Stability analysis of the vectorial lattice-Boltzmann method

Kévin Guillon, Romane Hélie, Philippe Helluy

STABILITY ANALYSIS OF THE VECTORIAL LATTICE-BOLTZMANN METHOD

KÉVIN GUILLON, ROMANE HÉLIE, PHILIPPE HELLUY

Abstract. We perform a stability analysis of the Vectorial Lattice-Boltzmann Method (VLBM). The VLBM has been introduced in [3, 1, 16, 7, 2]. It is a variant of the LBM with extended stability features: it allows handling compressible flows with shock waves, while the LBM is limited to low-Mach number regime. The stability analysis is based on the Legendre transform theory. We also propose a new tool: the equivalent system analysis, which we conjecture to contains both the stability and the consistency of the VLBM.

1. Introduction

The Vectorial Lattice-Boltzmann Method (VLBM) is a variation of the Lattice-Boltzmann Method (LBM) proposed by several authors [3, 1, 12, 16, 7, 2] for solving systems of conservation laws. Compared to the traditional LBM, which uses a scalar distribution function, the VLBM utilizes a vectorial distribution function and offers several advantages. For instance, it allows for a fully rigorous entropy stability analysis [3, 16], which ensures that the resulting schemes can compute shock waves [2] while the standard LBM is only stable on low-Mach number flows.

Both the LBM and the VLBM are based on a kinetic representation of the system of conservation laws, which consists of a set of transport equations coupled through a stiff relaxation term with a small parameter ε. Solving the transport and relaxation separately through a splitting method results in simple and efficient schemes.

However, this methodology raises several natural questions, three of which are listed below:

(1) Is the kinetic representation stable as $\varepsilon \to 0$?
(2) Is the kinetic representation consistent with the system of conservation laws as $\varepsilon \to 0$?
(3) Is the splitting scheme consistent with the kinetic representation and/or the system of conservation laws as $\varepsilon \to 0$?

The first question can be answered by an entropy analysis, which has been explored in prior work [3, 16]. However, the fully rigorous theory has not yet been achieved, especially for general systems of conservation laws in high-dimensional spaces. This is related to challenging questions, such as the mathematical theory of Navier-Stokes equations [29], strange behavior of the general solutions of conservation laws [10], and uncertainty quantification [19].

The second question regarding whether or not the kinetic model converges towards the approximated conservation laws can be fully answered in some specific cases. However, the general case is also difficult. There are some semi-heuristic arguments, based on the Chapman-Enskog expansion [6], that provide explanations for the convergence of the stiff relaxation approximation. Among many works in this direction, [11] presents one approach.

The third question is related to the fact that the stiff relaxation is never resolved in practice. Instead, a time splitting scheme with over-relaxation is used. In most situations, this splitting algorithm does not provide a good approximation of the relaxation model, but rather it approximates the initial conservation laws directly, which is the ultimate objective. Theoretically, this approach is justified with an equivalent equation analysis, performed directly on the time splitting scheme without the relaxation intermediary. The analysis relies on a combination of Taylor and Chapman-Enskog expansions [13] and generally provides relevant information on the consistency of the LBM or VLBM method.

However, the stability of the equivalent equation is generally not sufficient to ensure the stability of the VLBM [4]. Thus, separate studies are necessary for both the stability and the consistency of the method in practice.

Key words and phrases. vectorial lattice-Boltzmann, entropy stability, equivalent equation analysis.

In this work, we aim to provide a more comprehensive justification for the stability and consistency of the VLBM method. We begin by reviewing the entropy stability of the VLBM, which can be proven using the Legendre transform theory. We demonstrate that this approach can be naturally extended to prove the stability of the VLBM when an over-relaxed splitting scheme is used for time integration.

To analyze consistency, we propose a new algorithmic approach for constructing a system of equivalent Partial Differential Equations (PDEs) that includes not only the conservative variables but also the equilibrium deviation variables. In several examples, we show that the analysis of the equivalent system and the entropy analysis yield the same stability condition.

We then utilize an additional Chapman-Enskog expansion to remove the equilibrium deviation variables, under a smallness hypothesis. This enables us to recover the traditional equivalent equation provided by other authors [15, 20, 8].

Finally, we present numerical experiments to verify that the equivalent system with deviation variables provides more accurate information about the stability of the VLBM than the standard equivalent equation.

2. Hyperbolic conservation law and duality

2.1. **Hyperbolic systems of conservation laws.** In this work, we consider the numerical resolution of a system of conservation laws

$$\partial_t W + \sum_{i=1}^{d} \partial_i Q^i(W) = 0, \tag{2.1}$$

where the unknown is a vector of m conserved quantities $W(X,t) \in \mathbb{R}^m$, depending on a space variable $X \in \mathbb{R}^d$ and a time variable $t \in \mathbb{R}$. Let $N = (N_1, \dots, N_d) \in \mathbb{R}^d$ be a space vector. The flux of the system is defined by

$$Q(W,N) = \sum_{i=1}^{d} Q^i(W) N_i.$$

We assume that the system of conservation laws admits a Lax entropy $W \mapsto s(W)$. Therefore, there is an entropy flux $\sum g^i(W) N_i$ such that

$$\partial_t s(W) + \sum_{i=1}^{d} \partial_i g^i(W) = 0,$$

whenever W is a smooth solution of (2.1). This imposes that

$$D_W s(W) D_W Q^i(W) = D_W g^i(W), \tag{2.2}$$

where we have denoted by $D_W g(W)$ the Jacobian of $g(W)$. Let us recall that the Jacobian is the transpose of the gradient

$$D_W s(W) = \nabla_W s(W)^\intercal. \tag{2.3}$$

Thus $D_W s(W)$ is a row vector, while $\nabla_W s(W)$ is a column vector.

In addition, we assume a convexity hypothesis of s on a closed convex cone $\mathcal{K} \subset \mathbb{R}^m$:

$$s \text{ is strictly convex on } \mathcal{K}.$$

Finally, following a standard convex analysis convention, we extend s by an infinite value outside $\mathcal{K}$:

$$s(W) = +\infty, \quad W \in \complement\mathcal{K}. \tag{2.4}$$

Remark 1. For working in a fully practical framework, it is indeed necessary to consider cases where $\mathcal{K} \neq \mathbb{R}^d$. However this leads to mathematical questions that we have not yet investigated. Therefore, in the following, the reader can suppose that $\mathcal{K} = \mathbb{R}^d$.

2.2. **Entropy and symmetrization.** According to the Mock theorem [26, 27], the system (2.1) is symmetrizable and thus hyperbolic: for all $W \in \mathcal{K}$ and all $N \in \mathbb{R}^d$ the Jacobian of the flux

$$A(W, N) = D_W Q(W, N)$$

is diagonalizable with real eigenvalues.

Let us recall how to prove this result, because it will be useful. For this we introduce the conjugate of the entropy [23, 24] defined by

$$s^*(W^*) = \max_W (W^* \cdot W - s(W)), \tag{2.5}$$

where we denote by $\cdot$ the usual dot product:

$$W^* \cdot W = W^{*\intercal} W = W^\intercal W^*.$$

Thanks to (2.4) we can also write

$$s^*(W^*) = \max_{W \in \mathcal{K}} (W^* \cdot W - s(W)), \tag{2.6}$$

The components of W^* are called the dual variables, or entropy variables [21, 5, 9]. The function s^* is called the conjugate or the dual entropy. The definition of the conjugate (2.5) applies to functions that are not necessarily smooth or convex. In the regular case, when s is smooth and strictly convex on $\mathbb{R}^m$, for instance, s^* is defined implicitly by the following two relations

$$W^* = \nabla s(W(W^*)), \tag{2.7}$$

$$s^*(W^*) = W^* \cdot W(W^*) - s(W(W^*)). \tag{2.8}$$

The formula (2.7) determines in a unique way the map between the conservative variables W and the dual variables W^*. In the convex case, it can also be shown that $s^{**} = s$. Thus we have the reverse relations:

$$W = \nabla s^*(W^*(W)),$$
$$s(W) = W^*(W) \cdot W - s^*(W^*(W)).$$

We can also define the dual entropy flux by the relation

$$P^{i,\star}(W^*) = W^* \cdot Q^i(W(W^*)) - g^i(W(W^*)). \tag{2.9}$$

We do not use exactly the same symbol for the dual entropy $(*)$ and the dual flux $(\star)$, because the definitions are slightly different. An important fact is that the knowledge of $s^*(W^*)$ and $P^{i,\star}(W^*)$ is sufficient to reconstruct the system of conservation laws (2.1). Indeed,

$$\begin{aligned}\nabla_{W^*} P^{i,\star}(W^*) &= Q^i(W(W^*)) + W^* \cdot D_W Q^i(W(W^*)) D_{W^*} W(W^*) - D_W g^i D_{W^*} W(W^*),\\ &= Q^i(W(W^*)) + D_W s(W) D_W Q^i(W(W^*)) D_{W^*} W(W^*) - D_W g^i D_{W^*} W(W^*),\end{aligned}$$

(because of (2.3) and (2.2))

$$= Q^i(W(W^*)) + D_W g^i(W(W^*)) D_{W^*} W(W^*) - D_W g^i(W(W^*)) D_{W^*} W(W^*),$$

(because of (2.7)), and thus

$$\nabla_{W^*} P^{i,\star}(W^*) = Q^i(W(W^*)). \tag{2.10}$$

In short: the gradient of the dual entropy gives the conservative variables and the gradient of the dual flux gives the flux.

Theorem 2. *The change of variables $W^* \mapsto W(W^*)$ symmetrizes the system of conservation laws (2.1).*

Proof. Indeed

$$\partial_t W + \sum_i \partial_i Q^i(W) = 0,$$

can be rewritten

$$\partial_t \nabla_{W^*} s^*(W^*) + \sum_i \partial_i \nabla_{W^*} P^{i,\star}(W^*) = 0,$$

thanks to the above remark (2.7) and (2.10), or

$$D^2_{W^*W^*}s^*(W^*)\partial_t W^* + \sum_i D^2_{W^*W^*}P^{i,\star}(W^*)\partial_i W^* = 0, \tag{2.11}$$

where the Hessian matrices $D^2_{W^*W^*}s^*(W^*)$ and $D^2_{W^*W^*}P^{i,\star}(W^*)$ are obviously symmetric and $D^2_{W^*W^*}s^*(W^*)$ is positive definite when s^* is strictly convex. □

We have just proven the Mock theorem [27]. In the following sections we shall often try to guess directly a symmetrization of the system of conservation laws, rather than the full dual entropy theory to obtain the stability conditions of the relaxation scheme.

Remark 3. In the following, the practical stability will derive from a convexity condition imposed on s^*. This may sound strange, because the conjugate of any function is always convex, according to the standard definition (2.5). We need to be a little bit more subtle. Actually, it is possible to define two different duality transformations. Let us denote by Fenchel transform the s^* given by the first formula (2.5). The Fenchel transform derives from an optimization problem. Let us denote by Legendre transformation the s^* given by formula (2.7) and (2.8). The Legendre transform is thus defined by purely algebraic relations. It is unambiguously defined, as soon as $W \mapsto W^* = \nabla_W s(W)$ is an invertible map. In this algebraic definition, s^* is not necessarily convex, but the convexity of s^* is equivalent to the convexity of s. The Legendre and Fenchel transform are different in the general case. They coincide in the convex case.

Remark 4. In the most general case, the Legendre transform is a multivalued function. It has to be defined from differential geometry tools. The interested reader can refer to [17] for an introduction to this topic.

3. Vectorial kinetic model

3.1. **Direct construction.** In this section, we recall the entropy theory of the kinetic representation. This theory has a long history, see for instance [25, 27, 21, 14, 5, 9, 28]. In our context it has been analyzed by Bouchut in [3]. However, in his work, Bouchut avoids the use of the Legendre transform. The ideas have been rephrased in an easier way (to our opinion) in the work of Dubois [16] with the help of the Legendre transform. Let us now recall the theory.

We consider a formal vectorial kinetic representation of system (2.1)

$$\partial_t F_k + V_k \cdot \nabla F_k = \frac{1}{\varepsilon}(F_k^{\mathrm{eq}} - F_k), \quad k = 1 \dots n_v. \tag{3.1}$$

The approximate conservative vector is the sum of the kinetic vectors

$$W = \sum_{k=1}^{n_v} F_k, \tag{3.2}$$

and the kinetic equilibrium vectors (or Maxwellians) are functions of the conservative data

$$F_k^{\mathrm{eq}} = F_k^{\mathrm{eq}}(W). \tag{3.3}$$

The kinetic velocities V_k are n_v constant and given vectors of $\mathbb{R}^d$. In practice, it is interesting to introduce a positive parameter λ, whose purpose is to change the size of the kinetic velocities. So we shall often take

$$V_k = \lambda \tilde{V}_k, \tag{3.4}$$

where the directions $\tilde{V}_k$ of the kinetic velocities are fixed. In this way, the kinetic velocities can be dilated.

