COMMUN. MATH. SCI.
Vol. 20, No. 2, pp. 297–326

SOME PRELIMINARY RESULTS ON A HIGH ORDER ASYMPTOTIC PRESERVING COMPUTATIONALLY EXPLICIT KINETIC SCHEME*

RÉMI ABGRALL† AND DAVIDE TORLO‡

Abstract. In this short paper, we intend to describe one way to construct arbitrarily high order kinetic schemes on regular meshes. The method can be arbitrarily high order in space and time, run at least CFL one, is asymptotic preserving and computationally explicit, i.e., the computational costs are of the same order of a fully explicit scheme. We also introduce a nonlinear stability method that enables to simulate problems with discontinuities, and it does not kill the accuracy for smooth regular solutions.

Keywords. Kinetic scheme; asymptotic preserving; high order; stability analysis.

AMS subject classifications. 65M12; 65L04; 65M60.

1. Introduction

Let us specify first the context. We are given the PDE

$$\frac{\partial \mathbf{u}}{\partial t}+\frac{\partial \mathbf{f}(\mathbf{u})}{\partial x}=0 \tag{1.1a}$$

with the initial condition

$$\mathbf{u}(x,0)=\mathbf{u}_0(x), \tag{1.1b}$$

with $\mathbf{u}\in\mathbb{R}^p$ and $\mathbf{f}:\mathbb{R}^p\to\mathbb{R}^p$ a Lipschitz continuous flux. It is known, at least since the work of Jin [13] and then Natalini [17] and co-workers, that this system can formally be seen as the limit for $\varepsilon\to 0$ of a relaxation system:

$$\frac{\partial \mathbf{F}}{\partial t}+\Lambda\frac{\partial \mathbf{F}}{\partial x}=\frac{\mathbb{M}(\mathbb{P}\mathbf{F})-\mathbf{F}}{\varepsilon} \tag{1.2a}$$

with $\mathbf{F}\in\mathbb{R}^{k\times p}$, $\mathbb{M}$ is a Maxwellian and $\mathbb{P}$ is a linear operator such that $\mathbb{P}\mathbb{M}(\mathbb{P}\mathbf{F})=\mathbb{P}\mathbf{F}$. The constant matrix Λ and the flux $\mathbf{f}$ are linked by $\mathbb{P}\Lambda\mathbb{M}(\mathbb{P}\mathbf{F})=\mathbf{f}(\mathbb{P}\mathbf{F})$. The simplest example, due to Jin and Xin [13], is

$$\begin{aligned}&\frac{\partial u}{\partial t}+\frac{\partial v}{\partial x}=0\\&\frac{\partial v}{\partial t}+a^2\frac{\partial u}{\partial x}=\frac{f(u)-v}{\varepsilon}\end{aligned}$$

that can be rewritten in the form (1.2) with:

$$\begin{aligned}&\frac{\partial f_1}{\partial t}+a\frac{\partial f_1}{\partial x}=\frac{\mathbb{M}_1-f_1}{\varepsilon},\\&\frac{\partial f_2}{\partial t}-a\frac{\partial f_2}{\partial x}=\frac{\mathbb{M}_2-f_2}{\varepsilon},\end{aligned} \tag{1.3}$$

*Received: May 28, 2019; Accepted (in revised form): June 30, 2021. Communicated by Shi Jin.

†Institute of Mathematics and Institute of Computational Sciences, Universität Zürich, Winthertur-erstrasse 190, Zürich, Switzerland (remi.abgrall@math.uzh.ch).

‡Inria Bordeaux-Sud-Ouest, 200 avenue de la vieille tour, 33405 Talence, France (davide.torlo@inria.fr). https://davidetorlo.it/

i.e., where

$$\mathbf{F}=\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}, \quad \Lambda=\begin{pmatrix} a & 0 \\ 0 & -a \end{pmatrix}, \quad \mathbb{P}\mathbf{F}=f_1+f_2 \text{ and } \mathbb{M}=\begin{pmatrix} \mathbb{M}_1 \\ \mathbb{M}_2 \end{pmatrix}$$

where the Maxwellian is defined from the relations

$$\mathbb{M}_1+\mathbb{M}_2=f_1+f_2=u, \qquad a(\mathbb{M}_1-\mathbb{M}_2)=f(u),$$

i.e.

$$\mathbb{M}_1(f,a)=\frac{1}{2}\left(f_1+f_2+\frac{f(u)}{a}\right), \mathbb{M}_2(f,a)=\frac{1}{2}\left(f_1+f_2-\frac{f(u)}{a}\right).$$

We know that a must be larger than the max of $|f'(u)|$ because of the Whitham sub-characteristic condition, obtained via a formal Chapman Enskog expansion. Another argument is, as shown by [6], that under this condition the two Maxwellians $\mathbb{M}_1$ and $\mathbb{M}_2$ satisfy a monotonicity condition, i.e. the BGK model becomes compatible with entropy inequalities.

