_k^{\mathrm{eq}}$ and that the sum of the distribution functions under (2) approximates (1), see [2, 36].

For future use, we introduce the microscopic entropy [22], given by the sum of the kinetic entropies: $\Sigma(\boldsymbol{f}_1, \dots, \boldsymbol{f}_k) = \sum_{k=1}^{k=q} s_k(\boldsymbol{f}_k)$, where the kinetic entropies $s_1, \dots, s_q : \mathbb{R}^M \to \mathbb{R}$ are convex functions of their argument under the so-called characteristic condition. We also have

$$S(\boldsymbol{u}) = \min_{\boldsymbol{u} = \sum_{k=1}^{k=q} \boldsymbol{f}_k} \Sigma(\boldsymbol{f}_1, \dots, \boldsymbol{f}_k) = \Sigma(\boldsymbol{f}_1^{\mathrm{eq}}(\boldsymbol{u}), \dots, \boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u})), \tag{4}$$

meaning that the entropy S stems from a constrained optimization of the microscopic entropy Σ, and that the minimum is reached on the equilibrium. Furthermore, equation (4) tells us that the entropy is an inf-convolution of the kinetic entropies [28], which translates—thanks to the Legendre-Fenchel transform [43] that we shall indicate by a *—into $S^* = \sum_{k=1}^{k=q} s_k^*$. We also additionally request that [22]

$$G^{j,\star} := \boldsymbol{\varphi}^j(\boldsymbol{u}(\boldsymbol{p})) \cdot \boldsymbol{p} - G^j(\boldsymbol{u}(\boldsymbol{p})) = \sum_{k=1}^{q} V_k^j s_k^*, \quad j \in [\![1, d]\!], \tag{5}$$

where $\boldsymbol{p} = \nabla_{\boldsymbol{u}} S$ is the conjugate variable. We here warn the readers about the difference between Legendre-Fenchel–transformed quantities, denoted by the symbol *, and the notation employing a * in (5).

Let us introduce the choices of discrete velocities and relaxation systems that we address in the paper.

1.2.1 $d = 1$: One-dimensional problems

We consider a two-velocities model, so we set $q = 2$, having $V_1 = V > 0$ (indexed by $+$, for the associated distribution function is transported in the positive direction) and $V_2 = -V < 0$ (indexed by $-$). The only way of fulfilling (3) is to select the following equilibria:

$$\boldsymbol{f}_\pm^{\mathrm{eq}}(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u} \pm \frac{1}{2V}\boldsymbol{\varphi}(\boldsymbol{u}).$$

Using the change of basis $\boldsymbol{u} = \boldsymbol{f}_+ + \boldsymbol{f}_-$ and $\boldsymbol{v} = V(\boldsymbol{f}_+ - \boldsymbol{f}_-)$, (2) can be recast as

$$\begin{cases} \partial_t \boldsymbol{u} + \partial_x \boldsymbol{v} = 0, \\ \partial_t \boldsymbol{v} + V^2 \partial_x \boldsymbol{u} = -\frac{1}{\epsilon}(\boldsymbol{v} - \boldsymbol{\varphi}(\boldsymbol{u})), \end{cases} \tag{6}$$

being the well-known Jin-Xin relaxation system [32]. In the lattice Boltzmann nomenclature, this is a $\mathrm{D_1Q_2^M}$ relaxation system [27].

Let us provide a few examples of problems that can be tackled using this scheme.

Example 1 (Linear transport equation)**.** *Let $M = 1$ and $\varphi(u) = au$. We consider the classic quadratic entropy given by $S(u) = \frac{1}{2}u^2$, thus the entropy flux is given by $G(u) = \frac{a}{2}u^2$. We obtain $S^*(p) = \frac{1}{2}p^2$ and $G^\star(p) = \frac{a}{2}p^2$ (observe that $G^*(p) = \frac{1}{2a}p^2$). The dual kinetic entropies satisfying the imposed constraints are*

$$s_\pm^*(p) = \frac{1}{4}\Big(1 \pm \frac{a}{V}\Big)p^2, \qquad \textit{thus} \qquad s_\pm(f_\pm) = \frac{V}{V \pm a}(f_\pm)^2.$$

The condition providing the convexity of the kinetic entropy is the sub-characteristic condition $|a| < V$ that is found in many works, see [10, 2].

Example 2 (Burger' equation)**.** *Let $M = 1$ and $\varphi(u) = \frac{1}{2}u^2$. We take $S(u) = \frac{1}{2}u^2$, thus the entropy flux is given by $G(u) = \frac{1}{3}u^3$. Analogously to the linear case, we obtain*

$$s_\pm^*(p) = \frac{1}{4}p^2 \pm \frac{1}{12V}p^3, \qquad s_\pm(f_\pm) = \frac{V^2}{6}\Big(\Big(1 \pm \frac{4f_\pm}{V}\Big)^{3/2} \mp \frac{6f_\pm}{V} - 1\Big).$$

Again, convexity comes by $|u| < V$.

Example 3 (Shallow water system)**.** *Let $M = 2$, $\boldsymbol{u} = (u_1, u_2) = (h, hu)$, and $\boldsymbol{\varphi}(h, hu) = (hu, hu^2 + \frac{g}{2}h^2)$ where $g > 0$ is the gravity acceleration. As in [12, Section 3.2], we take $S(h, hu) = \frac{1}{2}hu^2 + \frac{g}{2}h^2$, hence $G(h, hu) = \frac{1}{2}hu^3 + \frac{g}{2}h^2u$. We have the dual variables $p_1 = -\frac{1}{2}u^2 + gh$ and $p_2 = u$ and the dual kinetic entropies given by*

$$s_\pm^*(\boldsymbol{p}) = \frac{(V \pm p_2)(2p_1 + p_2^2)^2}{16gV}.$$

Convexity comes under the condition $V > |u| + \sqrt{gh}$, see [28], which guarantees that the kinetic velocity is larger than the one of the fastest wave in the system. The kinetic entropies can be found analytically, albeit with complicated formulæ: they are given by $s_\pm(\boldsymbol{f}_\pm) = \boldsymbol{p}_\pm(\boldsymbol{f}_\pm) \cdot \boldsymbol{f}_\pm - s_\pm^*(\boldsymbol{p}_\pm(\boldsymbol{f}_\pm))$, where $\boldsymbol{p}_\pm(\boldsymbol{f}_\pm)$ is the solution of $\boldsymbol{f}_\pm = \nabla_{\boldsymbol{p}} s_\pm^*(\boldsymbol{p}_\pm)$. One can see that the first equation to solve is linear in $p_\pm^1$, thus we obtain

$$p_\pm^1(p_\pm^2) = \frac{4Vgf_\pm^1 - V(p_\pm^2)^2 \mp (p_\pm^2)^3}{2(V \pm p_\pm^2)},$$

which corresponds to a third-order equation on $p_\pm^2$ only:

$$\pm Vg(f_\pm^1)^2 + f_\pm^1(p_\pm^2)^3 - V^2 f_\pm^2 \pm (2Vf_\pm^1 \mp f_\pm^2)(p_\pm^2)^2 + (V^2 f_\pm^1 \pm 2Vf_\pm^2)p_\pm^2 = 0.$$

This equation can be solved with the well-known formula for cubic equations, upon choosing a specific branch.

1.2.2 $d = 2$: Two-dimensional problems

We consider a four-velocities model, so we set $q = 4$, having $\boldsymbol{V}_1 = (V, 0)$ (indexed by $+, x$) with $V > 0$, $\boldsymbol{V}_2 = (0, V)$ (indexed by $+, y$), $\boldsymbol{V}_3 = (-V, 0)$ (indexed by $-, x$), and $\boldsymbol{V}_4 = (0, -V)$ (indexed by $-, y$). There are several ways of enforcing (3): the one we select is [25, Chapter 3]

$$\boldsymbol{f}_{\pm,x/y}^{\mathrm{eq}}(\boldsymbol{u}) = \frac{1}{4}\boldsymbol{u} \pm \frac{1}{2V}\boldsymbol{\varphi}_{x/y}(\boldsymbol{u}).$$

Using the change of basis $\boldsymbol{u} = \boldsymbol{f}_{+,x} + \boldsymbol{f}_{+,y} + \boldsymbol{f}_{-,x} + \boldsymbol{f}_{-,y}$, $\boldsymbol{v}_x = V(\boldsymbol{f}_{+,x} - \boldsymbol{f}_{-,x})$, $\boldsymbol{v}_y = V(\boldsymbol{f}_{+,y} - \boldsymbol{f}_{-,y})$, and $\boldsymbol{w} = V^2(\boldsymbol{f}_{+,x} - \boldsymbol{f}_{+,y} + \boldsymbol{f}_{-,x} - \boldsymbol{f}_{-,y})$, we get

$$\begin{cases} \partial_t \boldsymbol{u} + \partial_x \boldsymbol{v}_x + \partial_y \boldsymbol{v}_y = 0, \\ \partial_t \boldsymbol{v}_x + V^2 \partial_x \big(\frac{1}{2}\boldsymbol{u} + \frac{1}{2V^2}\boldsymbol{w}\big) = -\frac{1}{\epsilon}(\boldsymbol{v}_x - \boldsymbol{\varphi}_x(\boldsymbol{u})), \\ \partial_t \boldsymbol{v}_y + V^2 \partial_y \big(\frac{1}{2}\boldsymbol{u} - \frac{1}{2V^2}\boldsymbol{w}\big) = -\frac{1}{\epsilon}(\boldsymbol{v}_y - \boldsymbol{\varphi}_y(\boldsymbol{u})), \\ \partial_t \boldsymbol{w} + V^2 \partial_x \boldsymbol{v}_x - V^2 \partial_y \boldsymbol{v}_y = -\frac{1}{\epsilon}\boldsymbol{w}. \end{cases} \tag{7}$$

This can be called a $\mathrm{D_2Q_4^M}$ relaxation system [23].

Example 4 (Linear transport equation)**.** *Let $M = 1$ and $\varphi^1(u) = (a_x u)$, $\varphi^2(u) = a_y u$. We consider $S(u) = \frac{1}{2}u^2$, thus the entropy flux is given by $G(u) = (\frac{a_x}{2}u^2, \frac{a_y}{2}u^2)$. We obtain $S^*(p) = \frac{1}{2}p^2$ and $G^\star(p) = (\frac{a_x}{2}p^2, \frac{a_y}{2}p^2)$. Possible dual kinetic entropies satisfying the constraints are*

$$s_{\pm,x/y}^*(p) = \frac{1}{4}\Big(\frac{1}{2} \pm \frac{a_{x/y}}{V}\Big)p^2, \qquad \textit{thus} \qquad s_{\pm,x/y}(f_{\pm,x/y}) = \frac{2V}{V \pm 2a_{x/y}}(f_{\pm,x/y})^2.$$

The conditions providing the convexity of the kinetic entropies read $|a_{x/y}| < V/2$, see [28].

Besides the specific choices of discrete velocities that we have presented hitherto, the techniques developed in the paper work as long as $\boldsymbol{V}_k \in V\mathbb{Z}^d$ for $k \in [\![1, q]\!]$ with a given $V \in \mathbb{R}$. For example, one could employ the well-known $\mathrm{D_2Q_9}$ scheme [37].

2 Numerical schemes

Now that we have set the preliminaries concerning relaxation systems at a continuous level, we are ready to propose several numerical schemes to tackle (1) inspired by (2).

The first step is—in Section 2.1—the introduction of transport and relaxation phases, yielding the standard lattice Boltzmann scheme. This scheme can be easily made second-order accurate; however, it is hard to push it towards higher accuracy because the scheme lacks time-symmetry. The time-symmetry property is indeed useful for increasing the order of the scheme through palindromic composition [41]. The second step in the process, presented in Section 2.2, is conducted by symmetrization, without increasing the actual order of the scheme. The latter is the aim of the third step in the process and is obtained by composition, as detailed in Section 2.3.

For the space discretization, we employ a uniform Cartesian mesh $\Delta x\mathbb{Z}^d$—also known as lattice—of step $\Delta x > 0$. The uniform time step is denoted by $\Delta t > 0$ and is specified in what follows.

2.1 Standard lattice Boltzmann schemes

The left-hand side of (2) is made up of linear transport equations with constant velocities $\boldsymbol{V}_k$, whereas the right-hand side represents local relaxations. It is therefore natural to split these two terms and let them undergo different treatments. In this part of the paper, we consider $k \in [\![1, q]\!]$ be the index of any discrete velocity.