For practical reasons we also introduce the kinetic vectors F and F^{eq}, which are column vectors made of all the stacked kinetic data:

$$F = (F_1^\intercal, \dots, F_{n_v}^\intercal)^\intercal, \quad F^{eq} = \left((F_1^{eq})^\intercal, \dots, \left(F_{n_v}^{eq}\right)^\intercal\right)^\intercal.$$

With the help of the following diagonal matrices (1_m is the identity matrix of size $m \times m$)

$$V^i = \begin{pmatrix} V_1^i 1_m & & \\ & \ddots & \\ & & V_{n_v}^i 1_m \end{pmatrix},$$

the kinetic system can also be written in the compact form

$$\partial_t F + \sum_{i=1}^{d} \partial_i \left(V^i F \right) = \frac{1}{\varepsilon}(F^{\mathrm{eq}} - F).$$

When $\varepsilon \to 0$, we expect that

$$F_k \simeq F_k^{eq}. \tag{3.5}$$

If we assume that

$$W = \sum_{k=1}^{n_v} F_k^{\mathrm{eq}}(W), \quad Q^i(W) = \sum_{k=1}^{n_v} V_k^i F_k^{\mathrm{eq}}(W), \tag{3.6}$$

then, summing (3.1) on k and using (3.5) we formally obtain

$$\partial_t W + \sum_{i=1}^{d} \partial_i Q^i(W) \simeq 0,$$

and we have obtained an approximation of the initial system of conservation laws.

Remark 5. For being more rigorous, we should have made more explicit in (3.1) and (3.2) the dependency of F_k and $W = \sum_k F_k$ with respect to ε. For instance, by using the notations $F_k^{(\varepsilon)}$ and $W^{(\varepsilon)}$. But we prefer to lighten the notations, and in the following the dependency with ε will be implicit. This means that from now on, $W = \sum_k F_k$ is not an exact solution of (2.1) but an approximate one.

For the moment, instead of relating the equilibrium (3.3) to the conservation laws (2.1), we assume that the equilibrium is obtained from an entropy optimization principle. For this we introduce a microscopic entropy

$$\sigma(F) = \sum_{k=1}^{n_v} s_k(F_k),$$

where the kinetic entropies s_k are strictly convex functions of the F_k on $\mathcal{K}$. The macroscopic entropy is obtained from the resolution of the following constrained optimization problem

$$s(W) = \min_{W = \sum_k F_k} \sigma(F). \tag{3.7}$$

In optimization theory, this operation is known as an inf-convolution operation [24]. The macroscopic entropy is the inf-convolution of the kinetic entropies. In many works, the inf-convolution operator is denoted with a $\square$. We thus have:

$$s = s_1 \square s_2 \ldots \square s_{n_v}.$$

We denote by $F_k^{\mathrm{eq}}(W)$ the (supposed to be unique) values of F_k that achieve the minimum

$$s(W) = \sum_{k=1}^{n_v} s_k(F_k^{\mathrm{eq}}(W)).$$

If we introduce the Lagrangian

$$L(F, \Lambda) = \sum_{k=1}^{n_v} s_k(F_k) + \Lambda \cdot \left(W - \sum_{k=1}^{n_v} F_k \right).$$

The minimizer $F^{\mathrm{eq}}(W)$ and the Lagrange multiplier $\Lambda(W)$ are characterized by

$$\nabla_{F_k} s_k(F_k^{\mathrm{eq}}) = \Lambda, \quad \sum_{k=1}^{n_v} F_k^{\mathrm{eq}} = W. \tag{3.8}$$

These relations are simply obtained by deriving the Lagrangian with respect to F_k or Λ.

An essential property of the inf-convolution is that the Fenchel transform changes it into a sum. We thus have

$$s^*(W^*) = \sum_{k=1}^{n_v} s_k^*(W^*). \tag{3.9}$$

Taking the Legendre transform of (3.8) we see that the couple $(F_k^{\mathrm{eq}}(W), \Lambda(W))$ is solution to

$$F_k^{\mathrm{eq}} = \nabla_\Lambda s_k^*(\Lambda(W)), \quad \sum_{k=1}^{n_v} F_k^{\mathrm{eq}} = W. \tag{3.10}$$

Summing over k we also have the Lagrange multiplier in an easier way

$$\sum_{k=1}^{n_v} F_k^{\mathrm{eq}} = \sum_{k=1}^{n_v} \nabla_\Lambda s_k^*(\Lambda(W)),$$

$$W = \nabla_\Lambda \sum_{k=1}^{n_v} s_k^*(\Lambda(W)),$$

$$= \nabla_\Lambda s^*(\Lambda(W)),$$

and thus

$$\Lambda(W) = W^*(W) = \nabla s(W).$$

The Lagrange multiplier of the constrained optimization problem (3.8) is simply the gradient of the macroscopic entropy.

Let us now assume the additional property

$$P^{i,\star}(W^*) = \sum_{k=1}^{n_v} V_k^i s_k^*(W^*). \tag{3.11}$$

Then we have

$$\nabla_{W^*} P^{i,\star}(W^*) = \sum_{k=1}^{n_v} V_k^i \nabla_{W^*} s_k^*(W^*),$$

$$\nabla_{W^*} P^{i,\star}(W^*) = \sum_{k=1}^{n_v} V_k^i F_k^{\mathrm{eq}}(W),$$

(from (3.10)) and thus, from (2.10)

$$Q^i(W) = \sum_{k=1}^{n_v} V_k^i F_k^{\mathrm{eq}}(W). \tag{3.12}$$

In the flux form, we can also write

$$Q(W, N) = \sum_{k=1}^{n_v} (V_k \cdot N) F_k^{\mathrm{eq}}(W).$$

We recover the relations (3.6) that impose the consistency of the kinetic model (3.1) with the system of conservation laws (2.1). But now the consistency derives from an entropy optimization principle. We can sum up the direct construction of a kinetic model (3.1) that is entropy consistent with (2.1):

(1) Compute the Legendre transform s^* of the entropy s by (2.7), (2.8).
(2) Compute the dual fluxes $P^{i,\star}$ by (2.9).
(3) Find n_v strictly convex functions s_k^* satisfying the consistency relations written in the dual variables:

$$\sum_{k=1}^{n_v} s_k^* = s^*, \quad \sum_{k=1}^{n_v} V_k^i s_k^* = P^{i,\star}. \tag{3.13}$$

(4) The equilibrium is given by

$$F_k^{eq} = \nabla s_k^*.$$

Remark 6. If $n_v \geqslant d+1$, there is generally at least one solution to the algebraic system (3.13). The difficulty is to ensure that the s_k^* are convex. For this property to hold, the scaling (3.4) is useful. It gives an additional degree of freedom for ensuring convexity.

3.2. **Reverse construction.** Now that we have recalled the entropy theory of the kinetic representation, we can proceed in the reverse way. We *choose* the equilibrium $F^{\rm eq}$ in such a way that the consistency relation (3.6) is satisfied. From the above theory, we expect that $F_k^{\rm eq}(W(W^*))$ is a gradient, when it is expressed in the entropy variables W^*. We can thus find dual kinetic entropies $s_k^*(W^*)$ such that

$$F_k^{\rm eq}(W(W^*)) = \nabla_{W^*} s_k^*(W^*).$$

By Legendre transform, we can (in principle) compute the kinetic entropies $s_k(F_k)$ and this gives us the microscopic entropy

$$\sigma(F) = \sum_{k=1}^{n_v} s_k(F_k).$$

The main point in the reverse construction is to ensure that the strict convexity is preserved. In practice, we will see that the microscopic entropy is convex under a condition that λ is large enough. This will be the subcharacteristic condition.

4. Splitting and over-relaxation

4.1. **Splitting scheme.** In this section we recall how to solve practically in time the kinetic system (3.1). Indeed, in (3.1) all the transport equations are coupled in a non-linear way. We introduce a splitting method for separating the transport equations. In this numerical method, the approximation of the kinetic data is computed at fixed times $t_n = n\Delta t$, where Δt is the time step. The approximation is not continuous at time t_n, i.e. we distinguish between the values of $F_k(X, t_n^-)$ and of $F_k(X, t_n^+)$. Suppose that we know the kinetic data $F_k(X, t_{n-1}^+)$ at the end of time step $n-1$. Computing the next time step consists first in solving the homogenous transport equations

$$\partial_t F_k + V_k \cdot \nabla F_k = 0,$$

for a duration of Δt, which defines $F_k(X, t_n^-)$. Indeed, using the characteristic method (and avoiding for the moment difficulties arising from the boundaries), we get the explicit formula

$$F_k(X, t_n^-) = F_k(X - \Delta t V_k, t_{n-1}^+). \tag{4.1}$$

In this way, we obtain the new conservative data by

$$W(X, t_n) = \sum_{k=1}^{n_v} F_k(X, t_n^-).$$

We then apply the following over-relaxation formula

$$F_k(X, t_n^+) = \omega F_k^{eq}(W(X, t_n)) + (1-\omega) F_k(X, t_n^-). \tag{4.2}$$

Several choices can be made for the relaxation parameter ω. The choice $\omega = 1$ corresponds to a projection of the kinetic data on the equilibrium at the end of each time step. The choice $\omega = 2$ (over-relaxation) is interesting, because, in this case, it can be shown that the time integration is second-order accurate (see for instance [7] and included references).

4.2. **Operator notations.** The splitting scheme is very simple to implement in a computer program and no additional information is needed. However, we introduce here some additional mathematical notations that will be useful for deriving the equivalent equation analysis.

The interesting output of the kinetic scheme is obviously the conservative variables vector of size m, given by

$$W = \sum_{k=1}^{n_v} F_k.$$

However, the kinetic representation introduces n_v kinetic vectors F_k instead of one vector W. For the analysis, it is necessary to supplement W with additional quantities. The equivalent equation could be written in an arbitrary set of variables $Y \in \mathbb{R}^{mn_v}$. We have done a particular choice, which helps the understanding, to our opinion.

The first components of Y are made of the conservative variables W. The next components will be made of the flux errors

$$\sum_{k=1}^{n_v} V_k^i \left(F_k - F_k^{eq}\right) = \sum_{k=1}^{n_v} V_k^i F_k - Q^i, \quad 1 \leqslant i \leqslant d.$$

If $n_v = d+1$, we have enough variables. If $n_v > d+1$ we still have to supplement Y with $n_c = (n_v - d - 1)$ independent linear combinations of the kinetic data

$$\sum_{k=1}^{n_v} \beta_k^\ell \left(F_k - F_k^{eq}\right), \quad 1 \leqslant \ell \leqslant n_v - d - 1 = n_c.$$

It is convenient to choose the coefficients β_k^ℓ of the linear combinations in order to cancel the contributions of F^{eq}

$$\sum_{k=1}^{n_v} \beta_k^\ell F_k^{eq} = 0. \tag{4.3}$$

Finally, Y is of the form

$$Y = \begin{pmatrix} W \\ \sum_k V_k^1 F_k \\ \vdots \\ \sum_k V_k^d F_k \\ \sum_k \beta_k^1 F_k \\ \vdots \\ \sum_k \beta_k^{n_v-d-1} F_k \end{pmatrix} - \begin{pmatrix} 0 \\ \sum_k V_k^1 F_k^{eq} \\ \vdots \\ \sum_k V_k^d F_k^{eq} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

While the m first components of Y are an approximate solution to (2.1)

$$Y^{1\cdots m} = W,$$

we expect the next components to be small

$$Y^{m+1\cdots mn_v} \simeq 0.$$

In the following, we denote the sub-vector $Y^{m+1\cdots mn_v}$ as the *flux error* (even if the last components of Y, as we have seen above, are not necessarily all related to fluxes).

We can write the Y representation in a matrix form

$$Y(F) = MF - CMF^{eq}(W(F)), \quad W(F) = \sum_{k=1}^{n_v} F_k. \tag{4.4}$$

We call the matrix M the matrix of moments. It is supposed to be invertible. The matrix C is the diagonal matrix

$$C = \begin{pmatrix} 0_m & & \\ & 1_{md} & \\ & & 0_{mn_c} \end{pmatrix},$$

where we denote by 1_r the diagonal identity matrix of size $r \times r$ and by 0_r the null matrix of size $r \times r$. The nonlinear mapping $F \mapsto Y$ can be inverted explicitly

$$F(Y) = M^{-1}Y + M^{-1}CMF^{eq}(W(Y)), \quad W(Y) = (Y^1, \ldots, Y^m)^\intercal. \tag{4.5}$$

The notations are tedious but simple. For making them clearer, let us consider three examples that we will use below.

4.2.1. *D1Q2 case.* In this case, the space dimension $d = 1$ and we take two kinetic velocities

$$V_1 = -\lambda, \quad V_2 = \lambda.$$

Because $n_v = d+1$ and from the consistency relation (3.12), the equilibrium is necessarily given by

$$F_1^{eq}(W) = \frac{1}{2}W - \frac{1}{2\lambda}Q(W), \quad F_2^{eq}(W) = \frac{1}{2}W + \frac{1}{2\lambda}Q(W).$$

Then

$$M = \begin{pmatrix} 1_m & 1_m \\ -\lambda 1_m & \lambda 1_m \end{pmatrix}.$$

4.2.2. *D2Q3 case.* In this case, the space dimension $d = 2$ and we take $n_v = 3$ kinetic velocities

$$V_1 = \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad V_2 = \frac{\lambda}{2} \begin{pmatrix} -1 \\ \sqrt{3} \end{pmatrix}, \quad V_3 = \frac{\lambda}{2} \begin{pmatrix} -1 \\ -\sqrt{3} \end{pmatrix}.$$

Because of $n_v = d + 1$ the equilibrium is uniquely defined. It is given by

$$F^{eq} = M^{-1} \begin{pmatrix} W \\ Q^1 \\ Q^2 \end{pmatrix},$$

with

$$M = \begin{pmatrix} 1_m & 1_m & 1_m \\ \lambda 1_m & -\frac{1}{2}\lambda 1_m & -\frac{\lambda}{2} 1_m \\ 0_m & \frac{\sqrt{3}}{2}\lambda 1_m & -\frac{\sqrt{3}}{2}\lambda 1_m \end{pmatrix}.$$

We find

$$F_k^{eq} = \frac{W}{3} + \frac{2}{3\lambda^2} V_k \cdot \begin{pmatrix} Q^1 \\ Q^2 \end{pmatrix}.$$

4.2.3. *D2Q4 case.* In this case, the space dimension $d = 2$ and we take four kinetic velocities

$$V_1 = \lambda \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad V_2 = \lambda \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad V_3 = \lambda \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad V_4 = \lambda \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

Because $n_v > d + 1$, we have one degree of freedom for the computation of the equilibrium. We can take, by analogy with the D1Q2 model:

$$F_{1,2}^{eq}(W) = \frac{1}{2} W \pm \frac{1}{2\lambda} Q^1(W), \quad F_{3,4}^{eq}(W) = \frac{1}{2} W \pm \frac{1}{2\lambda} Q^2(W).$$

Then the matrix of moments is given by

$$M = \begin{pmatrix} 1_m & 1_m & 1_m & 1_m \\ \lambda 1_m & -\lambda 1_m & 0_m & 0_m \\ 0_m & 0_m & \lambda 1_m & -\lambda 1_m \\ \lambda^2 1_m & \lambda^2 1_m & -\lambda^2 1_m & -\lambda^2 1_m \end{pmatrix}.$$

For the D2Q4 model, other choices are possible for the last row of the matrix of moments. We have made the choice proposed in [18], but we could have replaced λ^2 by one, for instance. Actually, we could have supplemented by any row that ensures that M is invertible, but then the condition (4.3) is generally not satisfied.