The questions we address in this paper are the following: Given a system (1.1) and a regular grid of spatial step $\Delta x>0$, can we construct a computationally explicit scheme that solves (1.2) with uniform accuracy of order $r>0$ for all $\varepsilon>0$ and with a CFL condition, based on the matrix Λ, that is larger than 1[1]. The answer is yes, and this paper proposes a simple construction in one dimension. By *computationally explicit* we mean that the solution of a certain scheme does not require any nonlinear solver, nor the inversion of a mass matrix.

High order accurate methods for kinetic problems à la Shi-Jin has received a lot of attention in the recent years. For a long time the state of the art was that of second order in time and space finite volume with a TVD-like stabilisation, see e.g [3]. For higher than second order, one may mention [15] where a splitting approach is adopted with a regular CFL stability condition for the overall finite volume scheme, [19] where relaxed upwind schemes are proposed running up to CFL $=1$ and up to third order in time/space, again in a finite volume context. In [4], a WENO approach is proposed. In [9] a discontinuous Galerkin approximation of the system (1.2) is developed (with a temporal scheme allowing very large CFL number). In the kinetic literature, where the fluid system is represented with the BGK approximation, so that dense and less dense flows can be simulated, there has also been a large effort towards high order schemes with asymptotic preserving properties. One may mention [5] for hyperbolic systems with diffusion, [23] where a high order conservative semi-Lagrangian technique is developed.

We want to go beyond that, with very simple and cheap numerical schemes that are potentially arbitrary high order and run at CFL $=1$, with an accuracy that is independent of the relaxation parameter ε. The format of the paper is as follows. We first introduce the general method which amounts to describing the discretisation of $\Lambda\dfrac{\partial \mathbf{F}}{\partial x}$ and a time discretisation. We take into account the source term. The scheme resulting from this discretisation is fully implicit. The next step is to show that, thanks to the operator $\mathbb{P}$, and using a particular time discretisation, we can make it computationally

[1]Initially, the first author was motivated by understanding in a better way the LBM method, even though the answer is not about the LBM method at all. The only remaining property between what we look for and the LBM method is the CFL condition.

explicit, and high order accurate, independently of the parameter ε. Several choices of Λ and Maxwellians $\mathbb{M}$ are described. We also address the question of the nonlinear stabilisation of the method when discontinuities appear. Several numerical examples, covering scalar and system cases, are then proposed to show the relevance of the method. The accuracy is checked for the scalar case.

2. General discretisation principle

Starting from (1.2), the idea is to discretise first in space $\Lambda\dfrac{\partial \mathbf{F}}{\partial x}$. This introduces an error which we assume to be $O(\Delta x^q)$,

$$\frac{\partial \mathbf{F}}{\partial t}+\frac{1}{\Delta x}\Lambda\delta\mathbf{F}=\frac{\mathbb{M}(\mathbb{P}\mathbf{F})-\mathbf{F}}{\varepsilon}+O(\Delta x^q). \tag{2.1}$$

The second step is to discretise in time, so that we expect that the resulting scheme will be of order p in space and time, at least for moderate values of ε. The problem is then two-fold: (i) how to define the discretisation operator δ for which a minimum requirement is the semi-discrete linear stability when there is no source term, (ii) how to discretise in time so that the accuracy is uniform in time and ε. We first discuss the issue of time discretisation, then space discretisation.

2.1. Time discretisation. One may use IMEX Runge-Kutta schemes, and more precisely SSP IMEX Runge-Kutta schemes, to have a better control of the stability properties of the method. Rewriting (1.2a) as the sum of a non-stiff term and a stiff one

$$\frac{dU}{dt}+\mathcal{F}(U)=\frac{\mathcal{G}(U)}{\varepsilon} \tag{2.2}$$

an IMEX method is defined by two Butcher's tableaux

$$\begin{array}{c|c} c & A \\ \hline 0 & b^T \end{array} \quad \text{and} \quad \begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline 0 & \tilde{b}^T \end{array}$$

where the first one is for the non-stiff part, while the second one is for the stiff part:

$$\begin{aligned} U_0&=U^n\\ &\vdots\\ U_k&=U_0+\Delta t\sum_{j=1}^{k-1}a_{kj}\mathcal{F}(U_j)+\frac{\Delta t}{\varepsilon}\sum_{j=1}^{s}\tilde{a}_{kj}\mathcal{G}(U_j)\\ &\vdots\\ U^{n+1}&=U^n+\Delta t\sum_{j=1}^{s}b_j\mathcal{F}(U_j)+\frac{\Delta t}{\varepsilon}\sum_{j=1}^{s}\tilde{b}_j\mathcal{G}(U_j) \end{aligned} \tag{2.3}$$

with various compatibility conditions so that a given order is reached, see [11, Chapter IV]. Anticipating a bit, if there exists a linear operator $\mathbb{P}$ such that $\mathbb{P}\mathcal{G}=0$ as here, we see that, applying $\mathbb{P}$ to (2.3), a necessary condition is that the explicit RK scheme defined by the explicit part is itself SSP. Since we want to have a running CFL number of at least one, this needs that the SSP RK scheme must have a CFL number of at least $1+\epsilon$, $\epsilon>0$. To our knowledge there are some explicit SSP RK schemes satisfying this condition, inter alia [20], but they are not generalizable to arbitrarily high order of accuracy, and no IMEX versions are available.

For this reason, we use an IMEX deferred correction (DeC) method. It is a general way of building arbitrarily high order Runge Kutta schemes. It also allows more freedom in the spatial discretization, for instance, allowing the use of lumped mass matrix [1]. Its implicit and IMEX versions allow to use a combination of more traditional low order IMEX schemes and arbitrarily high order implicit RK schemes, obtaining arbitrarily high order IMEX schemes. We leave the study of SSP version of these schemes for future research. The final IMEX DeC scheme we obtain is computationally explicit and it is also matrix-free.

2.1.1. Deferred correction. The DeC is an iterative procedure that was proposed and developed in its explicit version in [10] and in an implicit version in [16]. It was applied to hyperbolic PDE, for instance, in [1], with a new formalism which makes the proof of its properties more straightforward. An IMEX version of this algorithm applied to hyperbolic PDE is available in [2], and the algorithm we discuss in the following is a modification of this one. With the notation of [1], the DeC uses two operators: one high order accurate $\mathcal{L}^2$, which defines a fully implicit method, and a low order easy to solve $\mathcal{L}^1$ operator. The process allows to approximate with arbitrary accuracy the solution of the high order operator $\mathcal{L}^2$, with the simplicity of the operator $\mathcal{L}^1$. We start with the description of the high order operator $\mathcal{L}^2$.

Let us consider $q+1$ points in $[0,1]$, $c_0=0<c_1\ldots<c_i<\ldots<c_q=1$ and the quadrature formula

$$\int_{t_n}^{t_n+c_i\Delta t}\varphi(s)\,ds\approx\Delta t\sum_{j=0}^{q}a_{ij}\varphi(t_n+c_j\Delta t).$$

More precisely, if $\{\ell_j\}$ are the Lagrange polynomials associated to the partition $\{c_j\}_{j=0}^q$, if we take

$$a_{ij}=\int_0^{c_i}\ell_j(s)ds,$$

the quadrature formula is of order $q+1$. We will always require that the quadrature formula is consistent, i.e.

$$\sum_{j=0}^{q}a_{ij}=c_i. \tag{2.4}$$

Considering x_k, a grid point, and setting $\mathbf{F}_k^{n,j}\approx\mathbf{F}(x_k,t_n+c_j\Delta t)$ and $\mathbf{F}_k^{n,0}=\mathbf{F}_k^n$, an approximation of (1.2) is:

$$\mathbf{F}_k^{n,j}-\mathbf{F}_k^{n,0}+\frac{\Delta t}{\Delta x}\left(\sum_{l=0}^{q}a_{il}\Lambda\delta_k\mathbf{F}^{n,l}\right)-\mu\sum_{l=0}^{q}a_{il}\left(\mathbb{MP}\mathbf{F}_k^{n,l}-\mathbf{F}_k^{n,l}\right)=0,\quad j=1,\ldots,q \tag{2.5}$$

where $\mu=\frac{\Delta t}{\varepsilon}$ and $\frac{\delta\mathbf{F}}{\Delta x}$ is a consistent approximation of $\dfrac{\partial\mathbf{F}}{\partial x}$. We will set $\mathbf{F}_k^{n+1}=\mathbf{F}_k^{n,q}$. The relations (2.5) can be rewritten in matrix form, setting