2.1.1 Transport

The equations associated with the left-hand side of (2) are solved using any consistent one-step scheme for the linear transport equation. Recall that the kinetic velocities $\boldsymbol{V}_k$ are integer multiples of V, which is adjusted so that the kinetic velocity V fulfills $V\Delta t/\Delta x = \kappa \in \mathbb{N}^*$. In this way, the transport phase is indeed given by elementary shifts on the grid, which is the natural issue of any consistent one-step scheme in this peculiar framework. The fact of shifting data sticking to the discrete grid makes our approach a lattice Boltzmann approach. This reads

$$\boldsymbol{f}_k(\Delta t, \boldsymbol{x}) = \boldsymbol{f}_k(0, \boldsymbol{x} - \boldsymbol{V}_k\Delta t) = \boldsymbol{f}_k(0, \boldsymbol{x} - \underbrace{\kappa\tfrac{\boldsymbol{V}_k}{V}}_{\in\mathbb{Z}^d}\Delta x), \qquad \boldsymbol{x} \in \Delta x\mathbb{Z}^d. \tag{8}$$

Gathering all the distribution functions together, this transport phase is denoted by $\mathsf{T}(\Delta t)$. This operator is made up of exact schemes for the transport equations; however, one must be aware that we have performed an overall splitting between transport and relaxation, thus this does not ensure accuracy with respect to the original problem (2)—and *a fortiori* with (1)—above first-order.

2.1.2 Relaxation

The relaxation part, *i.e.* the right-hand side of (2), is solved using a trapezoidal quadrature, see [20], which is second-order accurate. Using the fact that the relaxation phase conserves $\boldsymbol{u}$ and thus any equilibrium fulfills $\boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}(\Delta t)) = \boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}(0))$, the algorithm can be fortunately kept explicit and thus reads

$$\boldsymbol{f}_k(\Delta t) = \frac{2\epsilon - \Delta t}{2\epsilon + \Delta t}\boldsymbol{f}_k(0) + \frac{2\Delta t}{2\epsilon + \Delta t}\boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}(0)) \xrightarrow[\epsilon\to 0]{} -\boldsymbol{f}_k(0) + 2\boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}(0)). \tag{9}$$

The space variable is not listed since the relaxation step is local and performed at each point of the spatial grid $\Delta x\mathbb{Z}^d$. We consider the limit $\epsilon \to 0$ in (9), thus a relaxation independent of Δt. Therefore, ϵ no longer appears in the numerical scheme. The relaxation system and its discretization must be seen as an intermediate step to propose a numerical scheme for another equation where no relaxation time exists, namely (1). More generally, a relaxation independent of Δt can be written as

$$\boldsymbol{f}_k(\Delta t) = (1-\omega)\boldsymbol{f}_k(0) + \omega\boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}(0)). \tag{10}$$

with a relaxation parameter $\omega \in (0, 2]$, and we indicate it by R_ω. Whenever we write R, we mean $\mathsf{R}_{\omega=2}$. Notice that $\mathsf{R}_{\omega=2}$ is an involution: $\mathsf{R}_{\omega=2}\mathsf{R}_{\omega=2} = \mathsf{Id}$, which is false for relaxation parameters $\omega < 2$. This feature of the relaxation operator is crucial in what follows.

2.1.3 Overall lattice Boltzmann scheme

One can show that the scheme $\boldsymbol{\psi}(\Delta t) = \mathsf{R}\mathsf{T}(\Delta t)$ (or $\boldsymbol{\psi}(\Delta t) = \mathsf{T}(\Delta t)\mathsf{R}$) is a second-order scheme to solve (1). This boils down to the standard SRT (Single-Relaxation-Time) lattice Boltzmann scheme with relaxation parameter equal to two [27], which is second-order accurate, see [24, 5].

However, these schemes are not time-symmetric. Time symmetry is defined by

$$\boldsymbol{\psi}(\Delta t)\boldsymbol{\psi}(-\Delta t) = \mathsf{Id} \qquad \text{and} \qquad \boldsymbol{\psi}(0) = \mathsf{Id}. \tag{11}$$

In the present case, $\boldsymbol{\psi}(\Delta t)\boldsymbol{\psi}(-\Delta t) \neq \mathsf{Id}$, where whenever we employ negative time-steps, it is like if we simply reverse $V \mapsto -V$ using a positive time-step, *i.e.* the distribution functions are transported in the opposite direction compared to what they would do when $V > 0$. Time-symmetry is a highly desirable feature that fosters the increase of the order by using composition procedures, in the spirit of what [41] presents. We now try to fix this problem.

2.2 Symmetric lattice Boltzmann schemes

The first idea is to use a sort of Strang formula that would read $\boldsymbol{\psi}(\Delta t) = \mathsf{T}(\frac{\Delta t}{2})\mathsf{R}\mathsf{T}(\frac{\Delta t}{2})$. Since the transport phase is made up of elementary shifts on the grid, we have that $\boldsymbol{\psi}(\Delta t)\boldsymbol{\psi}(-\Delta t) = \mathsf{Id}$. However, $\boldsymbol{\psi}(0) \neq \mathsf{Id}$, thus this operator is not suitable to be employed to increase the overall order of the numerical scheme. Juxtaposing two half-steps of this operator—following [19]—we can take advantage of the involution property of the relaxation operator R, and gain

$$\boldsymbol{\psi}(\Delta t) = \mathsf{T}\Big(\frac{\Delta t}{4}\Big)\mathsf{R}\mathsf{T}\Big(\frac{\Delta t}{4}\Big)\mathsf{T}\Big(\frac{\Delta t}{4}\Big)\mathsf{R}\mathsf{T}\Big(\frac{\Delta t}{4}\Big) = \mathsf{T}\Big(\frac{\Delta t}{4}\Big)\mathsf{R}\mathsf{T}\Big(\frac{\Delta t}{2}\Big)\mathsf{R}\mathsf{T}\Big(\frac{\Delta t}{4}\Big). \tag{12}$$

One can easily check that, thanks to the fact that R is an involution, we have (11) hence $\boldsymbol{\psi}$ defined through (12) is time-symmetric. So far, nothing special has been done to increase the order, so (12) is just another second-order accurate solver, as the one provided by the Strang formula $\mathsf{T}(\frac{\Delta t}{2})\mathsf{R}\mathsf{T}(\frac{\Delta t}{2})$.

2.3 Fourth-order lattice Boltzmann scheme

The symmetry property is crucial to obtain high-order schemes by composition. Let us assume that the symmetric lattice Boltzmann operator $\boldsymbol{\psi}(\Delta t)$ defined in (12) leads to a converging lattice Boltzmann scheme. In other words, this means that $\boldsymbol{\psi}(\Delta t)$ is an approximation of the flow of a formal differential equation $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_q)'(t, \cdot) = \boldsymbol{g}((\boldsymbol{f}_1, \dots, \boldsymbol{f}_q)(t, \cdot))$. In the Lie groups theory, *cf.* [41, Introduction] and [29, Chapter 2 and 3], it is common to denote the flow of the differential equation by the exponential notation $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_q)(t, \cdot) = e^{t\boldsymbol{g}}((\boldsymbol{f}_1, \dots, \boldsymbol{f}_q)(0, \cdot))$. This is a generalization of the matrix exponential from the linear case. With this, we have $\boldsymbol{\psi}(\Delta t) \approx e^{\Delta t \boldsymbol{g}}$. In [21, 28], it is shown that $\sum_{k=1}^{k=q} \boldsymbol{g}_k(\boldsymbol{f}_1, \dots, \boldsymbol{f}_q)$ depends only on $\boldsymbol{u} = \sum_{k=1}^{k=q} \boldsymbol{f}_k$, and

$$\sum_{k=1}^{q} \boldsymbol{g}_k(\boldsymbol{u}) = -\sum_{j=1}^{d} \frac{\partial}{\partial x_j} \boldsymbol{\varphi}^j(\boldsymbol{u}),$$

so that the scheme eventually solves (1), plus other equations that can be made explicit. It is possible to be more precise: by time-symmetry, using [41, Theorem 19], there exists a vector field $\boldsymbol{d}$ such that

$$\boldsymbol{\psi}(\Delta t) = e^{\Delta t \boldsymbol{g} + \Delta t^3 \boldsymbol{d}} + \mathcal{O}(\Delta t^5),$$

indicating that $\boldsymbol{\psi}$ is second-order accurate. Remark that there is no guarantee on the fact that $\boldsymbol{g}$ and $\boldsymbol{d}$ commute. Following [41, Equation (4.4)], we look for an overall operator—constructed by composition—under the form

$$\boldsymbol{\phi}(\Delta t) = \boldsymbol{\psi}(\alpha \Delta t)^n \boldsymbol{\psi}(\beta \Delta t) \boldsymbol{\psi}(\alpha \Delta t)^n, \tag{13}$$

where $n \in \mathbb{N}^*$. According to [41, Theorem 22], the operator $\boldsymbol{\phi}$ is such that

$$\boldsymbol{\phi}(\Delta t) = e^{\Delta t (2n\alpha + \beta)\boldsymbol{g} + \Delta t^3 (2n\alpha^3 + \beta^3)\boldsymbol{d}} + \mathcal{O}(\Delta t^5),$$

thus it has a local truncation error of order five—thus it is globally accurate at order four—provided that the conditions

$$2n\alpha + \beta = 1, \tag{14}$$

$$2n\alpha^3 + \beta^3 = 0, \tag{15}$$

are satisfied. For we want to deal with a lattice Boltzmann approach, characterized by the fact that the transport phase (8) is made up of integer shifts on the discrete spatial grid $\Delta x \mathbb{Z}^d$, we would like $\alpha, \beta \in \mathbb{Q}$.[1] In order to fulfill (15), one can easily see that either α or β has to be negative, meaning that steps with transport according to the sign of the discrete velocities are interspersed with steps in the opposite direction. Otherwise said, the price to pay to obtain fourth-order consistency is to alternate steps both forward and backward in time. Inserting (14) into (15) gives $2n\alpha^3 + (1 - 2n\alpha)^3 = 0$. For $n = 1$ the only real solution is irrational. The same holds for $n = 2, 3$, and these cases are not of interest in our setting, because once put back into (14), both α and β are irrational and eventually incommensurable. For $n = 4$, we have the rational solution $\alpha = 1/6$, hence $\beta = -1/3$. Therefore, the formula that we retain is

$$\boldsymbol{\phi}(\Delta t) = \boldsymbol{\psi}\Big(\frac{\Delta t}{6}\Big)^4 \boldsymbol{\psi}\Big(-\frac{\Delta t}{3}\Big) \boldsymbol{\psi}\Big(\frac{\Delta t}{6}\Big)^4. \tag{16}$$

Looking at (16) and (12), we see that the shortest transport phase features a time-step equal to $\Delta t/24$. This means that "particles" roughly travel $V\Delta t/(24\Delta x)$ gridpoints at each time the transport operator is called. To ensure that the scheme remains a lattice Boltzmann scheme, we enforce that $V\Delta t/(24\Delta x) = \kappa \in \mathbb{N}^*$. The time step is given by $\Delta t = 24\kappa \Delta x / V$. We consistently take $\kappa = 1$. The kinetic velocity is freely chosen, still ensuring that all the waves in (1) are resolved:

$$V \gtrsim \max_{i \in [\![1, M]\!]} |\lambda_i|, \tag{17}$$

where the λ_i are the eigenvalues of the Jacobian matrix of $\boldsymbol{\varphi}$, *i.e.* the velocities of the waves. The kinetic velocity should not be too large compared to the fastest wave in the system, to ensure accuracy.

[1] It could still be possible to obtain a lattice Boltzmann scheme whenever $\alpha, \beta \in \mathbb{R} \setminus \mathbb{Q}$, provided that α and β are commensurable.