4.2.4. *Transport operator in the Y variables.* With these notations, we can rewrite the transport operator in the Y variables. Let us define the $\tau_k(\Delta t)$ shift operators, applied on an arbitrary function $X \mapsto f(X)$ by

$$(\tau_k(\Delta t) f)(X) = f(X - \Delta t V_k).$$

The vectorial shift operator is then given by

$$\tau(\Delta t) = \begin{pmatrix} \tau_1(\Delta t) 1_m & & 0 \\ & \ddots & \\ 0 & & \tau_{n_v}(\Delta t) 1_m \end{pmatrix}.$$

The initial data $Y(\cdot, t_{n-1}^+)$ at the end of the time step $n-1$ are supposed to be known. The transport map $\mathcal{T}(\Delta t)$ is the procedure that takes $Y(\cdot, t_{n-1}^+)$ and produces $Y(\cdot, t_n^-)$ before the relaxation step:

$$Y(\cdot, t_n^-) = \mathcal{T}(\Delta t) Y(\cdot, t_{n-1}^+).$$

The procedure is as follows:

(1) Express the initial data Y in the kinetic variables thanks to (4.5)

$$F(\cdot, t_{n-1}^+) = F(Y(\cdot, t_{n-1}^+)).$$

(2) Apply the transport operator

$$F(\cdot, t_n^-) = \tau(\Delta t) F(\cdot, t_{n-1}^+).$$

(3) Go back to the Y variables thanks to (4.4)

$$Y(\cdot, t_n^-) = Y(F(\cdot, t_n^-)).$$

The transport is nonlinear functional operator, made of shift operators and nonlinear algebraic operations. We do not give all the details.

4.2.5. *Relaxation operator in the Y variables.* The relaxation operator in the kinetic variables is given by (4.2). In vectorial form it reads

$$F(\cdot, t_n^+) = \omega F^{eq}(W(\cdot, t_n^-)) + (1-\omega) F(\cdot, t_n^-). \tag{4.6}$$

Using the fact that $W = \sum_k F_k$ does not change during the relaxation, we simply obtain

$$Y(\cdot, t_n^+) = \mathcal{R}_\omega Y(\cdot, t_n^-), \tag{4.7}$$

where the relaxation operator $\mathcal{R}_\omega$ is the diagonal operator

$$\mathcal{R}_\omega = \begin{pmatrix} 1_m & 0 \\ 0 & (1-\omega) 1_{m(n_v-1)} \end{pmatrix}.$$

In other words, the relaxation operator leaves the conservative variables unchanged:

$$(\mathcal{R}_\omega Y)^i = Y^i = W^i, \quad 1 \leqslant i \leqslant m,$$

and the other components of Y are multiplied by $1-\omega$:

$$(\mathcal{R}_\omega Y)^i = (1-\omega) Y^i, \quad i > m.$$

When $\omega \simeq 2$, $1-\omega \simeq -1$ and thus, these components are numerically fluctuating around 0 at each time iteration. There is no damping of the oscillations when $\omega = 2$. The frequency of the fluctuations is $1/\Delta t$ and tends to infinity when Δt tends to zero.

4.2.6. *Formal numerical scheme.* Now that we have introduced the operator notation, we can give a more formal definition of the split kinetic approximation. For solving (2.1) with the initial data

$$W(X, 0) = W_0(X),$$

we start with

$$Y_0^+ = \begin{pmatrix} W_0(X) \\ 0 \end{pmatrix}.$$

Assume that we have reached Y_{n-1}^+ at time step $n-1$. Then the next value is given by

$$Y_n^+ = \mathcal{S}(\Delta t) Y_{n-1}^+,$$

where the split kinetic operator is

$$\mathcal{S}(\Delta t) = \mathcal{R}_\omega \circ \mathcal{T}(\Delta t).$$

Then we extract the first m components of Y_n^+ and we expect to obtain an approximation of W at time n:

$$Y_n^+(X) = \begin{pmatrix} W_n(X) \\ \vdots \end{pmatrix}, \quad W(X, t_n) \simeq W_n(X).$$

As stated above, the relaxation operator generates time fluctuations with frequency $1/\Delta t$. For the analysis, it is convenient to suppress these non-relevant time fluctuations. This is done by considering only an even number of time steps for instance. And this is equivalent to consider the modified split kinetic operator

$$\mathcal{S}(\Delta t) = \mathcal{T}\left(\frac{\Delta t}{4}\right) \circ \mathcal{R}_\omega \circ \mathcal{T}\left(\frac{\Delta t}{2}\right) \circ \mathcal{R}_\omega \circ \mathcal{T}\left(\frac{\Delta t}{4}\right). \tag{4.8}$$

In the case $\omega = 2$, this modified operator is time symmetric in the sense that

$$\mathcal{S}(0) = I_d, \quad \mathcal{S}(\Delta t)^{-1} = \mathcal{S}(-\Delta t). \tag{4.9}$$

This is this property that ensures that the split scheme is second order accurate when $\omega = 2$. We refer, for instance to [7] and to the consistency analysis given below.

5. Applications

In this section, we apply the entropy theory to practical examples.

5.1. **Application to the transport equation .** We first consider the construction of the D1Q2 model for the one-dimensional transport equation

$$\partial_t W + c\partial_x W = 0,$$

where the velocity c is supposed to be constant and $X = x$ is the one-dimensional space variable. In this case

$$Q(W) = cW, \quad V_1 = -\lambda, \quad V_2 = \lambda,$$

and we have no free choice for choosing the equilibrium kinetic data, which are given by

$$F_1^{\mathrm{eq}}(W) = \frac{W}{2} - \frac{cW}{2\lambda}, \quad F_2^{\mathrm{eq}}(W) = \frac{W}{2} + \frac{cW}{2\lambda}.$$

For this simple linear conservation law we can take the entropy associated to the L^2 norm

$$s(W) = \frac{W^2}{2}.$$

The dual entropy is simply

$$s^*(W^*) = \frac{W^{*2}}{2},$$

and the entropy variable is

$$W^* = \nabla_W s(W) = W.$$

Thus

$$F_1^{\mathrm{eq}}(W(W^*)) = \frac{W^*}{2} - \frac{cW^*}{2\lambda}, \quad F_2^{\mathrm{eq}}(W(W^*)) = \frac{W^*}{2} + \frac{cW^*}{2\lambda}.$$

From (3.10) we deduce the dual kinetic entropies

$$s_1^*(W^*) = \frac{1}{4}(1 - c/\lambda)W^{*2}, \quad s_2^*(W^*) = \frac{1}{4}(1 + c/\lambda)W^{*2}.$$

They are strictly convex under the subcharacteristic condition

$$\lambda > |c|\,. \tag{5.1}$$

We can then compute the kinetic entropies

$$s_1(F_1) = \frac{\lambda}{\lambda - c}F_1^2, \quad s_2(F_2) = \frac{\lambda}{\lambda + c}F_2^2.$$

The microscopic entropy is then

$$\sigma(F_1, F_2) = \frac{\lambda}{\lambda - c}F_1^2 + \frac{\lambda}{\lambda + c}F_2^2.$$

As expected, it is a diagonal quadratic form in the (F_1, F_2) variables.

Let us express the microscopic entropy with respect to the $Y = (W, y)^\intercal$ variables. We have

$$W = F_1 + F_2,$$

and

$$\begin{aligned} y &= -\lambda F_1 + \lambda F_2 - Q(W), \\ &= -\lambda F_1 + \lambda F_2 - c(F_1 + F_2). \end{aligned}$$

After simple computations, we find that the microscopic entropy is also

$$\widetilde{\sigma}(W, y) = \sigma(F_1, F_2) = \frac{W^2}{2} + \frac{y^2}{2(\lambda^2 - c^2)}. \tag{5.2}$$

It is a convex function of W and y under condition (5.1). As expected, it is minimal when the flux error y vanishes. In addition, in the relaxation step, the entropy is exactly conserved when $\omega = 2$ because

$$\widetilde{\sigma}(W, (1 - \omega)y) = \widetilde{\sigma}(W, -y) = \widetilde{\sigma}(W, y). \tag{5.3}$$

We are now in a position to prove the entropy stability of the over-relaxation scheme when $1 \leqslant \omega \leqslant 2$.

Theorem 7. *With periodic boundary conditions, or in an infinite domain, the over-relaxation scheme is entropy stable under the sub-characteristic condition (5.1) when* $1 \leqslant \omega \leqslant 2$.

Proof. It is sufficient to prove the decrease of the entropy

$$\mathcal{S}(t) = \int_x \sigma(F_1(x,t), F_2(x,t)) = \int_x s_1(F_1) + s_2(F_2),$$

for a single time step. In the transport step, one solves

$$\partial_t F_k + V_k \partial_x F_k = 0,$$

and thus the microscopic entropies

$$\overline{s}_k(t) = \int_x s_k(F_k),$$

are separately conserved

$$\overline{s}_k(t + \Delta t^-) = \overline{s}_k(t^+).$$

In the relaxation step, W is not changed and

$$y(x, t + \Delta t^+) = (1 - \omega) y(x, t + \Delta t^-),$$

because $|1 - \omega| \leqslant 1$ we see from the expression (5.2) of the entropy in the (W, y) variable that the microscopic entropy decreases pointwise, at each x. Therefore

$$\mathcal{S}(t + \Delta t^+) \leqslant \mathcal{S}(t + \Delta t^-).$$

□

5.2. **Application to the shallow water equations.** In order to show that the approach still works for non-linear systems of conservation laws, we try now to apply the above method to the shallow water model where the unknowns are the water height $h(x,t)$ and the velocity $u(x,t)$. It reads

$$\partial_t W + \partial_x Q(W) = 0,$$

with

$$W = \begin{pmatrix} h \\ hu \end{pmatrix}, \quad Q(W) = \begin{pmatrix} hu \\ hu^2 + gh^2/2 \end{pmatrix}.$$

We define the primitive variables

$$v = \begin{pmatrix} h \\ u \end{pmatrix}.$$

For smooth solutions, we also have

$$\partial_t v + B(v) \partial_x v = 0,$$

with

$$B(v) = \begin{pmatrix} u & h \\ g & u \end{pmatrix}.$$

Assume that the Lax entropy $s(W) = H(v)$ is expressed in the primitive variables, and that the entropy flux $G(W) = R(v)$. Then we must have

$$D_v H(v) B(v) = D_v R(v).$$

Denoting the partial derivatives with indices we obtain

$$\begin{pmatrix} H_h & H_u \end{pmatrix} \begin{pmatrix} u & h \\ g & u \end{pmatrix} = \begin{pmatrix} R_h & R_u \end{pmatrix}.$$

We search H under the form

$$H(h,u) = h\frac{u^2}{2} + e(h).$$

Because

$$H_h = \frac{u^2}{2} + e'(h), \quad H_u = hu,$$

this gives

$$\frac{u^3}{2} + ue' + ghu = R_h, \quad \frac{3hu^2}{2} + he' = R_u.$$

We take

$$R = h\frac{u^3}{2} + ue + gu\frac{h^2}{2}.$$

Then

$$R_u = \frac{3hu^2}{2} + e + g\frac{h^2}{2} = \frac{3hu^2}{2} + he'.$$

$e(h)$ is then solution of the differential equation

$$e - he' + gh^2/2 = 0.$$

A solution is

$$e(h) = \frac{gh^2}{2}.$$

In the end, we find

$$s(W) = h\frac{u^2}{2} + \frac{gh^2}{2}, \quad G(W) = h\frac{u^3}{2} + ugh^2.$$

This allows us to compute the entropy variables

$$W_1^* = gh - \frac{u^2}{2}, \quad W_2^* = u, \tag{5.4}$$

and the reverse change of variables

$$h = \frac{2W_1^* + W_2^{*2}}{2g}, \quad u = W_2^*.$$

The equilibrium kinetic vectors are given by

$$F_1^{\rm eq} = \frac{W}{2} - \frac{Q(W)}{2\lambda}, \quad F_2^{\rm eq} = \frac{W}{2} + \frac{Q(W)}{2\lambda}.$$