$$\mathcal{F}_k=\left(\mathbf{F}_k^{n,1},\ldots,\mathbf{F}_k^{n,q}\right)^T,\quad \mathcal{F}_k^{(0)}=\left(\mathbf{F}_k^{n,0},\ldots,\mathbf{F}_k^{n,0}\right)^T=\left(\mathbf{F}_k^{n},\ldots\mathbf{F}_k^{n}\right),$$

and neglecting the index of the timestep n, as

$$\mathcal{F}_k - \mathcal{F}_k^{(0)} + \frac{\Delta t}{\Delta x}\Lambda A \delta_k \mathcal{F} - \mu A\big(\mathbb{M}(\mathbb{P}\mathcal{F}_k) - \mathcal{F}_k\big) + \frac{\Delta t}{\Delta x}\Lambda \mathbf{a}_0 \otimes \delta_k \mathbf{F}^{n,0} - \mu \mathbf{a}_0 \otimes (\mathbb{M}(\mathbb{P}\mathbf{F}_k^{n,0}) - \mathbf{F}_k^{n,0}) = 0, \tag{2.6}$$

where, by abuse of language, we have written

$$\mathbb{M}(\mathbb{P}\mathcal{F}) = \big(\mathbb{M}(\mathbb{P}\mathbf{F}^1), \ldots, \mathbb{M}(\mathbb{P}\mathbf{F}^q)\big)^T.$$

The matrix A is

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1q} \\ \vdots & \vdots & \vdots \\ a_{q1} & \ldots & a_{qq} \end{pmatrix}$$

and we have

$$\mathbf{a}_0 = \begin{pmatrix} a_{0q} \\ \vdots \\ a_{01} \end{pmatrix}.$$

As a result, (2.6) is implicit, and in general nonlinear, because of the Maxwellian. In order to simplify the resolution, we consider a simpler scheme, where the source term discretisation remains the same and the forward Euler method is used on each sub-time step:

$$\mathbf{F}_k^{n,j} - \mathbf{F}_k^n + c_j \frac{\Delta t}{\Delta x}\Lambda \delta_k \mathbf{F}^{n,0} - \mu \sum_{l=0}^{q} a_{jl}\big(\mathbb{M}(\mathbb{P}\mathbf{F}_k^{n,l}) - \mathbf{F}_k^{n,l}\big) = 0, \quad j = 1, \ldots, q. \tag{2.7}$$

We rewrite this as:

$$\mathcal{F}_k - \mathcal{F}_k^{(0)} + \frac{\Delta t}{\Delta x} C\Lambda \delta_k \mathcal{F}^{(0)} - \mu A\big(\mathbb{M}(\mathbb{P}\mathcal{F}_k) - \mathcal{F}_k\big) - \mu \mathbf{a}_0 \otimes (\mathbb{M}\mathbb{P}\mathbf{F}_k^{n,0} - \mathbf{F}_k^{n,0}) = 0 \tag{2.8}$$

where $C = \operatorname{diag}\big(c_1, \ldots, c_q\big)$ and $\mathcal{F}^{(0)} = (\mathbf{F}^{n,0}, \ldots, \mathbf{F}^{n,0})^T$.

This leads to the introduction of two operators $\mathcal{L}^1$ and $\mathcal{L}^2$ acting on $\mathcal{F} = (\ldots, \mathcal{F}_k, \mathcal{F}_{k+1}, \ldots)$ and defined as:

$$\big[\mathcal{L}^1(\mathcal{F})\big]_k := \mathcal{F}_k - \mathcal{F}_k^{(0)} + \frac{\Delta t}{\Delta x} C\Lambda \delta_k \mathcal{F}^{(0)} - \mu A\big(\mathbb{M}(\mathbb{P}\mathcal{F}_k) - \mathcal{F}_k\big) - \mu \mathbf{a}_0 \otimes (\mathbb{M}(\mathbb{P}\mathbf{F}_k^{n,0}) - \mathbf{F}_k^{n,0}),$$

and

$$\big[\mathcal{L}^2(\mathcal{F})\big]_k := \mathcal{F}_k - \mathcal{F}_k^{(0)} + \frac{\Delta t}{\Delta x}\Lambda A \delta_k \mathcal{F} - \mu A\big(\mathbb{M}(\mathbb{P}\mathcal{F}_k) - \mathcal{F}_k\big) + \frac{\Delta t}{\Delta x}\Lambda \mathbf{a}_0 \otimes \delta_k \mathbf{F}^{n,0} - \mu \mathbf{a}_0 \otimes (\mathbb{M}(\mathbb{P}\mathbf{F}_k^{n,0}) - \mathbf{F}_k^{n,0}). \tag{2.9}$$

So that (2.8) is $\mathcal{L}^1(\mathcal{F}^{n,j})_k = 0$ while (2.6) is $\mathcal{L}^2(\mathcal{F}^{n,j})_k = 0$. In order to have more structure, we will require that $\delta_k \mathcal{F}$ has the following difference form:

$$\delta_k \mathcal{F} = \widehat{\mathcal{F}}_{k+1/2} - \widehat{\mathcal{F}}_{k-1/2} \tag{2.10}$$

where $\widehat{\mathcal{F}}_{k+1/2}$ depends on P arguments, is consistent with $\mathcal{F}$ and uniformly Lipschitz continuous with respect to its arguments. Examples will be given in Section 2.2.