Remark 1. *We see that according to* (16)*, at each time step, "particles" undergo* $1 \times 4 \times 4 \times \kappa = 16\kappa$ *shifts according to the sign of their velocities and eight relaxations, followed by* $2 \times 1 \times 4 \times \kappa = 8\kappa$ *shifts (of twice the length) in the opposite direction and two relaxations, followed again by* $1 \times 4 \times 4 \times \kappa = 16\kappa$ *shifts according to the sign of their velocities and eight relaxations.*

Remark 2. *By making the change of variable* $\Delta t \mapsto 24\Delta t$*, we can interpret things in another manner. The overall scheme is fourth-order accurate if we observe it every* 24 *time steps doing*

$$\overbrace{\mathbf{T}(\Delta t)\mathbf{R}\mathbf{T}(2\Delta t)\mathbf{R}\mathbf{T}(\Delta t) \times \cdots \times \mathbf{T}(\Delta t)\mathbf{R}\mathbf{T}(2\Delta t)\mathbf{R}\mathbf{T}(\Delta t)}^{4 \text{ times}} \times \mathbf{T}(-2\Delta t)\mathbf{R}\mathbf{T}(-4\Delta t)\mathbf{R}\mathbf{T}(-2\Delta t) \times \underbrace{\mathbf{T}(\Delta t)\mathbf{R}\mathbf{T}(2\Delta t)\mathbf{R}\mathbf{T}(\Delta t) \times \cdots \times \mathbf{T}(\Delta t)\mathbf{R}\mathbf{T}(2\Delta t)\mathbf{R}\mathbf{T}(\Delta t)}_{4 \text{ times}},$$

which means having done 32 *steps (of length* κ*) forward and* 8 *steps backward (of twice the step) with a specific interleaving.*

Remark 3 (Cost of the scheme)**.** *Considering* (16) *and the previous remark, the cost of the whole algorithm might seem very high. However, the time-marching procedure is totally made up of traditional transport and relaxation steps of a standard lattice Boltzmann method, and techniques to parallelize and deploy them of modern architectures (*e.g. *GPUs) are available and indeed employed. A crucial advantage of the splitting strategy is that it does not require additional storage, as it is the case with a Runge-Kutta approach. Moreover, the numerical solution in the inner sub-steps is consistent—*i.e. *meaningful—and simply second-order accurate. We can thus consider the method as an (almost) standard second-order lattice Boltzmann scheme, where fourth-order accuracy is observed at specific steps of the time-marching procedure.*

2.4 L^2 stability

We see in Section 4 that—with a simple procedure acting on the relaxation parameter—our scheme can possess excellent features concerning entropy stability, ensuring stability in a non-linear framework. This comes from the fact that entropy provides—reminding us of the work by [33]—the right weighted norm to take the effect of the relaxation into account. It is more involved to study the stability with respect to the L^2 norm, which furthermore applies only to a linear setting. Even for the standard $\mathrm{D_2Q_4}$ from Section 2.1, no explicit L^2 stability condition is known, to the best of our knowledge.

Nevertheless, we start by a brief study concerning L^2 stability. We consider the case of $d = 1$ with one conservation law $M = 1$, thus we use a $\mathrm{D_1Q_2}$ scheme. Moreover, we consider a linear problem: $\varphi(u) = au$. A polynomial with complex coefficients is said to be a simple *von Neumann* polynomial if its roots are in the closed unit disk and those on the unit circle are simple. Then, the corresponding Finite Difference scheme computed using the characteristic polynomial of $\boldsymbol{\phi}$ [5] is L^2 stable if the characteristic polynomial, upon considering its Fourier transform, is a simple *von Neumann* polynomial for every frequency. Conversely, the original lattice Boltzmann scheme $\boldsymbol{\phi}$ is L^2 stable if its minimal polynomial is a simple *von Neumann* polynomial for every frequency, see [7]. It is well-know [27] that the standard lattice Boltzmann scheme ($\mathbf{R}\mathbf{T}(\Delta t)$ or $\mathbf{T}(\Delta t)\mathbf{R}$) from Section 2.1 is stable for the L^2 norm under the strict condition

$$\frac{|a|\Delta t}{\kappa \Delta x} < 1, \tag{18}$$

which is the CFL condition of a leap-frog scheme, *cf.* [45]. For the new scheme $\boldsymbol{\phi}(\Delta t)$ given by (16), we cannot conclude that it is stable provided that $\boldsymbol{\psi}(\frac{\Delta t}{6})$ and $\boldsymbol{\psi}(-\frac{\Delta t}{3})$ are stable, because these operators are not simultaneously diagonalizable, for they do not commute. We have to study the eigenvalues of $\boldsymbol{\phi}(\Delta t)$. In particular, we focus on its characteristic polynomial—whose roots include those of the minimal polynomial—as long as it allows concluding. Otherwise, we switch to the minimal polynomial. The characteristic polynomial reads

$$\det(z\mathbf{Id} - \hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)) = z^2 - \mathrm{tr}(\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x))z + \det(\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)), \tag{19}$$

for $|\xi\Delta x| \leq \pi$. Here, hats denote Fourier-transformed quantities. Since $\det(\hat{\mathbf{T}}(\Delta t)(\xi\Delta x)) = 1$, $\det(\hat{\mathbf{R}}(\xi\Delta x)) = -1$, and $\boldsymbol{\phi}$ is made up of an even number of relaxations $\mathbf{R}$, we can use the formula for the determinant of a product of matrices and thus obtain $\det(\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)) = 1$. For the trace appearing in (19), less can

be said. Its explicit expression—computed using a computer algebra system—is involved and provided in Appendix A for the interested readers. Yet, we observe that $\mathrm{tr}(\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)) \in \mathbb{R}$. Using the results by [42], the characteristic polynomial is a simple *von Neumann* polynomial if and only if the sole root of $\frac{\mathrm{d}}{\mathrm{d}z}\det(z\mathbf{Id} - \hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x))$ is in the open unit disk. This reads

$$|\mathrm{tr}(\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x))| < 2. \tag{20}$$

One can be easily persuaded, *cf.* Appendix A, that (20) if fulfilled as long as $|a|/V < 1$ (except at $\xi\Delta x = 0, \frac{\pi}{2\kappa}$ which need to be analyzed separately) and, when going beyond this value, that the left-hand side of (20) is critically maximal at $\xi\Delta x = \frac{\pi}{4\kappa}$. Evaluating (20) at this value gives an inequality of degree 16 in a/V, which, solved using `sage-math`, exactly gives (21). This allows to conclude that under this condition, except for $\xi\Delta x = 0, \frac{\pi}{2\kappa}$, the characteristic polynomial of $\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)$ and thus the minimal polynomial are simple *von Neumann*. The two exceptional cases are exactly those where there is a gap between characteristic and minimal polynomials [7]. Indeed $\det(z\mathbf{Id} - \hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x))|_{\xi\Delta x=0,\pi/(2\kappa)} = (z-1)^2$, whereas $\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)|_{\xi\Delta x=0,\pi/(2\kappa)} = \mathbf{Id}$ indicates that $z-1$ is the minimal polynomial in this case, and it is simple. Finally, we observe that we cannot include the case $|a|/V = 1$, since $\det(z\mathbf{Id} - \hat{\boldsymbol{\phi}}(\Delta t)(\frac{\pi}{4\kappa}))|_{|a|/V=1} = (z-1)^2$ but

$$\hat{\boldsymbol{\phi}}(\Delta t)(\tfrac{\pi}{4\kappa})|_{|a|/V=1} = \begin{pmatrix} 1 & 32i \\ 0 & 1 \end{pmatrix}.$$

This allows concluding on the L^2 stability of the lattice Boltzmann scheme, see the following result.

Proposition 1. *Let $d = 1$, $M = 1$, and $\varphi(u) = au$. Consider a D_1Q_2 scheme. Then $\boldsymbol{\phi}(\Delta t)$ is L^2-stable under the condition*

$$\frac{|a|}{V} = \frac{|a|\Delta t}{24\kappa\Delta x} < 1. \tag{21}$$

Observe that there is no difference between (18) and (21), because in between, we have made the change of variable $\Delta t \mapsto \frac{1}{24}\Delta t$. This is indeed the sub-characteristic condition found in Example 1.

3 Numerical experiments: Order of the scheme

We now proceed to several numerical experiments to confirm the theoretical order of the method we have devised in a non-linear context. We consider both scalar problems and systems in one and two spatial dimensions. All the tests have been implemented and parallelized on GPUs using `OpenCL` [4].

3.1 Non-linear scalar problem: Burgers' equation in 1D

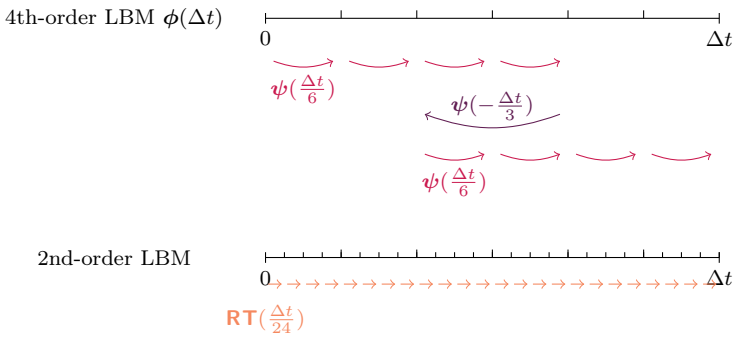

Figure 1: Way of devising a fair comparison between our new fourth-order lattice Boltzmann scheme (top) and the original second-order lattice Boltzmann scheme (bottom).

To test the fourth-order convergence of our solver $\boldsymbol{\phi}$ in a genuinely non-linear setting, we consider the Burgers' equation on a bounded domain $[0,1]$ endowed with periodic boundary conditions. The initial datum is the point-wise discretization of $u(t=0,x) = \sin(2\pi x)$ with data taken at equilibrium, which means that $\boldsymbol{f}_k(t=0) = \boldsymbol{f}_k^{\mathrm{eq}}(\boldsymbol{u}(t=0))$ for $k \in [\![1,q]\!]$. In order to fulfill the sub-characteristic condition,

the kinetic velocity is $V = 1.2$. The final time of the simulation, at which we measure errors, is $T = 1/10$, which is before the solution exhibits a shock wave.

We would also like to compare the accuracy of our approach against the standard second-order lattice Boltzmann method given by $\mathbf{RT}(\Delta t)$. To ensure a fair comparison between errors at roughly the same computational cost, we have to proceed as indicated in Figure 1, namely consider the scheme given by $(\mathbf{RT}(\frac{\Delta t}{24}))^{24}$. This allows to have roughly the same number of operations between the second-order scheme (24 steps) and the fourth-order scheme (32 steps) for advancing of Δt in time. The additional steps are needed for the fourth-order scheme sometimes goes backward in time.

Table 1: Errors and order of convergence in the L^2 norm for the Burgers' equation using the second-order lattice Boltzmann scheme and our new fourth-order scheme.

	2nd-order LBM		4th-order LBM	
Δx	L^2 error	Order	L^2 error	Order
2.000E-03	8.592E-05		3.370E-06	
1.250E-03	3.358E-05	2.00	1.552E-06	1.65
7.813E-04	1.404E-05	1.86	1.742E-07	4.65
4.883E-04	5.494E-06	2.00	3.365E-08	3.50
3.053E-04	2.160E-06	1.99	5.184E-09	3.98
1.908E-04	7.799E-07	2.17	8.688E-10	3.80
1.193E-04	3.057E-07	1.99	1.221E-10	4.18
7.454E-05	1.287E-07	1.84	2.109E-11	3.74

The results are given in Table 1 and show second-order convergence for the original lattice Boltzmann scheme and fourth-order convergence for the new scheme. Comparing the standard second-order lattice Boltzmann scheme to our new scheme in the framework where they have roughly the same computational cost, we see that even at a very coarse resolution, the fourth-order scheme outperforms the standard scheme being at least twenty times more accurate.

3.2 Non-linear system: Shallow water equations in 1D

Table 2: Error estimations and order of convergence in the L^2 metric for the shallow water system.

	Height h		Velocity u	
Δx	Err-Estim$_h$	Order	Err-Estim$_u$	Order
7.8125E-03	5.8333E-06		2.9538E-05	
3.9063E-03	7.9483E-07	2.88	1.6474E-06	4.16
1.9531E-03	1.0703E-07	2.89	4.8759E-08	5.08
9.7656E-04	7.6700E-09	3.80	2.9001E-09	4.07
4.8828E-04	4.9440E-10	3.96	1.8273E-10	3.99
2.4414E-04	3.1134E-11	3.99	1.1456E-11	4.00
1.2207E-04	1.9495E-12	4.00	7.1665E-13	4.00
6.1035E-05	1.2202E-13	4.00	4.5492E-14	3.98

We now test the order of the method for a system of equations. We consider the same setting as Section 3.1 except for the fact that we deal with the shallow water system with gravity $g = 1$ and initial datum $(h, u)(t = 0, x) = (1/2 + 1/5 \sin(2\pi x), 0)$. Simulations are carried until a final time of $T = 5/16$ with kinetic velocity $V = 1.2$. For the exact solution of the problem is difficult to find, the fourth-order accuracy of the method is demonstrated using the following error estimators:

$$\text{Err-Estim}_h = \sqrt{\sum_{k\in\mathbb{Z}} \Delta x |h_{\Delta x}(T, k\Delta x) - h_{\Delta x/2}(T, k\Delta x)|^2},$$

$$\text{Err-Estim}_u = \sqrt{\sum_{k\in\mathbb{Z}} \Delta x |u_{\Delta x}(T, k\Delta x) - u_{\Delta x/2}(T, k\Delta x)|^2},$$

where $h_{\Delta x}$ and $u_{\Delta x}$ indicate the discrete solution of our fourth-order scheme computed with space step Δx. We expect fourth-order convergence, which translates into $\text{Err-Estim}_h, \text{Err-Estim}_u = \mathcal{O}(\Delta x^4)$ as $\Delta x \to 0$. The numerical results in Table 2 give the expected trends for both height h and velocity u.