After some calculations, we can express this equilibrium in the entropy variables

$$F_1^{\rm eq} = \left[\begin{array}{cc} \frac{\left(W_2^{*2}+2W_1^*\right)(\lambda - W_2^*)}{4g\lambda} & -\frac{\left(W_2^{*2}+2W_1^*\right)\left(-4W_2^*\lambda+5W_2^{*2}+2W_1^*\right)}{16g\lambda} \end{array} \right]^{\mathsf{T}},$$

$$F_2^{\rm eq} = \left[\begin{array}{cc} \frac{\left(W_2^{*2}+2W_1^*\right)(\lambda + W_2^*)}{4g\lambda} & \frac{\left(W_2^{*2}+2W_1^*\right)\left(4W_2^*\lambda+5W_2^{*2}+2W_1^*\right)}{16g\lambda} \end{array} \right]^{\mathsf{T}}.$$

From the above theory, we know that

$$F_k^{\rm eq} = \nabla_{W^*} s_k^*,$$

for some dual kinetic entropies s_k^*. This is indeed the case and after more calculations we find

$$s_1^* = \frac{(\lambda - W_2^*)\left(W_2^{*2} + 2W_1^*\right)^2}{16g\lambda}, \quad s_2^* = \frac{\left(W_2^{*2} + 2W_1^*\right)^2(\lambda + W_2^*)}{16g\lambda}.$$

It is then possible to compute the Hessians of s_k^* and express them in the (h, u) variables with (5.4). We find

$$D_{W^*W^*}s_1^* = \left[\begin{array}{cc} \frac{\lambda - u}{2g\lambda} & \frac{-gh+\lambda u-u^2}{2g\lambda} \\ \frac{-gh+\lambda u-u^2}{2g\lambda} & \frac{\left(gh+u^2\right)\lambda - 3hug - u^3}{2g\lambda} \end{array} \right],$$

$$D_{W^*W^*}s_2^* = \left[\begin{array}{cc} \frac{\lambda + u}{2g\lambda} & \frac{gh+\lambda u+u^2}{2g\lambda} \\ \frac{gh+\lambda u+u^2}{2g\lambda} & \frac{\left(gh+u^2\right)\lambda + 3hug + u^3}{2g\lambda} \end{array} \right].$$

The two matrices are positive definite iff the first diagonal terms and the determinants are positive. This is equivalent to

$$\lambda > |u|,$$
$$(\lambda - u)^2 - gh > 0, \quad (\lambda + u)^2 - gh > 0,$$

which is again equivalent to

$$\lambda > |u| + \sqrt{gh}.$$

This is the expected sub-characteristic condition. It is difficult to go farther because the Legendre transforms s_1 and s_2 of s_1^* and s_2^* cannot be computed explicitly. However, we can reproduce the stability property of the linear case. The microscopic entropy is given by

$$\sigma(F_1, F_2) = s_1(F_1) + s_2(F_2).$$

Using the relations

$$W = F_1 + F_2,$$
$$y = -\lambda F_1 + \lambda F_2 - Q(F_1 + F_2),$$

we deduce

$$F_1 = \frac{W}{2} - \frac{Q(W)}{2\lambda} - \frac{y}{2\lambda}, \quad F_2 = \frac{W}{2} + \frac{Q(W)}{2\lambda} + \frac{y}{2\lambda},$$

and the microscopic entropy can be expressed in function of W and y

$$\widetilde{\sigma}(W,y) = s_1\left(\frac{W}{2} - \frac{Q(W)}{2\lambda} - \frac{y}{2\lambda}\right) + s_2\left(\frac{W}{2} + \frac{Q(W)}{2\lambda} + \frac{y}{2\lambda}\right).$$

For a fixed W the minimum is achieved for $y = 0$, therefore the macroscopic entropy is

$$s(W) = \widetilde{\sigma}(W,0),$$

and

$$\nabla_y \widetilde{\sigma}(W,0) = 0.$$

Then, with a Taylor expansion near to $y = 0$, we get

$$\widetilde{\sigma}(W,y) = \widetilde{\sigma}(W,-y) + O(|y|^3).$$

The relation (5.3) thus still holds but with a third-order term in y. This means that the relaxation scheme with $\omega = 2$ is entropy preserving up to third order in y. In principle, it is also possible to construct a scheme that preserves exactly the entropy in the non-linear case. It is sufficient to choose the relaxation parameter $\omega = \omega(W,y)$ in such way that

$$\widetilde{\sigma}(W,(1-\omega(W,y))y) = \widetilde{\sigma}(W,y). \tag{5.5}$$

In practice, this would not be very interesting, one would get

$$\omega(W,y) \simeq 2$$

and $\omega(W,y)$ would have to be computed numerically by first computing s_1 and $s_2$numerically and then by solving (5.5) also numerically.

What is interesting, however, is that the reasoning ensures the existence of a relaxation parameter $\omega(W,y) \simeq 2$, such that the whole scheme is entropy preserving. And if the scheme is run with a smaller relaxation parameter it is ensured to be entropy stable.

5.3. **Application to isothermal Euler equations.** Finally, we also apply the theory to the isothermal Euler model, which reads

$$\partial_t W + \partial_x Q(W) = 0,$$

with

$$W = \begin{pmatrix} \rho \\ \rho u \end{pmatrix}, \quad Q(W) = \begin{pmatrix} \rho u \\ \rho u^2 + c^2\rho \end{pmatrix}.$$

In primitive variables $v = (\rho, u)^\intercal$, the system reads

$$\partial_t v + B(v)\partial_x v = 0,$$

with

$$B(v) = \begin{pmatrix} u & \rho \\ \frac{c^2}{\rho} & u \end{pmatrix}.$$

First, let us find a Lax entropy $H(v)$ and an entropy flux $R(v)$. We must have

$$(H_\rho, H_u)\begin{pmatrix} u & \rho \\ \frac{c^2}{\rho} & u \end{pmatrix} = (R_\rho, R_u).$$

We search for $H(v)$ in the form

$$H = \rho\frac{u^2}{2} + e(\rho).$$

Then

$$R_\rho = \frac{u^3}{2} + e'(\rho)u + c^2 u,$$
$$R_u = \rho\frac{3u^2}{2} + e'(\rho)\rho.$$

Integrating the first equation with respect to ρ we get

$$R = \rho\frac{u^3}{2} + e(\rho)u + \rho c^2 u + C(\rho),$$

(we can take $C(\rho) = 0$). And deriving with respect to u we get

$$\rho\frac{3u^2}{2} + e'(\rho)\rho = \rho\frac{3u^2}{2} + e(\rho) + \rho c^2.$$

Therefore $e(\rho)$ is solution of the differential equation

$$e'(\rho)\rho = e(\rho) + \rho c^2.$$

We can take

$$e(\rho) = c^2\rho(\ln\rho - 1).$$

Finally

$$s(W) = \rho\frac{u^2}{2} + c^2\rho(\ln\rho - 1), \quad G(W) = u(s(W) + c^2\rho).$$

The entropy variables are

$$W_1^* = -\frac{u^2}{2} + c^2\ln\rho, \quad W_2^* = u.$$

The reverse change of variables is

$$\rho = W_1 = \exp\left(\frac{2W_1^* + W_2^{*2}}{2c^2}\right), \quad \rho u = W_2^*\exp\left(\frac{2W_1^* + W_2^{*2}}{2c^2}\right).$$

The equilibrium distribution is

$$F_1^{\rm eq} = \frac{W}{2} - \frac{Q(W)}{2\lambda} = \frac{1}{2}\begin{pmatrix} W_1 - W_2/\lambda \\ W_2 - (W_2^2/W_1 + c^2W_1)/\lambda \end{pmatrix}.$$

$$F_2^{\rm eq} = \frac{W}{2} + \frac{Q(W)}{2\lambda} = \frac{1}{2}\begin{pmatrix} W_1 + W_2/\lambda \\ W_2 + (W_2^2/W_1 + c^2W_1)/\lambda \end{pmatrix}.$$

In the dual variables we get

$$\nabla s_1^* = \frac{\exp\left(\frac{2W_1^*+W_2^{*2}}{2c^2}\right)}{2\lambda}\begin{pmatrix} \lambda - W_2^* \\ -W_2^{*2} + \lambda W_2^* - c^2 \end{pmatrix},$$

$$\nabla s_2^* = \frac{\exp\left(\frac{2W_1^*+W_2^{*2}}{2c^2}\right)}{2\lambda}\begin{pmatrix} \lambda + W_2^* \\ W_2^{*2} + \lambda W_2^* + c^2 \end{pmatrix},$$

and finally

$$s_1^* = \frac{\exp\left(\frac{2W_1^*+W_2^{*2}}{2c^2}\right)}{2\lambda}(\lambda - W_2^*), \quad s_2^* = \frac{\exp\left(\frac{2W_1^*+W_2^{*2}}{2c^2}\right)}{2\lambda}(\lambda + W_2^*).$$

With similar calculations as for the shallow water system we find the following sub-characteristic condition

$$\lambda > c + |u|.$$

As for the shallow water system, it is difficult to compute explicitly s_1 and s_2.

6. Equivalent equation analysis

The entropy analysis of the over-relaxation scheme ensures the stability of the scheme as soon as the sub-characteristic condition is satisfied. However, for the moment it is not obvious that the scheme provides an approximation of the system (2.1). Indeed, the consistency of the kinetic approximation with the system of conservation laws is ensured, as soon as

$$F(X,t) \simeq F^{eq}(W(X,t)),$$

and the over-relaxation formula (4.6) enforces $F = F^{eq}$ at the end of the time step only when $\omega = 1$.

The objective of the equivalent system analysis is to provide a consistency theory in the case $1 < \omega \leqslant 2$. This consistency analysis is based on two ingredients: a Taylor expansion followed by a Chapman-Enskog analysis.

6.1. **Taylor expansion: equivalent system in Y.** The first ingredient is a Taylor expansion of

$$\frac{\mathcal{S}(\Delta t) - \mathcal{S}^{-1}(\Delta t)}{2\Delta t}$$

when Δt tends to zero. The Taylor expansion provides an equivalent system of Partial Differential Equations (PDE) expressed on Y. This system involves the whole vector Y, which contains both W and the flux error. We denote it by the *equivalent system* in the following. It takes the general form

$$\partial_t Y + \frac{r(\omega)}{\Delta t}\begin{pmatrix} 0 \\ Y^{m+1\cdots} \end{pmatrix} + \sum_{1\leqslant i\leqslant d} A^i(Y,\omega)\partial_i Y - \Delta t \sum_{1\leqslant i,j\leqslant d} B^{i,j}(Y,\omega)\partial_{i,j} Y = O(\Delta t^2). \tag{6.1}$$

In the case $\omega = 2$, the formula is simpler and we shall find that

$$r(2) = 0, \quad B^{i,j}(Y,2) = 0, \tag{6.2}$$

and that the matrices A^i are of the form

$$A^i(Y,2) = \begin{pmatrix} D_W Q^i(W) & 0_m \\ 0_{m(n_v-1)} & \times \end{pmatrix}. \tag{6.3}$$

The Taylor expansion is thus sufficient to get the second-order time consistency of the split scheme with the initial system of conservation law when $\omega = 2$. This consistency holds even when the flux error is large. In addition, up to third order terms, the evolution of W is uncoupled from the evolution of the flux error. This surprising result relies essentially on the symmetry property (4.9), which ensures that the stiff terms in Δt^{-1} vanish in the Taylor expansion.

6.2. **Asymptotic analysis: equivalent equation in W.** In the case $\omega \neq 2$, the equivalent system contains stiff terms. And the evolution of W is no more uncoupled from the evolution of the flux error. In order to remove this coupling, the second step is to perform an additional asymptotic analysis (similar to the Hilbert or Chapman-Enskog expansion), with the assumption that the flux error is of order $O(\Delta t)$. From (6.1) we can deduce an algebraic relation between the flux error $Y^{m+1\cdots}$ and the gradient of W

$$Y^{m+1\cdots} = -\frac{\Delta t}{r(\omega)} T \sum_{1\leqslant i\leqslant d} A^i(Y,\omega)\partial_i \begin{pmatrix} W \\ 0 \end{pmatrix} + O(\Delta t^2), \quad T = \begin{pmatrix} 0_m & 0_{m(n_v-1)} \\ 0_{m(n_v-1)} & 1_{m(n_v-1)} \end{pmatrix}.$$

Reinjecting this approximation in the first row of (6.1) provides a simpler system involving only W. We denote it by the *equivalent equation* in the following. It takes the form

$$\partial_t W + \sum_{1\leqslant i\leqslant d} \partial_i Q^i(W) - \Delta t \sum_{1\leqslant i,j\leqslant d} \partial_i \left(D^{i,j}(W,\omega)\partial_j W \right) = O(\Delta t^2). \tag{6.4}$$

6.3. **Stability conditions.** Once we have obtained the equivalent system (6.1) and the equivalent equation (6.4) we can study their stability.

6.3.1. *Equivalent system.* From the theory developed in Section (3), we have stability under the subcharacteristic condition. This condition ensures dissipation of the kinetic entropy $\tilde{\sigma}(Y) = \tilde{\sigma}(W, Y^{m+1\cdots mn_v}) = \sum_k s_k(F_k)$. We also know that

$$\tilde{\sigma}(W,0) = s(W),$$

corresponds to the minimum of $\tilde{\sigma}$ with respect to the flux error $Y^{m+1\cdots mn_v}$. It is therefore expected that the change of variables $Y \mapsto \nabla_Y \tilde{\sigma}(Y)$ symmetrizes the equivalent system (6.1), which is thus hyperbolic. We have also seen that the computation of the kinetic entropies s_k is not necessarily easy. That is why in the following we try to find directly a symmetrization for the first order part of the equivalent system, rather than computing the kinetic entropy. We introduce the following definition.