Table 3: Error estimations and order of convergence in the L^2 metric for the Burgers' equation in 2D.

Δx	Err-Estim	Order
6.667E-02	8.784E-02	
3.226E-02	2.643E-02	1.65
1.587E-02	9.601E-03	1.43
7.874E-03	1.854E-03	2.35
3.922E-03	3.155E-04	2.54
1.957E-03	3.285E-05	3.25
9.775E-04	1.155E-06	4.82
4.885E-04	2.541E-08	5.50
2.442E-04	1.749E-09	3.86

3.3 Multidimensional non-linear scalar problem: Burgers' equation in 2D

To test our approach in 2D, we consider the Burgers' equation $\partial_t u + \partial_x(u^2/2) + \partial_y(3u^2/10) = 0$ on the bounded domain $[0,1]^2$ endowed with periodic boundary conditions. The initial datum is a narrow Gaussian profile, given by $u(t=0,\boldsymbol{x}) = \exp(-100|\boldsymbol{x} - (1/2,1/2)^T|^2)$. Simulations are carried until final time $T = 1/16$ where we measure

$$\text{Err-Estim} = \sqrt{\sum_{\boldsymbol{k}\in\mathbb{Z}^2} \Delta x^2 |u_{\Delta x}(T,\boldsymbol{k}\Delta x) - u_{\Delta x/2}(T,\boldsymbol{k}\Delta x)|^2},$$

given in Table 3. Once again we observe that our numerical method is fourth-order accurate, as expected.

4 Entropy stability

We now address a more useful notion of stability compared to the one studied in Section 2.4, which is going to be especially suitable for the non-linear framework. The total microscopic entropy in the domain—at each time—is given by

$$\sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} \Sigma(\boldsymbol{f}_1(\boldsymbol{x}),\dots,\boldsymbol{f}_q(\boldsymbol{x})).$$

If the computational domain is infinite or periodic boundary conditions on the distribution functions are enforced, this quantity is conserved throughout the transport phase $\mathsf{T}(\Delta t)$, since it is made up of shifts for each discrete velocity, without mixing the distribution functions between them. Mathematically, this reads

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} \Sigma(\mathsf{T}(\Delta t)(\boldsymbol{f}_1(\boldsymbol{x}),\dots,\boldsymbol{f}_q(\boldsymbol{x})))\\
&\qquad = \sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d}\sum_{k=1}^{q} s_k(\boldsymbol{f}_k(\boldsymbol{x} - \kappa\tfrac{\boldsymbol{V}_k}{V}\Delta x)) = \sum_{k=1}^{q}\sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} s_k(\boldsymbol{f}_k(\boldsymbol{x} - \kappa\tfrac{\boldsymbol{V}_k}{V}\Delta x))\\
&\qquad\qquad = \sum_{k=1}^{q}\sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} s_k(\boldsymbol{f}_k(\boldsymbol{x})) = \sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} \Sigma(\boldsymbol{f}_1(\boldsymbol{x}),\dots,\boldsymbol{f}_q(\boldsymbol{x})).
\end{aligned}$$

This is generally not true for the relaxation $\mathsf{R}_{\omega=2}$ that we have employed so far. There is an exception to this: the relaxation phase $\mathsf{R}_{\omega=2}$ preserves the microscopic entropy when the problem is linear. For example, in the context of the linear transport equation, *cf.* Example 1, simple computations give that

$$\Sigma(\mathsf{R}_{\omega=2}(f_+,f_-)) = \Sigma(f_+,f_-) = \frac{V}{V+a}(f_+)^2 + \frac{V}{V-a}(f_-)^2.$$

This is generalized by the following result.

Proposition 2. *Let* (1) *be linear, that is of the form*

$$\partial_t \boldsymbol{u} + \sum_{j=1}^{d} \boldsymbol{A}^j \frac{\partial \boldsymbol{u}}{\partial x_j} = 0.$$

Let $\boldsymbol{P}$ be a symmetrizer, that is, a symmetric definite positive matrix such that $\boldsymbol{P}\boldsymbol{A}^j$ is symmetric for all $j \in [\![1, d]\!]$. Consider the natural quadratic entropy-entropy flux given by

$$S(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{P}\boldsymbol{u}\cdot\boldsymbol{u}, \qquad G^j(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{P}\boldsymbol{A}^j\boldsymbol{u}\cdot\boldsymbol{u}.$$

Assume that the kinetic entropies $s_1, \dots, s_q$ are such that $s_1^*, \dots, s_q^*$ are quadratic and convex in their argument, and fulfill

$$\sum_{k=1}^{q} s_k^*(\boldsymbol{p}) = S^*(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{P}^{-1}\boldsymbol{p}\cdot\boldsymbol{p}, \qquad \sum_{k=1}^{q} \boldsymbol{V}_k^j s_k^*(\boldsymbol{p}) = G^{j,\star}(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{A}^j\boldsymbol{P}^{-1}\boldsymbol{p}\cdot\boldsymbol{p}.$$

Then the relaxation phase $\mathbf{R}_{\omega=2}$ conserves the microscopic entropy:

$$\Sigma(\mathbf{R}_{\omega=2}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_q)) = \Sigma(\boldsymbol{f}_1, \dots, \boldsymbol{f}_q),$$

hence the numerical scheme $\boldsymbol{\phi}$ is entropy preserving.

Proof. The given quadratic entropy-entropy flux are natural in the sense that

$$\boldsymbol{P}\partial_t\boldsymbol{u} + \sum_{j=1}^{d}\boldsymbol{P}\boldsymbol{A}^j\partial_{x_j}\boldsymbol{u} = 0, \qquad \text{then}$$

$$\partial_t(\boldsymbol{P}\boldsymbol{u})\cdot\boldsymbol{u} + \sum_{j=1}^{d}\partial_{x_j}(\boldsymbol{P}\boldsymbol{A}^j\boldsymbol{u})\cdot\boldsymbol{u} = \partial_t\big(\tfrac{1}{2}\boldsymbol{P}\boldsymbol{u}\cdot\boldsymbol{u}\big) + \sum_{j=1}^{d}\partial_{x_j}\big(\tfrac{1}{2}\boldsymbol{P}\boldsymbol{A}^j\boldsymbol{u}\cdot\boldsymbol{u}\big) = 0.$$

The dual entropy and entropy flux can be easily calculated and give the expected constraint on the dual kinetic entropies. By assumption, the microscopic entropy $\Sigma(\boldsymbol{f}_1, \dots, \boldsymbol{f}_q) = \sum_{k=1}^{k=q} s_k(\boldsymbol{f}_k)$ is a quadratic function in each argument. Its minimum under the conservation constraint is given by the equilibria, according to (4). The relaxation $\mathbf{R}_{\omega=2}$ given by (9) is nothing but a reflection with respect to the equilibrium and because Σ and its isolines respect this symmetry, the post-relaxation distribution functions yield the same value of Σ, concluding the proof. □

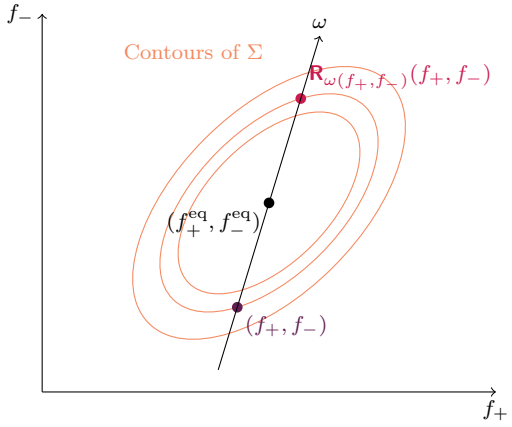

Figure 2: Idea—illustrated in the case of a two-velocities scheme—behind the procedure selecting a variable relaxation parameter $\omega = \omega(f_+, f_-)$, in order to make the pre and the post-relaxation datum lay on the same level-set of the microscopic entropy. Notice that the equilibrium is a minimizer.

However, we are interested in non-linear problems, where is it no longer true that $\mathbf{R}_{\omega=2}$ preserves the microscopic entropy. Our discussion is inspired by [31, 14, 3], who however employ a Boltzmann logarithmic entropy of the form $\sum_{k=1}^{k=q} f_k \log(f_k/\omega_k)$, where ω_k are positive weights, instead of the microscopic entropy Σ. Notice that the Boltzmann logarithmic entropy is well-defined as long as the distribution functions are positive, which is however almost never the case in practice, for they are merely "numerical" variables.

The microscopic entropy imbalance through relaxation, very similar to [3, Equation (4)], reads

$$\Delta\Sigma_{(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)}(\omega) = \Sigma(\mathbf{R}_\omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)) - \Sigma(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q).$$

At each time the relaxation operator is used, for every $\boldsymbol{x} \in \Delta x\mathbb{Z}^d$ and thus for every $\boldsymbol{f}_1(\boldsymbol{x}),\dots,\boldsymbol{f}_q(\boldsymbol{x})$, we solve the problem

$$\text{find} \quad \omega = \omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q) \quad \text{such that} \quad \Delta\Sigma_{(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)}(\omega) = 0, \tag{22}$$

and then relax using $\mathbf{R}_{\omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)}$, see Figure 2. We observe that—even this is practically what happens—there is no guarantee that (22) admits a solution. This highly depends on the structure of the underlying problem and the chosen entropies. Notice, and this is very important, that the new relaxation operator is an involution, namely

$$\mathbf{R}_{\omega(\mathbf{R}_{\omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)}(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q))}\mathbf{R}_{\omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)}(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q) = (\boldsymbol{f}_1,\dots,\boldsymbol{f}_q),$$

which can be seen by looking at Figure 2. The meaning of this formula is the following: from given distribution functions $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)$, one computes the non-linear $\omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)$ by (22), and relaxes with this rate. Then, the result of this relaxation feeds (22) once again, and a second relaxation with this new rate $\omega(\mathbf{R}_{\omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)}(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q))$ is performed. Eventually, the distribution functions go back to their original value $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)$. This guarantees that the overall scheme retainsfourth-order accuracy, because the involution property ensures that the basic brick by (12) is time-symmetric, both in the case $\omega = 2$ and when $\omega = \omega(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)$ adapted by (22).

Finally, we emphasize the fact that enforcing conservation of the microscopic entropy guarantees that the entropy S inside the domain decreases with respect to its initial value during the simulation. In particular, using (4), we have:

$$\begin{aligned}
\sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} S(\boldsymbol{u}(t,\boldsymbol{x})) &= \sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} \min_{\boldsymbol{u}(t,\boldsymbol{x})=\sum_{k=1}^{k=q}\boldsymbol{f}_k} \Sigma(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q) \\
&\le \sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} \Sigma(\boldsymbol{f}_1(t,\boldsymbol{x}),\dots,\boldsymbol{f}_q(t,\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} \Sigma(\boldsymbol{f}_1(0,\boldsymbol{x}),\dots,\boldsymbol{f}_q(0,\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} \Sigma(\boldsymbol{f}_1^{\mathrm{eq}}(\boldsymbol{u}(0,\boldsymbol{x})),\dots,\boldsymbol{f}_q^{\mathrm{eq}}(\boldsymbol{u}(0,\boldsymbol{x}))) = \sum_{\boldsymbol{x}\in\Delta x\mathbb{Z}^d} S(\boldsymbol{u}(0,\boldsymbol{x})),
\end{aligned}$$

where the last but one equality is valid upon selecting the initial datum at equilibrium, which is the most common choice.

5 Numerical experiments: Entropy conservation

The purpose of the numerical experiments in this section is two-fold. On the one hand, we would like to highlight the importance of the procedure presented in Section 4 to ensure stability. On the other hand, we want to check that the numerical scheme retains fourth-order accuracy as claimed.

5.1 Non-linear scalar problem: Burgers' equation in 1D

We start by considering the Burgers' equation. We employ the very same setting as Section 3.1, except for the choice $V = 10$ as far as the kinetic velocity is concerned. This is done in order to avoid violating the sub-characteristic condition when the first oscillations occur, which would of course drive the simulation to instability and would make the entropy correction unavailable due to the lack of convexity of the kinetic entropies. We consider a computational domain made up of 200 points with periodic boundary conditions. The result is in Figure 3, where we use a dichotomy to solve (22) at each point of the lattice. We see that the simulation which always employs $\mathbf{R}_{\omega=2}$ leads to instabilities (we stopped plotting the values when they strongly diverge), whereas the one where ω is adapted using (22) remains stable as claimed. This is due to the fact that once the shock is formed, the oscillations grow if no entropy correction is used, until some point where the sub-characteristic condition is violated, and instabilities savagely develop. Furthermore, one sees that when $\omega = 2$, the total microscopic entropy steadily increases in time, causing the eventual instability.

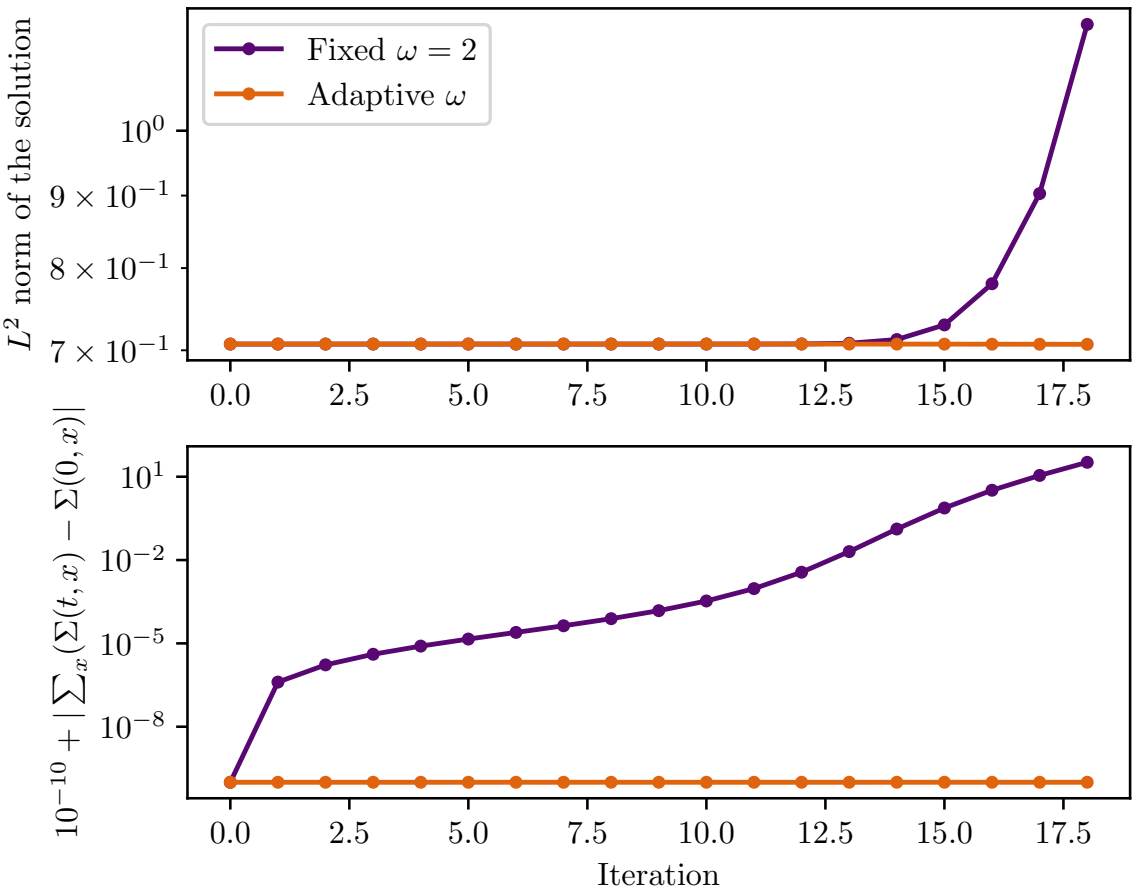

Figure 3: Norm of the solution (top) and difference between the total microscopic entropy at time zero and eventually in time (bottom), when simulating the Burgers' equation with and without entropy conservation during the relaxation. In the bottom plot, we add 10^{-10} to avoid taking the logarithm of zero.

Let us point out a practical yet fundamental point on the computation of a solution to (22). Indeed, we can have $\boldsymbol{f}_k \approx \boldsymbol{f}_k^{\mathrm{eq}}$, thus making $\Delta\Sigma_{(\boldsymbol{f}_1,\dots,\boldsymbol{f}_q)}(\omega)$ quite close to zero with almost zero derivative in ω. This frequently cause issues when iterative methods (Newton's, dichotomy, *etc.*) are employed to solve (22). The idea is to factorize the distance from the equilibrium in the problem: $\Delta\Sigma_{(f_+,f_-)}(\omega) = (f_+ - f_+^{\mathrm{eq}})\Delta\tilde{\Sigma}_{(f_+,f_-)}(\omega) = 0$, so that one eventually solves $\Delta\tilde{\Sigma}_{(f_+,f_-)}(\omega) = 0$.

Table 4: Errors and order of convergence in the L^2 norm for the Burgers' equation using (22).

Δx	L^2 error	Order
2.000E-03	3.725E-06	
1.250E-03	1.764E-06	1.59
7.813E-04	1.965E-07	4.67
4.883E-04	3.826E-08	3.48
3.053E-04	5.898E-09	3.98
1.908E-04	9.914E-10	3.79
1.193E-04	1.390E-10	4.18
7.454E-05	2.410E-11	3.73

Finally, we check that the entropy conservation procedure (22) does not change the fourth-order convergence of the method. We operate in the very same setting as in Section 3.1, also re-establishing $V = 1.2$. The results in Table 4 confirm that no order reduction is experienced and the scheme retains fourth-order accuracy, as claimed.

5.2 Non-linear system: Shallow water system in 1D

For testing the entropy correction on the shallow water system with gravity $g = 1$, we take the initial datum

$$(h,u)(t=0,x) = \begin{cases} (2,0), & x < 1/2, \\ (3/2,0), & x \geq 1/2, \end{cases}$$

and the kinetic velocity $V = 6$ with a spatial grid made up of 100 points. The results are given in Figure 4, where problem (22) has been solved using a *quasi*-Newton's method. This confirms the stabilizing power of the entropy conservation procedure and highlights—once again—that the growth of the total microscopic entropy makes solutions eventually diverge in time.

6 Variations on the numerical scheme and additional numerical experiments

We now propose variations on the basic fourth-order numerical scheme that we have proposed—whose interest is justified—and additional numerical experiments.

6.1 Projections on the equilibrium

We consider different schemes—where we introduce projections on the equilibrium at different stages. This way of proceeding can be used to reduce oscillations and enhance stability when shocks form. Somehow, these projections can help the numerical scheme to decrease entropy. Before proceeding, notice that $\mathbf{R}_{\omega=1}$ is the projection on the equilibrium. The name "projection" perfectly fits for $\mathbf{R}_{\omega=1}\mathbf{R}_{\omega=1} = \mathbf{R}_{\omega=1}$. We consider

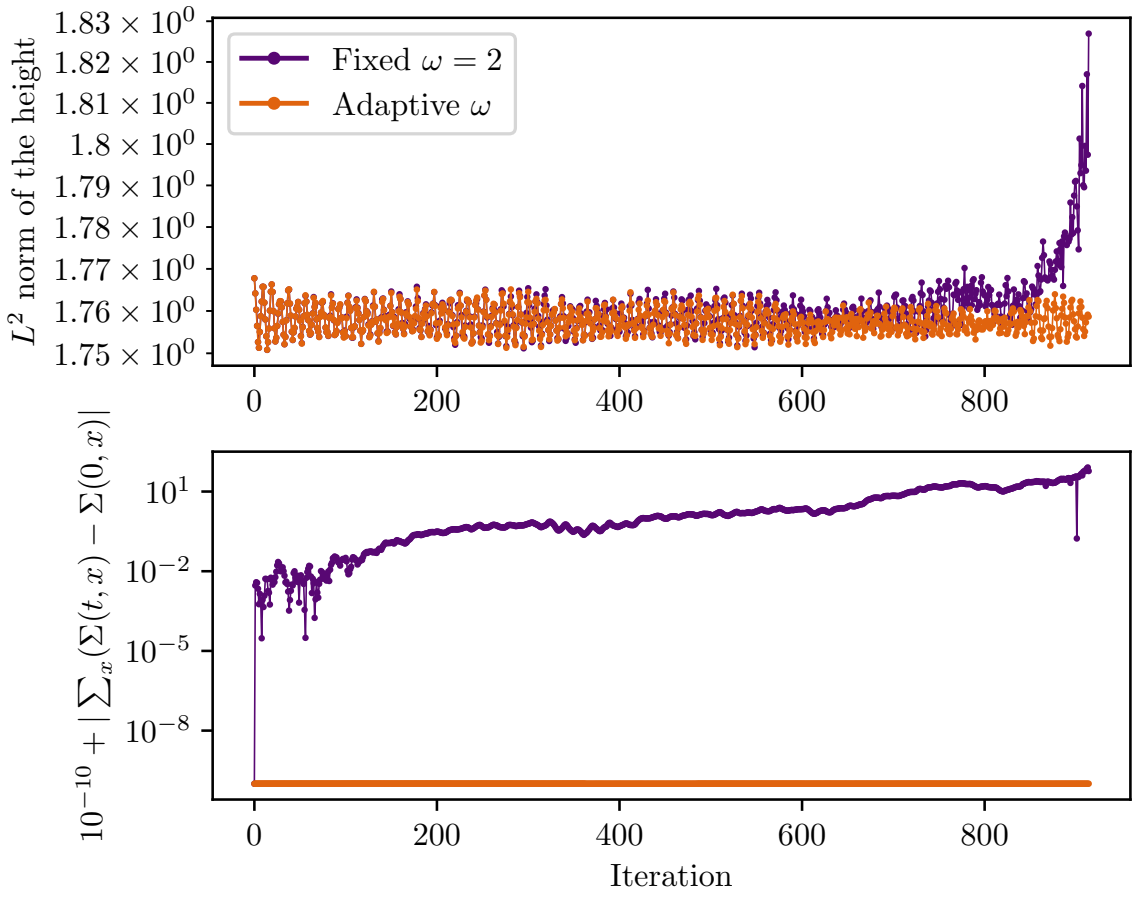

Figure 4: Norm of the solution (height, top) and difference between the total microscopic entropy at time zero and eventually in time (bottom), when simulating the shallow water equations with and without entropy conservation during the relaxation. In the bottom plot, we add 10^{-10} to avoid taking the logarithm of zero.

$$\text{Scheme (I)} \quad \begin{cases} \boldsymbol{\phi}(\Delta t) \text{ by (16)}, \\ \boldsymbol{\psi}(\Delta t) \text{ by (12)}. \end{cases}$$

$$\text{Scheme (II)} \quad \begin{cases} \boldsymbol{\phi}(\Delta t) = \mathbf{R}_{\omega=1}\boldsymbol{\psi}\big(\tfrac{\Delta t}{6}\big)^4\boldsymbol{\psi}\big(-\tfrac{\Delta t}{3}\big)\boldsymbol{\psi}\big(\tfrac{\Delta t}{6}\big)^4, \\ \boldsymbol{\psi}(\Delta t) \text{ by (12)}. \end{cases}$$

$$\text{Scheme (III)} \quad \begin{cases} \boldsymbol{\phi}(\Delta t) = \big(\mathbf{R}_{\omega=1}\boldsymbol{\psi}\big(\tfrac{\Delta t}{6}\big)\big)^4\mathbf{R}_{\omega=1}\boldsymbol{\psi}\big(-\tfrac{\Delta t}{3}\big)\big(\mathbf{R}_{\omega=1}\boldsymbol{\psi}\big(\tfrac{\Delta t}{6}\big)\big)^4, \\ \boldsymbol{\psi}(\Delta t) \text{ by (12)}. \end{cases}$$

$$\text{Scheme (IV)} \quad \begin{cases} \boldsymbol{\phi}(\Delta t) \text{ by (16)}, \\ \boldsymbol{\psi}(\Delta t) = \mathbf{R}_{\omega=1}\mathbf{T}\big(\tfrac{\Delta t}{4}\big)\mathbf{R}_{\omega=2}\mathbf{T}\big(\tfrac{\Delta t}{4}\big)\mathbf{R}_{\omega=1}\mathbf{T}\big(\tfrac{\Delta t}{4}\big)\mathbf{R}_{\omega=2}\mathbf{T}\big(\tfrac{\Delta t}{4}\big). \end{cases}$$

Let us briefly comment on these schemes. The first one is the original fourth order scheme we have proposed. For Scheme (I), (II), and (III), the basic brick $\boldsymbol{\psi}$ is left unchanged. Scheme (II) just performs a projection on the equilibrium at the end of each fourth-order solver. Scheme (III) does so after each employ of the basic brick $\boldsymbol{\psi}$. Finally, Scheme (IV) acts in a radically different fashion, for it combines a modified basic brick $\boldsymbol{\psi}$ with the same composition. The basic brick $\boldsymbol{\psi}$ is modified as a pairing of two Strang formulæ followed by projections on the equilibrium.