Definition 8. We shall say that the equivalent system (6.1) is *hyperbolic* iff we can find a matrix $P^0(Y)$, $0 \leqslant i \leqslant d$, such that $P^0(Y)$ is symmetric positive definite and $P^0(Y)A^i(Y,\omega)$ is symmetric for $1 \leqslant i \leqslant d$.

Remark 9. Our definition of hyperbolicity is stronger than the usual one, which only states that

$$\sum_{i=1}^{d} N_i A^i(Y,\omega)$$

is diagonalizable with real eigenvalues for all directions N. We expect that

$$P^0(Y) = \nabla^2 \tilde{\sigma}(Y),$$

at least in the case of the linear transport equation discussed in Section 5.1. But we don't know if the stability condition of Definition 8 is equivalent to the convexity of the dual kinetic entropies. Maybe that it is true. For a discussion around these questions, we refer to [4].

6.3.2. *Equivalent equation.* For the equivalent equation, we already know that its first order part is hyperbolic. The stability thus depends on the sign of the second-order terms. This gives another stability criterion. For obtaining this criterion, we multiply (6.4) on the left by $D_W s(W)$ and we integrate by part the second-order term. The entropy is dissipated if

$$\sum_{ij,k,\ell} \partial^2_{k,\ell} s \partial_i W^k D^{i,j} \partial_j W^\ell \geqslant 0.$$

We thus introduce the quadratic form acting on a second-order tensor α_i^k:

$$\alpha \mapsto q(\alpha) = \sum_{i,j,k,\ell} \partial^2_{k,\ell} s(W) D^{i,j}(W,\omega) \alpha_i^k \alpha_j^l.$$

Definition 10. The equivalent equation (6.4) is *dissipative* iff the quadratic form $q(\alpha)$ is positive.

Remark 11. This definition amounts to checking that the Hessian matrix $\nabla^2_\alpha q(\alpha)$ is positive. This (symmetric) matrix is of size $md \times md$. In the scalar case $m = 1$, which we study below, the condition is simpler. It simply states that the quadratic form

$$x \mapsto \sum_{i,j} D^{i,j} x_i x_j$$

is positive.

7. Applications to the transport equation

Now we apply the equivalent system analysis and the equivalent equation analysis to the simple scalar transport equation (thus $m = 1$)

$$\partial_t w + \sum_{i=1}^{d} \partial_i (v_i w) = 0,$$

where the velocity vector $(v_1, \dots, v_d)$ is supposed to be constant. We consider the cases $d = 1$ or $d = 2$ and the D1Q2, D2Q3 and D2Q4 models.

For each model, we compute the equivalent system. The expansion of

$$\frac{\mathcal{S}(\Delta t) - \mathcal{S}^{-1}(\Delta t)}{2\Delta t} Y(\cdot, t) = \partial_t Y(\cdot, t) + O(\Delta t^2),$$

is performed with the Computer Algebra System (CAS) Maple. This is done by entering the explicit definition of the symmetric operator given in Section 4.2, step by step. Without a CAS, the calculations would be extremely tedious...

7.1. **D1Q2.** For the D1Q2 model, we use the notations

$$Y = \begin{pmatrix} w \\ y \end{pmatrix}, \quad v_1 = v.$$

The equivalent system reads

$$\begin{aligned}
&\partial_t \begin{pmatrix} w \\ y \end{pmatrix} - \frac{\omega(\omega-2)(\omega^2-2\omega+2)}{2\Delta t(\omega-1)^2}\begin{pmatrix} 0 \\ y \end{pmatrix} + \begin{pmatrix} v & \gamma_1 \\ (\lambda^2-v^2)\gamma_1 & \frac{-v(\omega^4-4\omega^3+6\omega^2-4\omega+2)}{2(\omega-1)^2} \end{pmatrix}\partial_x \begin{pmatrix} w \\ y \end{pmatrix} \\
&+ \begin{pmatrix} -(\omega^2-6\omega+6)(\lambda^2-v^2) & 3v(\omega^2-2\omega+2) \\ 3v(\lambda^2-v^2)(\omega^2-2\omega+2) & -5v^2\omega^2-3\lambda^2\omega^2+6v^2\omega+10\lambda^2\omega-6v^2-10\lambda^2 \end{pmatrix} \\
&\times \frac{\Delta t\omega(\omega-2)}{32(\omega-1)^2}\partial_{xx}\begin{pmatrix} w \\ y \end{pmatrix} = O(\Delta t^2),
\end{aligned} \tag{7.1}$$

with $\gamma_1 = \frac{(\omega-2)^2(\omega^2-2\omega+2)}{8(\omega-1)^2}$. We can check that when $\omega = 2$, we indeed obtain an equivalent system with the simplification (6.2) and (6.3). The stiff term vanishes and the evolution of w is uncoupled from that of y at order 2. This means that the consistency is achieved even when the flux error y is large.

For obtaining the equivalent equation we assume that we have $y = O(\Delta t)$. Let us write $y = \Delta t\tilde{y}$. We obtain

$$\frac{\omega(\omega-2)(\omega^2-2\omega+2)}{2(\omega-1)^2}\tilde{y} = \frac{(\lambda^2-v^2)(\omega-2)^2(\omega^2-2\omega+2)}{8(\omega-1)^2}\partial_x w + O(\Delta t).$$

By simplifying, we have

$$y = \frac{(\lambda^2-v^2)(\omega-2)}{4\omega}\Delta t\partial_x w + O(\Delta t^2). \tag{7.2}$$

By reinjecting this expression of y in the first equation of equivalent system 7.1, we obtain the equivalent equation on w

$$\begin{aligned}
&\partial_t w + v\partial_x w + \frac{(\omega-2)^2(\omega^2-2\omega+2)}{8(\omega-1)^2}\frac{(\lambda^2-v^2)(\omega-2)}{4\omega}\Delta t\partial_{xx} w \\
&- \frac{\Delta t\omega(\omega-2)}{32(\omega-1)^2}(\omega^2-6\omega+6)(\lambda^2-v^2)\partial_{xx} w = O(\Delta t^2),
\end{aligned}$$

which can be simplified in

$$\partial_t w + v\partial_x w = \frac{1}{2}\left(\frac{1}{\omega}-\frac{1}{2}\right)(\lambda^2-v^2)\Delta t\partial_{xx} w + O(\Delta t^2). \tag{7.3}$$

We can notice that we recover the equivalent equation given in [15, 20, 8]

Theorem 12. *When $\omega \neq 2$, the sub-characteristic diffusive stability condition of the $D1Q2$ model is*

$$|v| < \lambda.$$

Proof. The equivalent equation on w of the $D1Q2$ model (7.3) is stable if the diffusion term is positive. As $\omega \in [1, 2]$, the term $\left(\frac{1}{\omega}-\frac{1}{2}\right)$ is positive. The positivity of the diffusion term is then equivalent to

$$\lambda^2 - v^2 > 0,$$

which gives us the stability condition

$$|v| < \lambda.$$

□

Remark 13. When $\omega = 2$, the diffusion term of the equivalent equation of the $D1Q2$ model disappears, which gives us

$$\partial_t w + v\partial_x w = O(\Delta t^2).$$

We obtain that the solution given by the $D1Q2$ model is an approximation of order 2 of the solution of the initial equation.

Theorem 14. *The matrix*

$$P = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\lambda^2-v^2} \end{pmatrix},$$

symmetrizes the equivalent system of the $D1Q2$ model (7.1), if the diffusive sub-characteristic stability condition is satisfied. Consequently, the equivalent system (7.1) is hyperbolic if

$$|v| < \lambda.$$

Proof. We search a matrix $P = \begin{pmatrix} p_1 & p_2 \\ p_2 & p_3 \end{pmatrix}$ such as PA is symmetric and P is symmetric positive definite. We have

$$\begin{aligned} PA &= \begin{pmatrix} p_1 & p_2 \\ p_2 & p_3 \end{pmatrix} \begin{pmatrix} v & \gamma_1 \\ (\lambda^2 - v^2)\gamma_1 & -v\gamma_2 \end{pmatrix}, \\ &= \begin{pmatrix} vp_1 + (\lambda^2 - v^2)\gamma_1 p_2 & \gamma_1 p_1 - v\gamma_2 p_2 \\ vp_2 + (\lambda^2 - v^2)\gamma_1 p_3 & \gamma_1 p_2 - v\gamma_2 p_3 \end{pmatrix}. \end{aligned}$$

with $\gamma_1 = \frac{(\omega-2)^2(\omega^2-2\omega+2)}{8(\omega-1)^2}$, and $\gamma_2 = \frac{(\omega^4-4\omega^3+6\omega^2-4\omega+2)}{2(\omega-1)^2}$. As we want PA to be symmetric, we need to satisfy the condition

$$\gamma_1 p_1 - v\gamma_2 p_2 = vp_2 + (\lambda^2 - v^2)\gamma_1 p_3,$$

which is equivalent to

$$p_3 = \frac{1}{(\lambda^2 - v^2)} p_1 - v \frac{1 + \gamma_2}{(\lambda^2 - v^2)\gamma_1} p_2.$$

Let us choose $p_2 = 0$ and $p_1 = 1$. We obtain

$$P = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\lambda^2 - v^2} \end{pmatrix}.$$

As its eigenvalues are 1 and $\frac{1}{\lambda^2-v^2}$, P is definite positive if

$$|v| < \lambda.$$

□

Remark 15. We obtain the same condition on v and λ as for the diffusive stability condition given in Proposition 12. In this case, the diffusive analysis and the hyperbolicity analysis give the same stability condition.

7.2. **D2Q3.** For the D2Q3, we use the notations

$$Y = \begin{pmatrix} w \\ y_1 \\ y_2 \end{pmatrix}, \quad v_1 = a, \quad v_2 = b.$$

The equivalent system of the $D2Q3$ model is

$$\begin{aligned} &\partial_t \begin{pmatrix} w \\ y_1 \\ y_2 \end{pmatrix} - \frac{\omega(\omega-2)(\omega^2-2\omega+2)}{2\Delta t(\omega-1)^2} \begin{pmatrix} 0 \\ y_1 \\ y_2 \end{pmatrix} \\ &\quad + \begin{pmatrix} a & -2\gamma_1 & 0 \\ \gamma_1(2a+\lambda)(a-\lambda) & \gamma_2(2a-\lambda) & 0 \\ \gamma_1 b(2a+\lambda) & 2b\gamma_2 & \gamma_2\lambda \end{pmatrix} \partial_1 \begin{pmatrix} w \\ y_1 \\ y_2 \end{pmatrix} \\ &\quad + \begin{pmatrix} b & 0 & -2\gamma_1 \\ \gamma_1 b(2a+\lambda) & 0 & \gamma_2(2a+\lambda) \\ \gamma_1(a\lambda + 2b^2 - \lambda^2) & \gamma_2\lambda & 2b\gamma_2 \end{pmatrix} \partial_2 \begin{pmatrix} w \\ y_1 \\ y_2 \end{pmatrix} = O(\Delta t), \end{aligned} \tag{7.4}$$

with $\gamma_1 = -\frac{1}{16}\frac{(\omega^2-2\omega+2)(\omega-2)^2}{(\omega-1)^2}$ and $\gamma_2 = -\frac{1}{4}\frac{\omega^4-4\omega^3+6\omega^2-4\omega+2}{(\omega-1)^2}$.

For getting the equivalent equation, we assume that $(y_1, y_2) = O(\Delta t)$. As above, we express the flux error in function of the gradient of w up to order 2. This gives:

$$y_1 = \Delta t \left(\frac{1}{\omega} - \frac{1}{2}\right) \frac{(2a+\lambda)}{2} \left((a-\lambda)\,\partial_1 w + b\partial_2 w\right) + O(\Delta t^2),$$

and

$$y_2 = \frac{\Delta t}{2} \left(\frac{1}{\omega} - \frac{1}{2}\right) \left((2ab + b\lambda)\,\partial_1 w + \left(\lambda a + 2b^2 - \lambda^2\right)\partial_2 w\right) + O(\Delta t^2).$$

We reinject these expressions of y_1 and y_2 in the first equation of the equivalent system (7.4), we obtain

$$\partial_t w + \sum_{i=1}^{2} \partial_i(v_i w) = \frac{\Delta t}{2}\left(\frac{1}{\omega} - \frac{1}{2}\right) \nabla \cdot (\mathcal{D}_3 \nabla w) + O(\Delta t^2), \tag{7.5}$$

with the diffusion matrix

$$\mathcal{D}_3 = \begin{pmatrix} \frac{\lambda}{2}(\lambda + a) - a^2 & -\frac{\lambda}{2}b - ab \\ -\frac{\lambda}{2}b - ab & \frac{\lambda}{2}(\lambda - a) - b^2 \end{pmatrix}.$$

Theorem 16. *The sub-characteristic stability condition of the D2Q3 model is*

$$\lambda^2 - a^2 - b^2 - \sqrt{(a^2+b^2)^2 + \lambda(-2a^3 + 6ab^2) + \lambda^2(a^2+b^2)} > 0.$$

Remark 17. This condition can also be written as the intersection of

$$a^2 + b^2 < \lambda^2,$$

and

$$\lambda^3 - 3\lambda(a^2+b^2) + 2a^3 - 6ab^2 > 0.$$

Proof. Indeed, with a linear flux, we have

$$\partial_t w + \sum_{i=1}^{2} \partial_i(v_i w) = \frac{\Delta t}{2}\left(\frac{1}{\omega} - \frac{1}{2}\right) \nabla \cdot (\mathcal{D}_3 \nabla w) + O(\Delta t^2),$$

with the diffusion matrix

$$\mathcal{D}_3 = \begin{pmatrix} \frac{\lambda}{2}(\lambda + a) - a^2 & -\frac{\lambda}{2}b - ab \\ -\frac{\lambda}{2}b - ab & \frac{\lambda}{2}(\lambda - a) - b^2 \end{pmatrix}.$$

The eigenvalues of this diffusion matrix are

$$d_{1,2} = \frac{1}{2}\left(\lambda^2 - a^2 - b^2 \pm \sqrt{(a^2+b^2)^2 + \lambda(-2a^3 + 6ab^2) + \lambda^2(a^2+b^2)}\right).$$

Finally, the model $D2Q3$ is stable if $\mathcal{D}_3$ is positive definite, namely if $d_1 > 0$ and $d_2 > 0$. □

Theorem 18. *The matrix*

$$P = \begin{pmatrix} \frac{\lambda}{2}(a^2 - 2a\lambda - 3b^2 + \lambda^2)(2a + \lambda) & 0 & 0 \\ 0 & -(a\lambda + 2b^2 - \lambda^2) & b(2a+\lambda) \\ 0 & b(2a+\lambda) & -(a-\lambda)(2a+\lambda) \end{pmatrix}.$$

symmetrizes the equivalent system of the D2Q3 model (7.4)*, if the diffusive sub-characteristic stability condition* (16) *is verified. Consequently, the equivalent system* (7.4) *is hyperbolic if*

$$\lambda^2 - a^2 - b^2 - \sqrt{(a^2+b^2)^2 + \lambda(-2a^3 + 6ab^2) + \lambda^2(a^2+b^2)} > 0.$$

Proof.