6.2 Non-linear scalar problem: Burgers' equation in 1D

Table 5: Errors and order of convergence in the L^2 norm for the Burgers' equation using different variation on our new fourth-order scheme.

	Scheme (I)		Scheme (II)		Scheme (III)		Scheme (IV)	
Δx	L^2 error	Order	L^2 error	Order	L^2 error	Order	L^2 error	Order
Initial datum at equilibrium								
2.000E-03	3.370E-06		3.374E-06		3.246E-06		1.147E-05	
1.250E-03	1.552E-06	1.65	1.551E-06	1.65	1.476E-06	1.68	5.686E-06	1.49
7.813E-04	1.742E-07	4.65	1.742E-07	4.65	1.677E-07	4.63	1.193E-06	3.32
4.883E-04	3.365E-08	3.50	3.365E-08	3.50	3.275E-08	3.47	3.471E-07	2.63
3.053E-04	5.184E-09	3.98	5.184E-09	3.98	5.091E-09	3.96	8.691E-08	2.95
1.908E-04	8.688E-10	3.80	8.688E-10	3.80	8.586E-10	3.79	2.283E-08	2.84
1.193E-04	1.221E-10	4.18	1.221E-10	4.18	1.212E-10	4.17	5.327E-09	3.10
7.454E-05	2.109E-11	3.74	2.109E-11	3.74	2.099E-11	3.73	1.418E-09	2.82
Initial datum off-equilibrium								
2.000E-03	6.432E-06		5.355E-06		4.115E-05		1.653E-04	
1.250E-03	1.800E-06	2.71	1.639E-06	2.52	1.069E-05	2.87	6.785E-05	1.90
7.813E-04	2.182E-07	4.49	1.825E-07	4.67	2.576E-06	3.03	2.579E-05	2.06
4.883E-04	3.945E-08	3.64	3.425E-08	3.56	6.346E-07	2.98	1.015E-05	1.98
3.053E-04	6.061E-09	3.99	5.234E-09	4.00	1.551E-07	3.00	3.958E-06	2.00
1.908E-04	9.954E-10	3.84	8.730E-10	3.81	3.796E-08	3.00	1.551E-06	1.99
1.193E-04	1.423E-10	4.14	1.225E-10	4.18	9.245E-09	3.01	6.033E-07	2.01
7.454E-05	2.396E-11	3.79	2.112E-11	3.74	2.264E-09	2.99	2.369E-07	1.99

To start testing the four different schemes proposed hitherto, we consider exactly the same setting as Section 3.1 and we shall test both by initializing the distribution functions at equilibrium and off-equilibrium. The results are given in Table 5 and show fourth-order convergence—when the initial datum is taken at equilibrium—for all numerical schemes except the last one, where order is reduced to three because of the projection on the equilibrium in the basic brick $\boldsymbol{\psi}$. Therefore, it is not advisable to employ Scheme (IV). Trying not to initialize at equilibrium, by taking $(f_+, f_-)(t=0) = (\frac{1}{4}, \frac{3}{4})u(t=0)$, we observe fourth-order convergence for the first two schemes, third-order for the third scheme, and second-order for the last one, this phenomenon is explained in what follows. We deduce that Scheme (III) needs to be used carefully, in particular, initializing at equilibrium.

In this non-linear case, the order reduction induced by the projections on the equilibrium could be seen using the modified equations [47, 28] but this would lead to very tedious calculations. Alternatively, this phenomenon can be easily understood in the case of linear transport, where $\varphi(u) = au$, using Fourier analysis [45]. In this case, we can look at the expansions of the two roots z_1 and z_2 of $\det(z\mathbf{Id} - \hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x))$ in the limit $|\xi\Delta x| \ll 1$, to theoretically understand the different convergence rates. This

provides

Scheme (I)

$$z_1(\xi\Delta x) = e^{-ia\xi\Delta t} + \frac{ia}{622080}(24a^4 - 25a^2V^2 + V^4)(\xi\Delta t)^5 + \mathcal{O}(|\xi\Delta t|^6),$$
$$z_2(\xi\Delta x) = e^{ia\xi\Delta t} - \frac{ia}{622080}(24a^4 - 25a^2V^2 + V^4)(\xi\Delta t)^5 + \mathcal{O}(|\xi\Delta t|^6).$$

Scheme (II)

$$z_1(\xi\Delta x) = e^{-ia\xi\Delta t} + \frac{ia}{622080}(24a^4 - 25a^2V^2 + V^4)(\xi\Delta t)^5 + \mathcal{O}(|\xi\Delta t|^6),$$
$$z_2(\xi\Delta x) = 0.$$

Scheme (III)

$$z_1(\xi\Delta x) = e^{-ia\xi\Delta t} + \frac{ia}{622080}(24a^4 - 25a^2V^2 + V^4)(\xi\Delta t)^5 + \mathcal{O}(|\xi\Delta t|^6),$$
$$z_2(\xi\Delta x) = 0.$$

Scheme (IV)

$$z_1(\xi\Delta x) = e^{-ia\xi\Delta t} + \frac{a^2}{3456}(a^2 - V^2)(\xi\Delta t)^4 + \mathcal{O}(|\xi\Delta t|^5),$$
$$z_2(\xi\Delta x) = 0.$$

We see that only the first scheme allows two discrete modes in the system, because no relaxation on the equilibrium is performed. One mode is the one carrying the accurate part of the solution, whereas $z_2(\xi\Delta x)$ corresponds to a parasitic numerical mode which is globally fourth-order accurate with respect to a transport equation with opposite velocity $-a$. Moreover, as all the leading order reminders vanish whenever $a = V$ and typically increase with V, these expansions suggest that one should take $V \gtrsim a$ but as close as possible to the velocity of the fastest wave, *cf.* (17), in order to minimize the truncation errors. Notice that in the first three schemes, we could observe fifth-order results provided that $|a| = \sqrt{6}/12V < V$. Indeed, we have even more: since $z_1(\xi\Delta x) = e^{-ia\xi\Delta t} + \frac{ia}{622080}(24a^4 - 25a^2V^2 + V^4)(\xi\Delta t)^5 + \frac{a^2}{622080}(24a^4 - 25a^2V^2 + V^4)(\xi\Delta t)^6 + \mathcal{O}(|\xi\Delta t|^7)$ in the case of Schemes (I) and (II), the method can be sixth-order and this is what we observe through simulations. For Scheme (III), we have $z_1(\xi\Delta x) = e^{-ia\xi\Delta t} + \frac{ia}{622080}(24a^4 - 25a^2V^2 + V^4)(\xi\Delta t)^5 + \frac{a^2}{1244160}(63a^4 - 65a^2V^2 + 2V^4)(\xi\Delta t)^6 + \mathcal{O}(|\xi\Delta t|^7)$, hence the scheme remains only fifth-order accurate when $|a| = \sqrt{6}/12V$.

Table 6: Errors and order of convergence in the L^2 norm for the transport equation with $|a| = \sqrt{6}/12V$ and initial datum at equilibrium.

	Scheme (I)		Scheme (II)		Scheme (III)		Scheme (IV)	
Δx	L^2 error	Order	L^2 error	Order	L^2 error	Order	L^2 error	Order
5.000E-02	2.283E-02		2.283E-02		2.307E-02		1.591E-01	
3.125E-02	1.890E-03	5.30	1.890E-03	5.30	2.660E-03	4.60	5.120E-02	2.41
1.961E-02	1.239E-04	5.85	1.239E-04	5.85	2.662E-04	4.94	1.321E-02	2.91
1.235E-02	8.066E-06	5.90	8.066E-06	5.90	2.713E-05	4.94	3.394E-03	2.94
7.752E-03	5.040E-07	5.96	5.040E-07	5.96	2.685E-06	4.97	8.506E-04	2.97
4.854E-03	3.081E-08	5.97	3.081E-08	5.97	2.616E-07	4.98	2.111E-04	2.98
3.040E-03	1.855E-09	6.00	1.855E-09	6.00	2.514E-08	5.00	5.171E-05	3.00
1.901E-03	1.111E-10	6.00	1.111E-10	6.00	2.406E-09	5.00	1.265E-05	3.00

These predictions are actually met by the results of Table 6. They are obtained exactly in the same setting as for the Burgers's equation, taking a final time $T = 10$ and quite coarse meshes in order to avoid very small errors below machine epsilon in double precision, since the numerical methods are now extremely accurate. Of course, this is of limited interest since valid only in the linear setting and does not extend to the case of the Burgers' equation. However, a similar idea could be utilized in the simulation of low-Mach–flows, where the wave speed is roughly constant in the domain, in order to obtain, if not sixth-order schemes, very accurate fourth-order ones.

To understand the order reductions experienced when the initial datum is not at equilibrium, *cf.* the bottom half of Table 5, again in the linear setting, we follow the procedure by [6], which has allowed to explain the behavior of the standard second-order lattice Boltzmann scheme as far as initializations are concerned. Considering that $(f_+, f_-)(t = 0) = (\theta, 1 - \theta)u(t = 0)$ with $\theta \in \mathbb{R}$, we study the low-frequency

limit of $\hat{\mathrm{g}}(\xi\Delta x) = (\boldsymbol{e}_1^T + \boldsymbol{e}_2^T)\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)(\theta, 1-\theta)^T$, the amplification factor giving the approximation of the conserved variable u after one time step, as function of the initial datum of the Cauchy problem. We have

$$\text{Scheme (I) and (II)} \qquad \hat{\mathrm{g}}(\xi\Delta x) = e^{-ia\xi\Delta t} + \mathcal{O}(|\xi\Delta t|^5),$$

independently of θ, which explains why fourth-order is indeed kept. For the other schemes

$$\text{Scheme (III)} \qquad \hat{\mathrm{g}}(\xi\Delta x) = e^{-ia\xi\Delta t} - \frac{ia^2}{1728}(a + V - 2V\theta)(\xi\Delta t)^3 + \mathcal{O}(|\xi\Delta t|^4),$$

hence we understand why we observe third-order convergence except when the initial datum is at equilibrium, that is $\theta = \frac{1}{2}(1 + \frac{a}{V})$. Finally, we have

$$\text{Scheme (IV)} \qquad \hat{\mathrm{g}}(\xi\Delta x) = e^{-ia\xi\Delta t} + \frac{a}{288}(a + V - 2V\theta)(\xi\Delta t)^2 + \mathcal{O}(|\xi\Delta t|^3),$$

yielding the same conclusion at second-order.

6.3 Solution of the Euler equations in 2D with Riemann problems

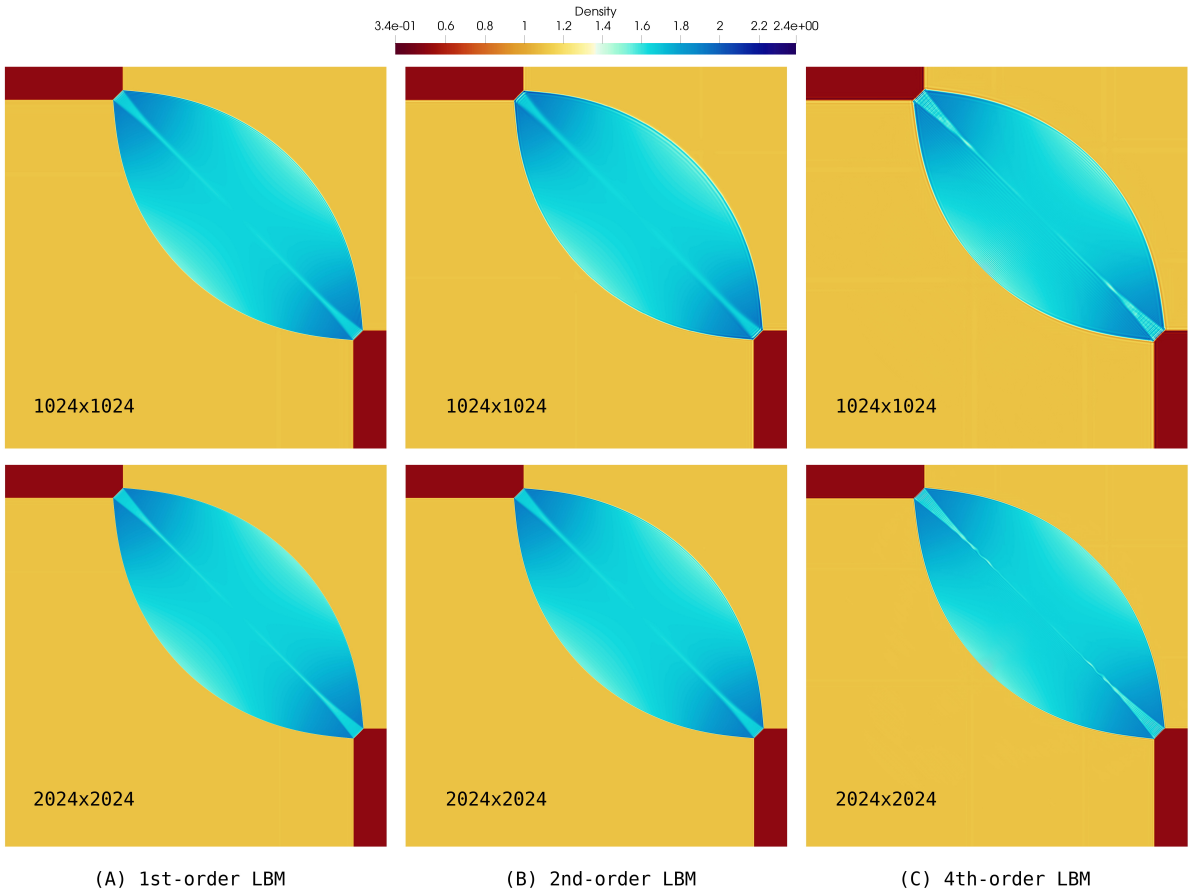

Figure 5: Densities at final time $T = 0.25$ with Riemann problems for the Euler equations, using Configuration 4 from [38] and employing schemes (A), (B), and (C).

We finish the paper on a test concerning the full Euler system in 2D with discontinuous solutions. Therefore, we have $d = 2$, $M = 4$ with $\boldsymbol{u} = (u_1, u_2, u_3, u_4) = (\rho, \rho u, \rho v, E)$ under the fluxes $\boldsymbol{\varphi}^1(\rho, u, v, p) = (\rho u, \rho u^2 + p, \rho uv, u(E+p))$ and $\boldsymbol{\varphi}^2(\rho, u, v, p) = (\rho v, \rho uv, \rho v^2 + p, v(E+p))$, where the link between energy E and pressure p is given by the polytropic equation of state

$$E = \frac{1}{2}\rho(u^2 + v^2) + \frac{p}{\gamma - 1},$$

with γ the gas constant.

We consider an initial datum made up of a Riemann problem with four constant states, given by the configuration 4 from [38], with gas constant $\gamma = 1.4$. The grids are made up of 1024 and 2048 points per direction, with a kinetic velocity $V = 6.21$. We would like to compare the second-order standard lattice Boltzmann scheme against our new fourth-order scheme. In the considered framework, the standard lattice Boltzmann method $\mathbf{R}_\omega \mathbf{T}(\Delta t)$ can only be used—for stability reasons—with $\omega = 1.93$, hence it is stuck to first-order accuracy, though with little dissipation. We utilize the scheme with $\Delta t = \Delta x/V$ and call it scheme (A). In this case, projecting on the equilibrium at each iteration would result in relaxing with $\omega = 1$, which gives extremely diffusive schemes. We therefore do not make this choice. To build a second-order stable scheme, which is however not the standard lattice Boltzmann scheme, we consider $\mathbf{R}_{\omega=1}\mathbf{T}\left(\frac{\Delta t}{4}\right)\mathbf{R}_{\omega=2}\mathbf{T}\left(\frac{\Delta t}{2}\right)\mathbf{R}_{\omega=2}\mathbf{T}\left(\frac{\Delta t}{4}\right)$, which boils down to perform one time-step by applying the basic brick (12) plus a relaxation on the equilibrium. We select $\Delta t = 4\Delta x/V$ and call the scheme (B). With our fourth-order scheme, we are able to use $\omega = 2$, hence making the scheme fourth-order accurate, provided that we employ Scheme (III). Indeed, the projection on the equilibrium at each call of the basic brick $\boldsymbol{\psi}$ has a very positive effect on stability. Notice that one could add shock detection algorithms in order to set $\omega = 2$ far from shocks and ω slightly smaller than two, ensuring the elimination of spurious oscillations. However, as claimed at the very beginning of the manuscript, this is beyond the proof-of-concept status of the paper and shall not be investigated here. The simulations are not stable when using Scheme (I) and $\omega = 2$.

After $T = 0.25$, the density field is plotted in Figure 5. The results are in agreement with those of [38]. We see that, surprisingly, at the fixed grid sizes that we consider, the lower the order of the scheme, the sharper the shocks appear on the picture, especially at coarse resolution. This is due to the (slightly) dissipative character of the scheme of the first-order scheme (A), contrarily to the dispersive nature of the second (B) and fourth-order (C) schemes, which generate visible wiggles around the shocks. However, this is not very important since, in the case of scalar 1D conservation laws solved with monotone (thus first-order accurate) finite difference schemes, the rate of convergence for solutions with shocks is $\mathcal{O}(\sqrt{\Delta x})$ in the L^1 norm, see [35, 46, 44]. When the problem is linear, for the same norm, we can expect rates $\mathcal{O}(\Delta x^{2/3})$ for second-order schemes and $\mathcal{O}(\Delta x^{4/5})$ for fourth-order ones, see [13]. Therefore, we cannot hope that either the standard lattice Boltzmann scheme or the 2nd-order scheme beats our fourth-order scheme. More interestingly, our fourth-order scheme allows observing hydrodynamic instabilities in the smooth area of the flow, along the central axis of the almond-shaped structure. This indicates that the underlying numerical scheme is a high-order one, whereas these structures cannot be observed with the standard lattice Boltzmann method and the second-order scheme at the given resolutions. In terms of computational time on GPUs, the standard lattice Boltzmann algorithm—scheme (A)—took 7.213 seconds to run with 1024 points per direction, and 59.697 seconds with 2048. Scheme (B) took 6.072 seconds with 1024 and 53.270 seconds with 2048 points. The fourth-order method—scheme (C)—took 9.470 seconds with 1024 and 64.529 seconds with 2048 points.

7 Conclusions

In this study, we have introduced a general framework for constructing fourth-order lattice Boltzmann schemes tailored to handle hyperbolic systems of conservation laws as long as their solution remains smooth. Our procedure relies on time-symmetric operators, combined together to increase the order of the method. For we employ a kinetic relaxation approximation, we can adjust the kinetic velocities to ensure that the resulting scheme adheres to the lattice Boltzmann method principles. Numerical simulations have been conducted to validate our theoretical findings. Furthermore, we have proposed modifications to the local relaxation phase to maintain entropy stability without compromising the order of the method.

Future research directions include the development of limitation strategies—both *a priori* and *a posteriori*—for these lattice Boltzmann schemes to effectively address numerical oscillations arising from shocks. Additionally, techniques to ensure positivity, particularly when dealing with shallow water equations, will be of interest. Finally, exploring methods to further increase the order of the scheme, potentially up to six or beyond, holds promise for enhancing accuracy and computational efficiency.

Acknowledgments

This work of the Interdisciplinary Thematic Institute IRMIA++, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm, was supported by IdEx Unistra (ANR-10-IDEX-0002), and by SFRI-STRAT'US project (ANR-20-SFRI-0012) under the framework of the French Investments for the Future Program.

The authors thank the two anonymous referees for their useful suggestions to improve the work. T. Bellotti also thanks O. A. Boolakee from ETH Zürich for having pointed out a typo in the first version of the manuscript.

References

[1] R. Abgrall and F. Nassajian Mojarrad, *An arbitrarily high order and asymptotic preserving kinetic scheme in compressible fluid dynamic*, Communications on Applied Mathematics and Computation, (2023), pp. 1–29.

[2] D. Aregba-Driollet and R. Natalini, *Discrete kinetic schemes for multidimensional systems of conservation laws*, SIAM Journal on Numerical Analysis, 37 (2000), pp. 1973–2004.

[3] M. Atif, P. K. Kolluru, C. Thantanapally, and S. Ansumali, *Essentially entropic lattice Boltzmann model*, Physical Review Letters, 119 (2017), p. 240602.

[4] H. Baty, F. Drui, P. Helluy, E. Franck, C. Klingenberg, and L. Thanhäuser, *A robust and efficient solver based on kinetic schemes for Magnetohydrodynamics (MHD) equations.*, Applied Mathematics and Computation, 440 (2023), p. 127667, `https://doi.org/10.1016/j.amc.2022.127667`, `https://hal.science/hal-02965967`.

[5] T. Bellotti, *Truncation errors and modified equations for the lattice Boltzmann method via the corresponding Finite Difference schemes*, ESAIM: Mathematical Modelling and Numerical Analysis, 57 (2023), pp. 1225–1255.

[6] T. Bellotti, *Initialisation from lattice Boltzmann to multi-step Finite Difference methods: modified equations and discrete observability*, Journal of Computational Physics, 504 (2024), p. 112871.

[7] T. Bellotti, *The influence of parasitic modes on stable lattice Boltzmann schemes and weakly unstable multi-step Finite Difference schemes.* working paper or preprint, Apr. 2024, `https://hal.science/hal-04358349`.

[8] T. Bellotti, L. Gouarin, B. Graille, and M. Massot, *Multidimensional fully adaptive lattice Boltzmann methods with error control based on multiresolution analysis*, Journal of Computational Physics, 471 (2022), p. 111670.

[9] T. Bellotti, L. Gouarin, B. Graille, and M. Massot, *Multiresolution-based mesh adaptation and error control for lattice Boltzmann methods with applications to hyperbolic conservation laws*, SIAM Journal on Scientific Computing, 44 (2022), pp. A2599–A2627.

[10] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation laws*, Journal of Statistical Physics, 95 (1999), pp. 113–170.

[11] F. Bouchut, *Entropy satisfying flux vector splittings and kinetic BGK models*, Numerische Mathematik, 94 (2003), pp. 623–672.

[12] F. Bouchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced schemes for Sources*, Springer Science & Business Media, 2004.

[13] P. Brenner, V. Thomée, and L. B. Wahlbin, *Besov spaces and applications to difference methods for initial value problems*, vol. 434, Springer, 2006.

[14] R. Brownlee, A. N. Gorban, and J. Levesley, *Stability and stabilization of the lattice Boltzmann method*, Physical Review E, 75 (2007), p. 036711.

[15] Y. Chen, Z. Chai, and B. Shi, *Fourth-order multiple-relaxation-time lattice Boltzmann model and equivalent finite-difference scheme for one-dimensional convection-diffusion equations*, Physical Review E, 107 (2023), p. 055305.

[16] Y. Chen, Z. Chai, and B. Shi, *A general fourth-order mesoscopic multiple-relaxation-time lattice Boltzmann model and its macroscopic finite-difference scheme for two-dimensional diffusion equations*, Journal of Computational Physics, 509 (2024), p. 113045.

[17] Y. Chen, X. Liu, Z. Chai, and B. Shi, *A Cole-Hopf transformation based fourth-order multiple-relaxation-time lattice Boltzmann model for the coupled Burgers' equations*, arXiv preprint arXiv:2309.02825, (2023).