We are searching for a matrix $P = \begin{pmatrix} p_1 & p_2 & p_3 \\ p_2 & p_4 & p_5 \\ p_3 & p_5 & p_6 \end{pmatrix}$ such as PA^1 and PA^2 are symmetric and P is symmetric positive definite. When we compute the matrices PA^1 and PA^2, the symmetry imposes 6 equations on the unknown $p_1, p_2, p_3, p_4, p_5, p_6$. This gives us the matrix

$$P = \begin{pmatrix} \frac{\lambda}{2}\frac{(a^2 - 2a\lambda - 3b^2 + \lambda^2)p_5}{b} & 0 & 0 \\ 0 & -\frac{(a\lambda + 2b^2 - \lambda^2)p_5}{b(2a+\lambda)} & p_5 \\ 0 & p_5 & -\frac{(a-\lambda)p_5}{b} \end{pmatrix},$$

where p_5 must be chosen. We choose $p_5 = b(2a + \lambda)$. We obtain

$$P = \begin{pmatrix} \frac{\lambda}{2}(a^2 - 2a\lambda - 3b^2 + \lambda^2)(2a + \lambda) & 0 & 0 \\ 0 & -(a\lambda + 2b^2 - \lambda^2) & b(2a+\lambda) \\ 0 & b(2a+\lambda) & -(a-\lambda)(2a+\lambda) \end{pmatrix}.$$

The eigenvalues of P are

$$e_1 = \frac{\lambda}{2}(a^2 - 2a\lambda - 3b^2 + \lambda^2)(2a + \lambda),$$

$$e_2 = \lambda^2 - a^2 - b^2 + \sqrt{(a^2+b^2)^2 + \lambda(-2a^3+6ab^2) + \lambda^2(a^2+b^2)},$$

and

$$e_3 = \lambda^2 - a^2 - b^2 - \sqrt{(a^2+b^2)^2 + \lambda(-2a^3+6ab^2) + \lambda^2(a^2+b^2)}.$$

By noticing that $e_2 > e_3$ and $e_2 e_3 = 2e_1$, we deduce that P is definite positive if $e_3 > 0$. □

Remark 19. The hyperbolicity condition on a, b and λ is the same as the diffusive stability condition given in the Proposition 16. Here again, the diffusive analysis and the hyperbolicity analysis are equivalent.

7.3. **D2Q4.** For the D2Q4, we use the notations

$$Y = \begin{pmatrix} w \\ y_1 \\ y_2 \\ z_3 \end{pmatrix}, \quad v_1 = a, \quad v_2 = b.$$

We can also compute the equivalent system on (w, y_1, y_2, z_3) of the $D2Q4$ model. We obtain

$$\begin{aligned} &\partial_t \begin{pmatrix} w \\ y_1 \\ y_2 \\ z_3 \end{pmatrix} - \frac{\omega(\omega-2)(\omega^2-2\omega+2)}{2\Delta t(\omega-1)^2} \begin{pmatrix} 0 \\ y_1 \\ y_2 \\ z_3 \end{pmatrix} + \begin{pmatrix} a & 2\gamma_1 & 0 & 0 \\ \gamma_1(\lambda^2-2a^2) & -2a\gamma_2 & 0 & \gamma_2 \\ -2ab\gamma_1 & -2b\gamma_2 & 0 & 0 \\ 2\lambda^2 a\gamma_1 & 2\lambda^2\gamma_2 & 0 & 0 \end{pmatrix} \partial_1 \begin{pmatrix} w \\ y_1 \\ y_2 \\ z_3 \end{pmatrix} \\ &\quad + \begin{pmatrix} b & 0 & 2\gamma_1 & 0 \\ -2ab\gamma_1 & 0 & -2a\gamma_2 & 0 \\ \gamma_1(\lambda^2-2b^2) & 0 & -2b\gamma_2 & -\gamma_2 \\ -2\lambda^2 b\gamma_1 & 0 & -2\lambda^2\gamma_2 & 0 \end{pmatrix} \partial_2 \begin{pmatrix} w \\ y_1 \\ y_2 \\ z_3 \end{pmatrix} = O(\Delta t), \end{aligned} \tag{7.6}$$

with $\gamma_1 = \frac{(\omega-2)^2(\omega^2-2\omega+2)}{16(\omega-1)^2}$ and $\gamma_2 = \frac{\omega^4-4\omega^3+6\omega^2-4\omega+2}{4(\omega-1)^2}$.

With the same method as above, we derive the equivalent equation

$$\partial_t w + \sum_{i=1}^{2} \partial_i(v_i w) = \frac{\Delta t}{2}\left(\frac{1}{\omega} - \frac{1}{2}\right) \nabla \cdot (\mathcal{D}_4 \nabla w) + O(\Delta t^2),$$

with the diffusion matrix

$$\mathcal{D}_4 = \begin{pmatrix} \frac{\lambda^2}{2} - a^2 & -ab \\ -ab & \frac{\lambda^2}{2} - b^2 \end{pmatrix}.$$

Theorem 20. *The D2Q4 model is stable if* $a^2 + b^2 \leqslant \frac{\lambda^2}{2}$.

Proof. We have

$$\partial_t w + \sum_{i=1}^{2} \partial_i(v_i w) = \frac{\Delta t}{2}\left(\frac{1}{\omega} - \frac{1}{2}\right) \nabla \cdot (\mathcal{D}_4 \nabla w) + O(\Delta t), \tag{7.7}$$

with $\mathcal{D}_4 = \begin{pmatrix} \frac{\lambda^2}{2} - a^2 & -ab \\ -ab & \frac{\lambda^2}{2} - b^2 \end{pmatrix}$. The model is stable if the diffusion matrix $\mathcal{D}_4$ is positive. Its eigenvalues are:

$$e_1 = \frac{1}{2}\left(\lambda^2 - a^2 - b^2 - \sqrt{\left(\lambda^2 - a^2 - b^2\right)^2 - \lambda^4 + 2\lambda^2 b^2 + 2\lambda^2 a^2}\right) \quad \text{and}$$

$$e_2 = \frac{1}{2}\left(\lambda^2 - a^2 - b^2 + \sqrt{\left(\lambda^2 - a^2 - b^2\right)^2 - \lambda^4 + 2\lambda^2 b^2 + 2\lambda^2 a^2}\right).$$

As $e_1 \leqslant e_2$, the eigenvalues are both positive if $e_1 \geqslant 0$, which means if

$$a^2 + b^2 \leqslant \frac{\lambda^2}{2}.$$

□

Theorem 21. *The matrix*

$$P = \begin{pmatrix} \lambda^2(4a^2-\lambda^2)(4b^2-\lambda^2) & 0 & 0 & 0 \\ 0 & -2\lambda^2(4b^2-\lambda^2) & 0 & 2a(4b^2-\lambda^2) \\ 0 & 0 & -2\lambda^2(4a^2-\lambda^2) & -2b(4a^2-\lambda^2) \\ 0 & 2a(4b^2-\lambda^2) & -2b(4a^2-\lambda^2) & -2a^2-2b^2+\lambda^2 \end{pmatrix},$$

symmetrizes the equivalent system of the D2Q4 model (7.6)*, if*

$$4\max(a^2,b^2) < \lambda^2. \tag{7.8}$$

Consequently, under this condition, the equivalent system (7.6) *is hyperbolic.*

Proof. We are searching for a matrix

$$P = \begin{pmatrix} p_1 & p_2 & p_3 & p_4 \\ p_2 & p_5 & p_6 & p_7 \\ p_3 & p_6 & p_8 & p_9 \\ p_4 & p_7 & p_9 & p_{10} \end{pmatrix},$$

such as PA^1 and PA^2 are symmetric and P is symmetric positive definite. We can compute PA^1 and PA^2. As we want these matrices to be symmetric, we obtain conditions on the coefficients p_i. We deduce that

$$P = \begin{pmatrix} \frac{1}{2a}\lambda^2(2a-\lambda)(2a+\lambda)p_7 & 0 & 0 & 0 \\ 0 & -p_7\lambda^2/a & 0 & p_7 \\ 0 & 0 & -\frac{p_7\lambda^2(2a-\lambda)(2a+\lambda)}{(a(2b-\lambda)(2b+\lambda))} & -\frac{bp_7(2a-\lambda)(2a+\lambda)}{(a(2b-\lambda)(2b+\lambda))} \\ 0 & p_7 & -\frac{bp_7(2a-\lambda)(2a+\lambda)}{(a(2b-\lambda)(2b+\lambda))} & -\frac{1(2a^2+2b^2-\lambda^2)p_7}{2(a(2b-\lambda)(2b+\lambda))} \end{pmatrix}.$$

By choosing $p_7 = 2a(2b-\lambda)(2b+\lambda)$, we obtain

$$P = \begin{pmatrix} \lambda^2(4a^2-\lambda^2)(4b^2-\lambda^2) & 0 & 0 & 0 \\ 0 & -2\lambda^2(4b^2-\lambda^2) & 0 & 2a(4b^2-\lambda^2) \\ 0 & 0 & -2\lambda^2(4a^2-\lambda^2) & -2b(4a^2-\lambda^2) \\ 0 & 2a(4b^2-\lambda^2) & -2b(4a^2-\lambda^2) & -2a^2-2b^2+\lambda^2 \end{pmatrix}. \tag{7.9}$$

As P is symmetric, according to the Sylvester's criterion, P is positive definite if and only if all of the leading principal minors are positive, that is to say if the following conditions are satisfied

$$\begin{cases} |P_1| &=& \lambda^2(4b^2-\lambda^2)(4a^2-\lambda^2) &>& 0, \\ |P_2| &=& -2\lambda^4(4a^2-\lambda^2)(4b^2-\lambda^2)^2 &>& 0, \\ |P_3| &=& 4\lambda^6(4a^2-\lambda^2)^2(4b^2-\lambda^2)^2 &>& 0, \\ |P_4| &=& 4\lambda^4(4a^2-\lambda^2)^3(4b^2-\lambda^2)^3 &>& 0. \end{cases}$$

This is equivalent to

$$\begin{cases} 4a^2 &<& \lambda^2, \\ 4b^2 &<& \lambda^2, \end{cases}$$

which can be rewritten

$$2\max(|a|,|b|) < \lambda.$$

□

Remark 22. The hyperbolicity condition obtained is more restrictive than the diffusive stability condition obtained in Proposition 20. We can see in Figure 7.1 the values of a/λ and b/λ for which the diffusive stability condition is verified, the circle colored in yellow, are included in the blue square, for which the hyperbolicity condition is checked. This is coherent with the review of stability conditions given by Bouchut in [4].

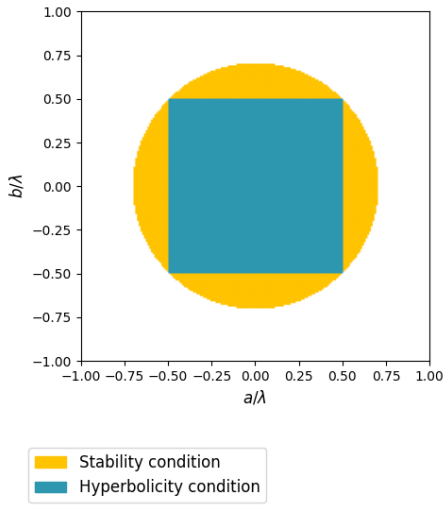


FIGURE 7.1. Graphic representation of the diffusive stability and hyperbolicity condition of the $D2Q4$ model.

8. Some numerical results

8.1. **D1Q2: consistency error.** Now that we have obtained the equivalent equations, we wish to quantify numerically how they are close to the kinetic equations. We shall compute analytic solutions of the equivalent equations and compare them with the solutions of the kinetic equation. We shall also compute the error between the two solutions.