[18] S. Clain, S. Diot, and R. Loubère, *A high-order finite volume method for systems of conservation laws–Multi-dimensional Optimal Order Detection (MOOD)*, Journal of Computational Physics, 230 (2011), pp. 4028–4050.

[19] D. Coulette, E. Franck, P. Helluy, M. Mehrenberger, and L. Navoret, *High-order implicit palindromic discontinuous Galerkin method for kinetic-relaxation approximation*, Computers & Fluids, 190 (2019), pp. 485–502.

[20] P. J. Dellar, *An interpretation and derivation of the lattice Boltzmann method using Strang splitting*, Computers & Mathematics with Applications, 65 (2013), pp. 129–141.

[21] F. Drui, E. Franck, P. Helluy, and L. Navoret, *An analysis of over-relaxation in a kinetic approximation of systems of conservation laws*, Comptes Rendus Mécanique, 347 (2019), pp. 259–269.

[22] F. Dubois, *Stable lattice Boltzmann schemes with a dual entropy approach for monodimensional nonlinear waves*, Computers & Mathematics with Applications, 65 (2013), pp. 142–159.

[23] F. Dubois, *Simulation of strong nonlinear waves with vectorial lattice Boltzmann schemes*, International Journal of Modern Physics C, 25 (2014), p. 1441014.

[24] F. Dubois, *Nonlinear fourth order Taylor expansion of lattice Boltzmann schemes*, Asymptotic Analysis, 127 (2022), pp. 297–337.

[25] T. Février, *Extension et analyse des schémas de Boltzmann sur réseau : les schémas à vitesse relative*, theses, Université Paris Sud - Paris XI, Dec. 2014, `https://theses.hal.science/tel-01126994`.

[26] E. Godlewski and P.-A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws*, vol. 118, Springer Science & Business Media, 2013.

[27] B. Graille, *Approximation of mono-dimensional hyperbolic systems: A lattice Boltzmann scheme as a relaxation method*, Journal of Computational Physics, 266 (2014), pp. 74–88.

[28] K. Guillon, R. Hélie, and P. Helluy, *Stability analysis of the vectorial lattice-Boltzmann method*. working paper or preprint, Feb. 2023, `https://hal.science/hal-03986533`.

[29] B. C. Hall, *Lie groups, Lie algebras, and representations*, Springer, 2013.

[30] F. J. Higuera and J. Jiménez, *Boltzmann approach to lattice gas simulations*, Europhysics Letters, 9 (1989), p. 663.

[31] S. A. Hosseini, M. Atif, S. Ansumali, and I. V. Karlin, *Entropic lattice Boltzmann methods: A review*, Computers & Fluids, (2023), p. 105884.

[32] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Communications on Pure and Applied Mathematics, 48 (1995), pp. 235–276.

[33] M. Junk and W.-A. Yong, *Weighted L^2-Stability of the Lattice Boltzmann Method*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 1651–1665.

[34] K. Kozhanova, R. Loubère, P. Boivin, and S. Zhao, *A hybrid* a posteriori *MOOD limited Lattice Boltzmann method to solve compressible fluid flows – LBMOOD*, Available at SSRN 4755400, (2024).

[35] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation*, USSR Computational Mathematics and Mathematical Physics, 16 (1976), pp. 105–119.

[36] P. Lafitte, W. Melis, and G. Samaey, *A high-order relaxation method with projective integration for solving nonlinear systems of hyperbolic conservation laws*, Journal of Computational Physics, 340 (2017), pp. 1–25.

[37] P. Lallemand and L.-S. Luo, *Theory of the lattice Boltzmann method: Dispersion, dissipation, isotropy, Galilean invariance, and stability*, Physical Review E, 61 (2000), p. 6546.

[38] P. D. Lax and X.-D. Liu, *Solution of two-dimensional Riemann problems of gas dynamics by positive schemes*, SIAM Journal on Scientific Computing, 19 (1998), pp. 319–340.

[39] R. J. LeVeque, *Finite volume methods for hyperbolic problems*, vol. 31, Cambridge university press, 2002.

[40] Y. Lin, N. Hong, B. Shi, and Z. Chai, *Multiple-relaxation-time lattice Boltzmann model-based four-level finite-difference scheme for one-dimensional diffusion equations*, Physical Review E, 104 (2021), p. 015312.

[41] R. I. McLachlan and G. R. W. Quispel, *Splitting methods*, Acta Numerica, 11 (2002), pp. 341–434.

[42] J. J. Miller, *On the location of zeros of certain classes of polynomials with applications to numerical analysis*, IMA Journal of Applied Mathematics, 8 (1971), pp. 397–406.

[43] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970, `https://doi.org/doi:10.1515/9781400873173`, `https://doi.org/10.1515/9781400873173`.

[44] F. Sabac, *The optimal convergence rate of monotone finite difference methods for hyperbolic conservation laws*, SIAM Journal on Numerical Analysis, 34 (1997), pp. 2306–2318.

[45] J. C. Strikwerda, *Finite difference schemes and partial differential equations*, SIAM, 2004.

[46] T. Tang and Z. H. Teng, *The sharpness of kuznetsov's $o(\sqrt{\Delta x})$ l^1-error estimate for monotone difference schemes*, Mathematics of Computation, 64 (1995), pp. 581–589.

[47] R. F. Warming and B. Hyett, *The modified equation approach to the stability and accuracy analysis of finite-difference methods*, Journal of Computational Physics, 14 (1974), pp. 159–179.

A Trace of the amplification matrix for the linear D_1Q_2 scheme

Introducing $C = a\Delta t/(\kappa\Delta x)$ and $\mu(\xi\Delta x) = \sin(\kappa\xi\Delta x) \in [-1, 1]$, we have

$$\begin{aligned}
\operatorname{tr}(\hat{\boldsymbol{\phi}}(\Delta t)(\xi\Delta x)) = &-\frac{\left(C^{18}-576C^{16}\right)1}{101559956668416}\mu(\xi\Delta x)^{36} + \frac{17\left(C^{18}-576C^{16}\right)}{203119913336832}\mu(\xi\Delta x)^{34} \\
&-\frac{\left(8C^{18}-4491C^{16}-67392C^{14}\right)1}{25389989167104}\mu(\xi\Delta x)^{32} + \frac{\left(35C^{18}-18414C^{16}-1005696C^{14}\right)}{50779978334208}\mu(\xi\Delta x)^{30} \\
&-\frac{\left(49C^{18}-22554C^{16}-3223152C^{14}-24634368C^{12}\right)}{50779978334208}\mu(\xi\Delta x)^{28} \\
&+\frac{\left(91C^{18}-31500C^{16}-11499408C^{14}-315767808C^{12}\right)}{101559956668416}\mu(\xi\Delta x)^{26} \\
&-\frac{\left(14C^{18}-2079C^{16}-3074112C^{14}-213077952C^{12}-1108546560C^{10}\right)}{25389989167104}\mu(\xi\Delta x)^{24} \\
&+\frac{\left(11C^{18}+2358C^{16}-3889944C^{14}-623464128C^{12}-11824496640C^{10}\right)}{50779978334208}\mu(\xi\Delta x)^{22} \\
&-\frac{\left(5C^{18}+4932C^{16}-2516832C^{14}-1053025920C^{12}-51399608832C^{10}-203166351360C^{8}\right)}{101559956668416}\mu(\xi\Delta x)^{20} \\
&+\frac{\left(C^{18}+3384C^{16}-204768C^{14}-992466432C^{12}-115473600000C^{10}-1712402104320C^{8}\right)}{203119913336832}\mu(\xi\Delta x)^{18} \\
&-\frac{\left(C^{16}+792C^{14}-494208C^{12}-160807680C^{10}-6429570048C^{8}-20316635136C^{6}\right)}{470184984576}\mu(\xi\Delta x)^{16} \\
&+\frac{\left(3C^{14}+8C^{12}-871200C^{10}-91228032C^{8}-1128701952C^{6}\right)}{8707129344}\mu(\xi\Delta x)^{14} \\
&-\frac{\left(3C^{12}-1100C^{10}-445392C^{8}-15894144C^{6}-41803776C^{4}\right)}{120932352}\mu(\xi\Delta x)^{12} \\
&+\frac{\left(11C^{10}-9720C^{8}-997920C^{6}-10450944C^{4}\right)}{20155392}\mu(\xi\Delta x)^{10} + \frac{\left(C^{8}+224C^{6}+2496C^{4}+13824C^{2}\right)}{31104}\mu(\xi\Delta x)^{8} \\
&-\frac{\left(7C^{6}-24C^{4}-2304C^{2}\right)}{2592}\mu(\xi\Delta x)^{6} + \frac{\left(C^{4}-4C^{2}\right)}{12}\mu(\xi\Delta x)^{4} - C^{2}\mu(\xi\Delta x)^{2} + 2.
\end{aligned}$$

This is an even polynomial of degree 36 to study on a bounded set $[0, 1]$. Since only even powers of C appear, we can just analyze $C > 0$. From Figure 6, we are confident that the polynomials stay in the band $[-2, 2]$ as long as $C \leq 24$, and they depart from it through a maximum forming at $\mu(\xi\Delta x) = \sqrt{2}/2$, namely for $\xi\Delta x = \pi/(4\kappa)$.

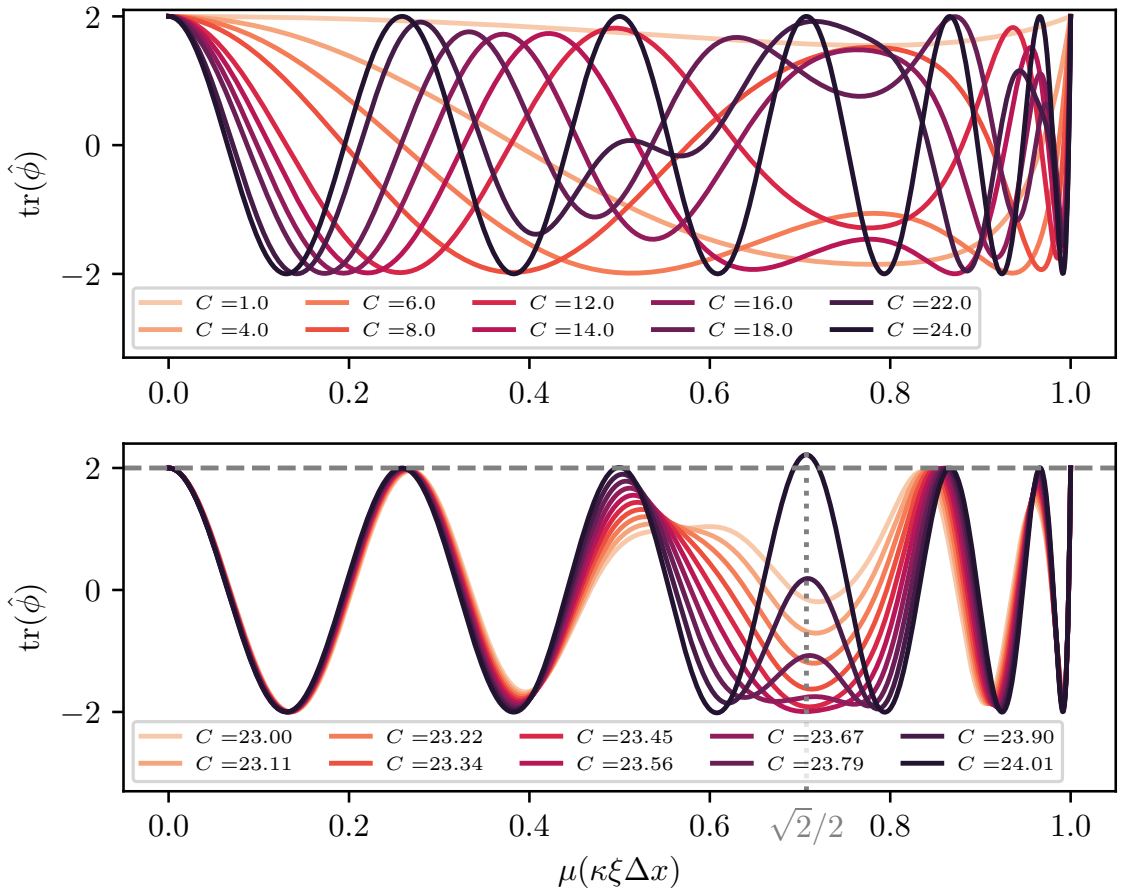

Figure 6: Trace of the amplification matrix for the linear D_1Q_2 scheme for different values of C.