For the numerical experiments, we use particular solutions of the form

$$\begin{pmatrix} w \\ y \end{pmatrix}(x,t) = \begin{pmatrix} w_0 \\ y_0 \end{pmatrix} e^{\gamma t} e^{ikx}, \tag{8.1}$$

with $k \in \mathbb{N}$ and $\gamma \in \mathbb{C}$.

8.1.1. *Particular solution of the equivalent equation.* If we inject this particular solution (8.1) in the equivalent equation (7.3) on w, we obtain

$$\gamma w + ivkw = -\frac{\Delta t}{2}\left(\frac{1}{\omega} - \frac{1}{2}\right) k^2 \left(\lambda^2 - v^2\right) w.$$

It gives us the value of γ with respect to k and v

$$\gamma = -\frac{\Delta t}{2}\left(\frac{1}{\omega} - \frac{1}{2}\right) k^2 \left(\lambda^2 - v^2\right) - ivk.$$

A particular solution of the equivalent equation (7.3) is then

$$w = w_0 e^{-\left(\frac{\Delta t}{2}\left(\frac{1}{\omega}-\frac{1}{2}\right)k^2\left(\lambda^2-v^2\right)+ivk\right)t} e^{ikx}.$$

In order to deal with real solutions, we compute the real part of this particular solution, that we denote w_{eqeq} and which is still a solution of (7.3)

$$w_{\text{eqeq}} = \Re(w) = w_0 e^{-\frac{\Delta t}{2}\left(\frac{1}{\omega}-\frac{1}{2}\right)k^2\left(\lambda^2-v^2\right)t} \cos\left(k(x-vt)\right).$$

To compute the equivalent equation, we assume that we have the relation between y and $\partial_x w$ given by (7.2)

$$y = \frac{(\lambda^2 - v^2)(\omega - 2)}{4\omega} \Delta t \partial_x w. \tag{8.2}$$

We denote y_{eqeq} the real part of y

$$\begin{aligned} y_{\text{eqeq}} &= \Re(y), \\ &= -\frac{(\lambda^2 - v^2)(\omega - 2)}{4\omega} \Delta t k w_0 e^{-\frac{\Delta t}{2}\left(\frac{1}{\omega}-\frac{1}{2}\right)k^2\left(\lambda^2-v^2\right)t} \sin\left(k(x-vt)\right). \end{aligned}$$

8.1.2. *Particular solution of the equivalent system.* Now, we inject the expression of the particular solution (8.1) in the equivalent system (7.1). We obtain

$$\left(\gamma I_2 + \frac{1}{\Delta t}R(\omega) + ikA(Y,\omega) + \Delta t k^2 B(Y,\omega)\right)\begin{pmatrix} w \\ y \end{pmatrix} = 0,$$

with $R(\omega) = \begin{pmatrix} 0 & 0 \\ 0 & r(\omega) \end{pmatrix}$.

The previous system admits two solutions $\gamma_1(k)$ and $\gamma_2(k)$ depending on k, which are the eigenvalues of $-\frac{1}{\Delta t}R(\omega) - ikA(Y,\omega) - \Delta t k^2 B(Y,\omega)$. We have

$$\gamma_1(k) = \frac{1}{\Delta t}\frac{16\omega^4 - 64\omega^3 + 96\omega^2 - 64\omega}{32(\omega-1)^2} + O(\Delta t^0),$$

and

$$\gamma_2(k) = -ikD(W,\omega) - \frac{D(W,\omega)^2(\lambda^2\omega^2 - 2) + 2\lambda^2}{4\omega}k^2\Delta t + O(\Delta t^2).$$

One of the solutions, $\gamma_1(k)$, behaves as $O(\frac{1}{\Delta t})$ when $\Delta t \to 0$, and the real part of the other solution $\gamma_2(k)$ behaves as $O(\Delta t)$ when $\Delta t \to 0$. If we compute the particular solution (8.1) with the eigenvalue γ_1 in $O(\frac{1}{\Delta t})$, we observe that y decreases rapidly toward 0. If we consider instead, the solution given by the second eigenvalue γ_2, y stays small and has slower variations. We choose to keep this eigenvalue γ_2 for a relevant comparison with the expected behavior.

A particular solution of the equivalent system (7.1) is then

$$\begin{pmatrix} w \\ y \end{pmatrix}(x,t) = \begin{pmatrix} w_0 \\ y_0 \end{pmatrix} e^{\gamma_2 t} e^{ikx}.$$

To test, we compute the real part of w, that we denote w_{syseq}

$$w_{\text{syseq}} = \Re(w) = w_0 e^{\Re(\gamma_2)t}\cos\left(\Im(\gamma_2)t + kx\right),$$

and the real part of y, denoted by y_{syseq}

$$y_{\text{syseq}} = \Re(y) = e^{\Re(\gamma_2)t}\left(\Re(y_0)\cos\left(\Im(\gamma_2)t + kx\right) - \Im(y_0)\sin\left(\Im(\gamma_2)t + kx\right)\right).$$

8.1.3. *Numerical comparison of w.* We take $k = 2$. We choose $w_0 = 1$, and we take y_0 such as (w_0, y_0) belong to the kernel of the matrix $-\frac{1}{\Delta t}R(\omega) - ikA(Y,\omega) - \Delta t k^2 B(Y,\omega)$.

We denote w_{LB} the solution given by the $D1Q2$ model with the initialization

$$\begin{pmatrix} w_{\text{LB}} \\ y_{\text{LB}} \end{pmatrix}(x,0) = \begin{pmatrix} w_0 \\ y_0 \end{pmatrix}\cos(kx).$$

We compute the relative L^2 error between the solution of the equivalent equation w_{eqeq} and the solution given by the $D1Q2$ model w_{LB} at the final time

$$\sqrt{\frac{\sum_{i=0}^{Nx}\left(w_{\text{LB}}^{i,Nt} - w_{\text{eqeq}}^{i,Nt}\right)^2}{\sum_{i=0}^{Nx}\left(w_{\text{LB}}^{i,Nt}\right)^2}},$$

and the relative L^2 error between the solution of the equivalent system w_{syseq} and w_{LB}

$$\sqrt{\frac{\sum_{i=0}^{Nx}\left(w_{\text{LB}}^{i,Nt} - w_{\text{syseq}}^{i,Nt}\right)^2}{\sum_{i=0}^{Nx}\left(w_{\text{LB}}^{i,Nt}\right)^2}}.$$

We compute the solution for different amounts of time steps $Nt = 16, 32, 64, 128, 256, 512, 1024$ and 2048, which gives us different time steps $\Delta t = \frac{T}{Nt}$, with $T = \pi$.

We obtain the relative errors of Figure 1, for different relaxation parameters ω.

The equivalent equation and the equivalent system both converge at the order 2 toward the solution given by the $D1Q2$ model. When $\omega \in [1.8, 2]$, the equivalent equation and the equivalent

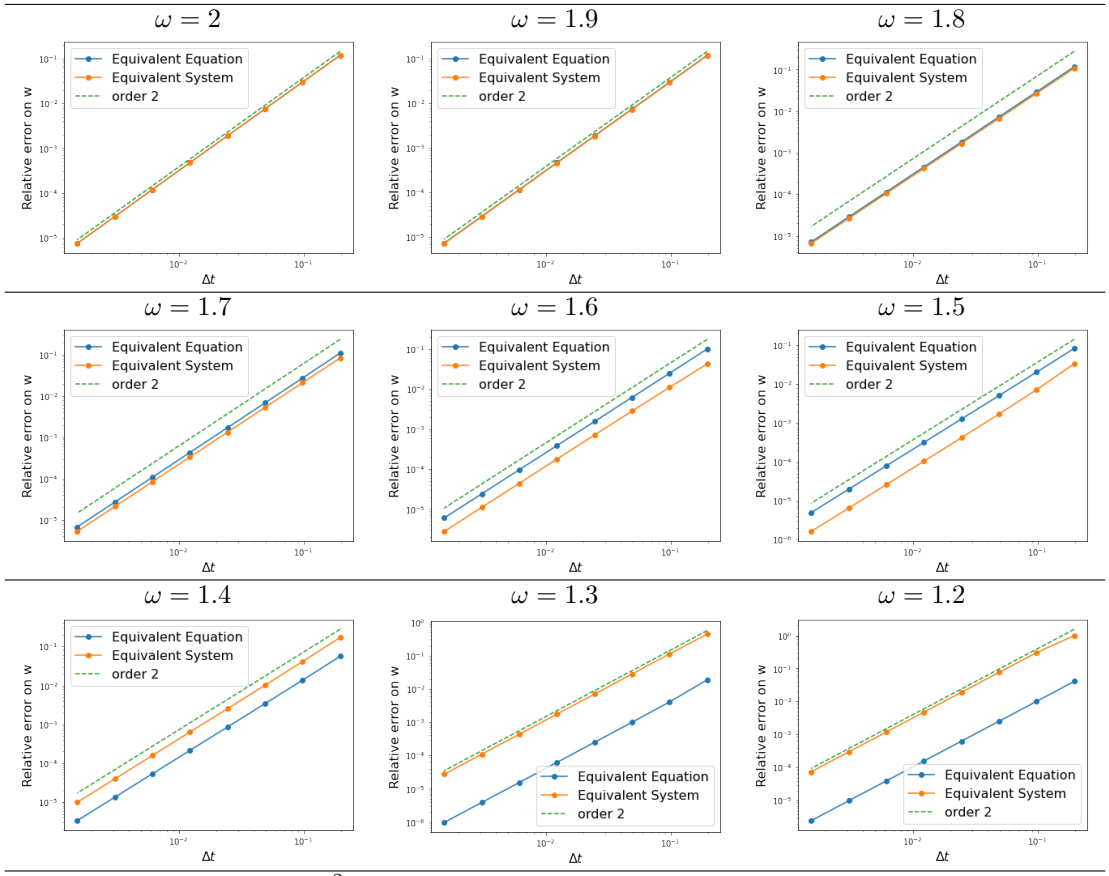


TABLE 1. Relative L^2 error on w with respect to the time step Δt, for different relaxation parameters ω.

system give similar accuracy. When $\omega \in [1.5, 1.8]$, the equivalent system is a better approximation of the solution given by the $D1Q2$ model, while when $\omega \leqslant 1.4$, the equivalent equation is more accurate.

8.1.4. *Numerical comparison of y.* Now, we want to compute the error on the flux error y.

We can compute the relative L^2 errors between y_{LB} and the flux error y_{eqeq} that we assume to have in order to compute the equivalent equation and between y_{LB} and the solution of the equivalent system y_{syseq}

$$\sqrt{\frac{\sum_{i=0}^{Nx}\left(y_{\text{LB}}^{i,Nt}-y_{\text{eqeq}}^{i,Nt}\right)^2}{\sum_{i=0}^{Nx}\left(y_{\text{LB}}^{i,Nt}\right)^2}} \quad \text{and} \quad \sqrt{\frac{\sum_{i=0}^{Nx}\left(y_{\text{LB}}^{i,Nt}-y_{\text{syseq}}^{i,Nt}\right)^2}{\sum_{i=0}^{Nx}\left(y_{\text{LB}}^{i,Nt}\right)^2}}.$$

We obtain the Figure 2. We can observe that the flux error y_{eqeq} given by the equivalent equation converges at the order 1 toward the y_{LB} given by the $D1Q2$ model, while the y_{syseq} given by the equivalent system converges at the order 2.

Remark 23. When $\omega = 2$, the error between the flux error y given by the equivalent equation and the one given by the $D1Q2$ model is constant. This is due to the fact that y_{eqeq} is given by (8.2), which is equal to 0 when $\omega = 2$. Indeed, when $\omega = 2$, w and y are independent, so we do not have to assume the smallness hypothesis $y = O(\Delta t)$ to deduce the equivalent equation from the equivalent system.

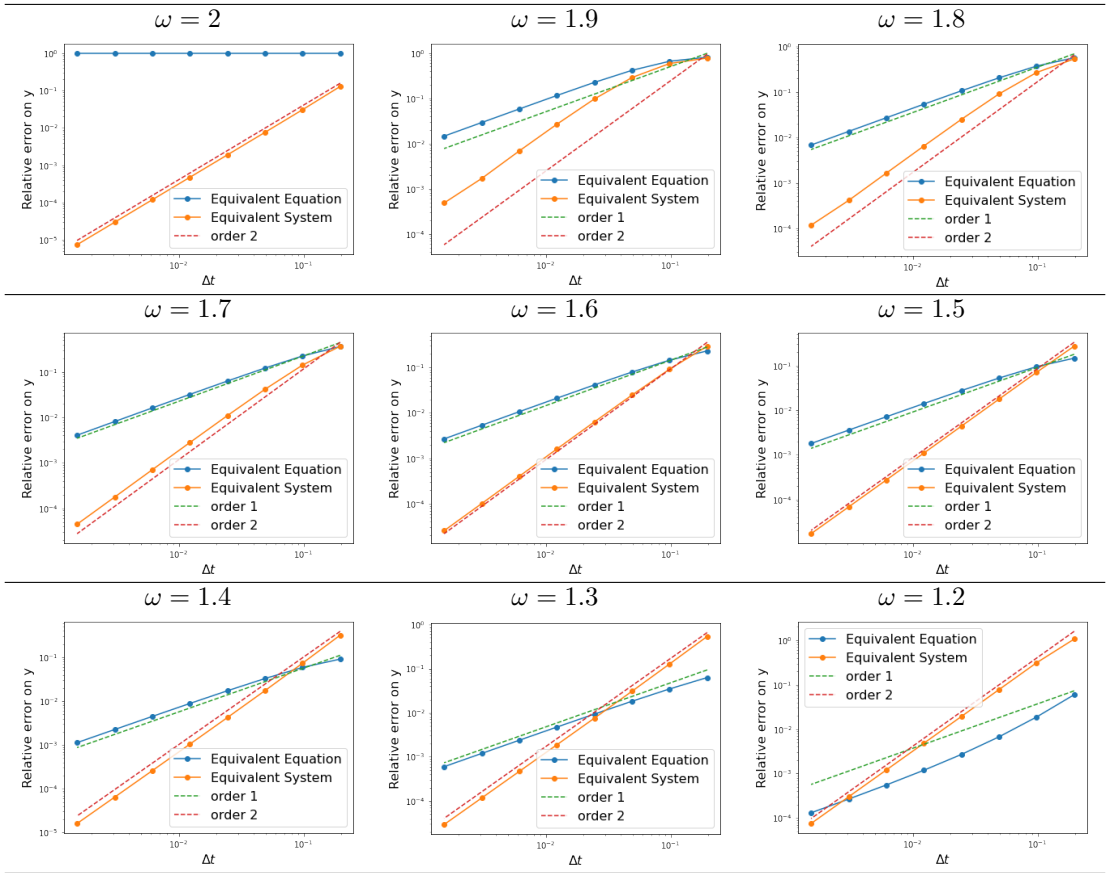


TABLE 2. Relative L^2 error on y with respect to the time step Δt, for different relaxation parameters ω.

8.2. **D2Q4: numerical stability.** As we can see in Figure 7.1, for some choice of velocity $(v_1, v_2) = (a, b)$ and norm of the kinetic velocity λ, the diffusive stability condition can be satisfied, but not the hyperbolicity condition. We want to test numerically what happened when we are in this case.

We consider a square geometry $[0, 1] \times [0, 1]$ with periodic boundary conditions. We consider $Nx = 200$ space steps in both directions. We denote by $\Delta x = 1/Nx$ the grid step. We initialized w with a Gaussian function centered in the middle of the square

$$w(x, y, 0) = e^{-80((x-0.5)^2+(y-0.5)^2)}.$$

Let us choose $(a, b) = (1, 0)$. The stability condition is satisfied if

$$\lambda > \sqrt{2(a^2 + b^2)} = \sqrt{2}.$$

The hyperbolicity condition is satisfied if

$$\lambda > 2 \max(|a|, |b|) = 2.$$

We are going to compare the solution obtained with $\lambda = 1.6$, that is when the diffusive stability condition is satisfied, but not the hyperbolicity condition, and $\lambda = 2.2$, namely when both the diffusive stability and the hyperbolicity conditions are satisfied.

We draw the solutions $w(x, y, T)$ at time $T = 1$, for different values of the relaxation parameter: $\omega = 2$, $\omega = 1.6$ and $\omega = 2$.

As we are solving the transport step of time step $\frac{\Delta t}{4}$ with a Lattice-Boltzmann method, we need to have the relation between the time and space step

$$\Delta t = \frac{4\Delta x}{\lambda} = \frac{4}{\lambda Nx}.$$

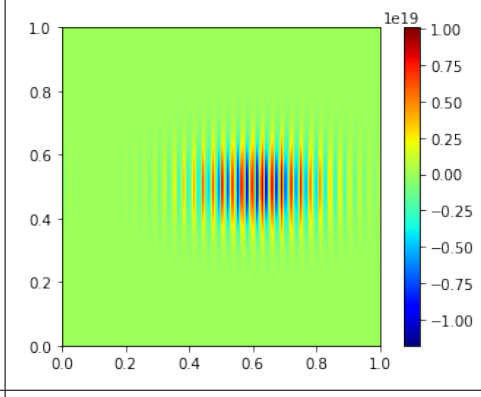
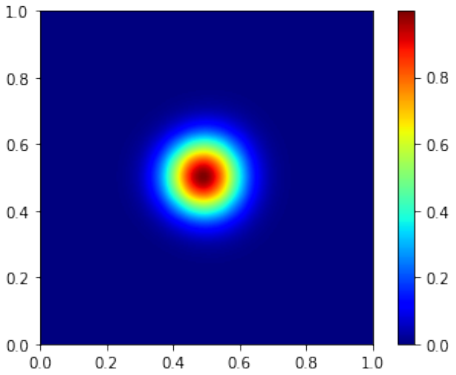
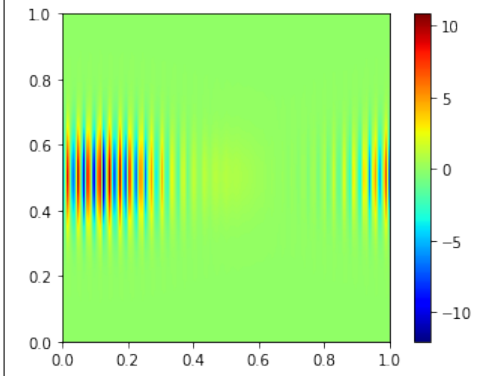
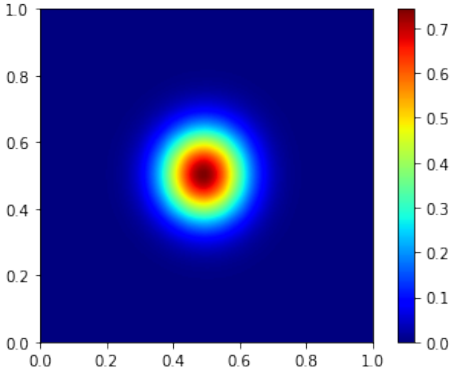
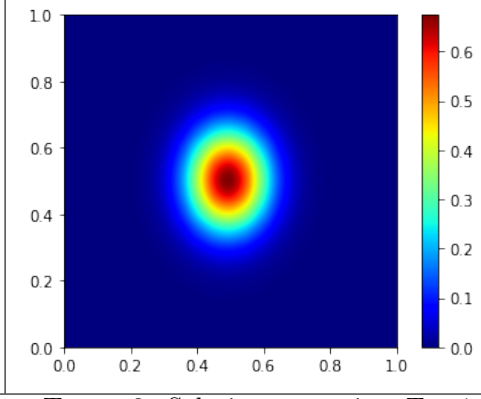
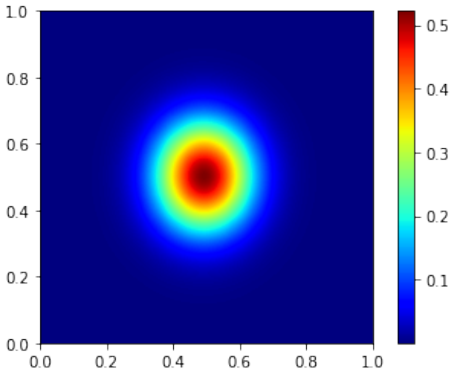
	$\lambda = 1.6$	$\lambda = 2.2$
$\omega = 2$		
$\omega = 1.6$		
$\omega = 1.2$		

TABLE 3. Solutions w at time $T = 1$ for different values of λ and ω.

Consequently, the number of time step is

$$Nt = \frac{T}{\Delta t} = \frac{\lambda N x}{4},$$

and which depends on the λ chosen: we do $Nt = 80$ time steps when $\lambda = 1.6$ and $Nt = 110$ steps when $\lambda = 2.2$.

We obtained the solution w of the Table 3. When $\lambda = 2.2$, that is to say when both the diffusive stability and the hyperbolicity conditions are verified, we obtained a Gaussian centered in the middle of the square, as expected. However, the closer the relaxation parameter ω is to 1, the more the Gaussian function dampens due to the relaxation step. When $\omega = 2$ or $\omega = 1.6$, the solutions obtained with $\lambda = 1.6$, namely when the diffusive stability condition is satisfied but not the hyperbolicity condition, are not stable. Oscillations appear and grow over time. When $\omega = 1.2$ and $\lambda = 1.6$, we obtained a solution close to the expected Gaussian function, but a little distorted. Moreover, this solution is stable, we do not observe any oscillations.

9. Conclusion

In this work we have provided a general methodology for studying the stability and the consistency of the Vectorial Lattice-Boltzmann Method (VLBM). We have first shown that the dual entropy analysis of [3, 16] can be applied for a direct and rigorous proof of the stability of the over-relaxed time splitting algorithm. It is not necessary to pass through the stiff relaxation intermediary.

Secondly, we have proposed an automatic way to construct an equivalent system of PDE, consistent with the VLBM. This equivalent system contains stiff terms in Δt. The classical equivalent equation can be derived from the equivalent system by a Chapman-Enskog analysis when Δt is small and the kinetic data close to equilibrium. It seems, but it is a conjecture, that the hyperbolicity condition of the equivalent system is exactly the entropy stability condition.

In future works, we plan to investigate this conjecture, perform additional numerical experiments on truly non-linear systems, beyond the simple transport equation. Finally, an important question is to extend the stability analysis for handling boundary conditions correctly (see [22] for preliminary results).

References

[1] Denise Aregba-Driollet and Roberto Natalini. Discrete kinetic schemes for multidimensional systems of conservation laws. *SIAM Journal on Numerical Analysis*, 37(6):1973–2004, 2000.

[2] Hubert Baty, Florence Drui, Philippe Helluy, Emmanuel Franck, Christian Klingenberg, and Lukas Thanhäuser. A robust and efficient solver based on kinetic schemes for magnetohydrodynamics (mhd) equations. *Applied Mathematics and Computation*, 440:127667, 2023.

[3] François Bouchut. Construction of BGK models with a family of kinetic entropies for a given system of conservation laws. *Journal of Statistical Physics*, 95(1-2):113–170, 1999.

[4] François Bouchut. Stability of relaxation models for conservation laws. In *European Congress of Mathematics*, pages 95–101. Eur. Math. Soc., 2005.

[5] Françoise Bourdel, Pierre-Alain Mazet, Jean-Pierre Croisille, and Philippe Delorme. On the approximation of k-diagonalizable hyperbolic systems by finite elements-applications to the euler equations and to gaseous mixtures. *La Recherche Aerospatiale (English Edition)(ISSN 0379-380X*, 5:15–34, 1989.

[6] Sydney Chapman and Thomas George 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.

[7] David Coulette, Emmanuel Franck, Philippe Helluy, Michel Mehrenberger, and Laurent Navoret. High-order implicit palindromic discontinuous galerkin method for kinetic-relaxation approximation. *Computers & Fluids*, 190:485–502, 2019.

[8] Clémentine Courtès, David Coulette, Emmanuel Franck, and Laurent Navoret. Vectorial kinetic relaxation model with central velocity. application to implicit relaxations schemes. *Communications in Computational Physics*, 27(4), 2020.

[9] Jean-Pierre Croisille. *Contribution à l'étude théorique et à l'approximation par éléments finis du système hyperbolique de la dynamique des gaz multidimensionnelle et multiespèces*. PhD thesis, Paris 6, 1990.

[10] Camillo De Lellis and László Székelyhidi Jr. The h-principle and the equations of fluid dynamics. *Bulletin of the American Mathematical Society*, 49(3):347–375, 2012.

[11] Paul J Dellar. Bulk and shear viscosities in lattice boltzmann equations. *Physical Review E*, 64(3):031203, 2001.

[12] Paul J Dellar. Lattice kinetic schemes for magnetohydrodynamics. *Journal of Computational Physics*, 179(1):95–126, 2002.

[13] Paul J Dellar. An interpretation and derivation of the lattice boltzmann method using strang splitting. *Computers & Mathematics with Applications*, 65(2):129–141, 2013.

[14] Suresh M Deshpande. A second-order accurate kinetic-theory-based method for inviscid compressible flows. Technical report, 1986.

[15] François Dubois. Equivalent partial differential equations of a lattice boltzmann scheme. *Computers & Mathematics with Applications*, 55(7):1441–1449, 2008.

[16] François Dubois. Simulation of strong nonlinear waves with vectorial lattice boltzmann schemes. *International Journal of Modern Physics C*, 25(12):1441014, 2014.

[17] Ivar Ekeland. Legendre duality in nonconvex optimization and calculus of variations. *SIAM Journal on Control and Optimization*, 15(6):905–934, 1977.

[18] Tony Février. *Extension et analyse des schémas de Boltzmann sur réseau: les schémas à vitesse relative*. PhD thesis, Paris 11, 2014.

[19] Ulrik S Fjordholm, Roger Käppeli, Siddhartha Mishra, and Eitan Tadmor. Construction of approximate entropy measure-valued solutions for hyperbolic systems of conservation laws. *Foundations of Computational Mathematics*, 17:763–827, 2017.

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

[21] Amiram Harten. On the symmetric form of systems of conservation laws with entropy. *Journal of computational physics*, 49, 1983.

[22] Romane Hélie. *Schéma de relaxation pour la simulation de plasmas dans les tokamaks*. PhD thesis, Strasbourg, 2023.
[23] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.
[24] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media, 2013.
[25] Peter Lax. Shock waves and entropy. In *Contributions to nonlinear functional analysis*, pages 603–634. Elsevier, 1971.
[26] Peter D Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*. SIAM, 1973.
[27] Michael S Mock. Systems of conservation laws of mixed type. *Journal of Differential equations*, 37(1):70–88, 1980.
[28] Benoît Perthame. Boltzmann type schemes for gas dynamics and the entropy property. *SIAM Journal on Numerical Analysis*, 27(6):1405–1421, 1990.
[29] Laure Saint-Raymond. *Hydrodynamic limits of the Boltzmann equation*. Number 1971. Springer Science & Business Media, 2009.