Thanks to (2.4), we see that

$$\mathcal{L}^2(\mathcal{F})_k - \mathcal{L}^1(\mathcal{F})_k = \frac{\Delta t}{\Delta x}\Lambda A\big(\delta_k\mathcal{F} - \delta_k\mathcal{F}^{(0)}\big), \tag{2.11}$$

the important fact is that ε plays no role here.

We will solve the problem (2.6) with the following defect correction (DeC) method:

- Set, for any k, $\mathcal{F}_k^{(0)} = (\mathbf{F}_k^n, \dots, \mathbf{F}_k^n)^T$,
- Solve for $p = 0, \dots, M-1$ the problem

$$\mathcal{L}^1(\mathcal{F}^{(p+1)}) = \mathcal{L}^1(\mathcal{F}^{(p)}) - \mathcal{L}^2(\mathcal{F}^{(p)}), \tag{2.12}$$

- Set $\mathcal{F}^{n+1} = \mathcal{F}^{(M)}$.

We remark that the operator $\mathcal{L}^2$ will never be solved directly as it will be applied to the previously computed iteration $\mathcal{F}^{(p)}$. The DeC procedure will converge to the solution of the $\mathcal{L}^2(\mathcal{F}^*) = 0$ problem by solving iteratively (2.12). We show that if the problem (2.6) has a unique solution $\mathcal{F}^\star$ and taking $M = q$, we have a formal error of Δt^q, i.e., for a norm to be defined,

$$\|\mathcal{F}^{(q)} - \mathcal{F}^\star\| \le C\Delta t^q,$$

so that the formal accuracy is the same as solving (2.6) exactly. Before doing that, we first have to explain how we solve for $\mathcal{L}^1$ and, hence, (2.12), then we show the error estimate (and define the proper norm).

2.1.2. Solution of $\mathcal{L}^1(\mathcal{F}) = \mathcal{G}$ and (2.12). Let us first start with $\mathcal{L}^1(\mathcal{F}) = \mathcal{G}$ for any $\mathcal{G}$. Applying $\mathbb{P}$ to this equation, we get, for any $k \in \mathbb{Z}$,

$$\mathbb{P}\mathcal{F}_k = \mathbb{P}\mathcal{G}_k + \mathbb{P}\mathcal{F}_k^{(0)} - \frac{\Delta t}{\Delta x}\mathbb{P}C\Lambda\delta_k\mathcal{F}^{(0)} = \mathbb{P}\mathcal{G}_k + \mathcal{K},$$

with

$$\mathcal{K} = \mathbb{P}\mathcal{F}_k^{(0)} - \frac{\Delta t}{\Delta x}\mathbb{P}C\Lambda\delta_k\mathcal{F}^{(0)}.$$

The found equation is explicit for $\mathbb{P}\mathcal{F}$ and we can in practice compute this term and use it to obtain the solution of the whole operator. Substituting $\mathbb{P}\mathcal{F}$ into the Maxwellian, we obtain

$$\mathcal{F}_k = \mathcal{G}_k + \mathcal{F}_k^{(0)} - \frac{\Delta t}{\Delta x}C\Lambda\delta_k\mathcal{F}^{(0)} + \mu\left(A\mathbb{M}\big(\mathbb{P}\mathcal{G}_k + \mathcal{K}\big) - A\mathcal{F}_k\right) + \mu\mathbf{a}_0 \otimes (\mathbb{M}(\mathbb{P}\mathbf{F}_k^{n,0}) - \mathbf{F}_k^{n,0}),$$

where all the unknown terms $\mathcal{F}$ depend only linearly on some coefficients which we can collect on the left-hand side

$$\big(\mathrm{Id}_{q\times q} + \mu A\big)\mathcal{F}_k = \mathcal{G}_k + \mathcal{F}_k^{(0)} - \frac{\Delta t}{\Delta x}C\Lambda\delta_k\mathcal{F}^{(0)} + \mu A\mathbb{M}\big(\mathbb{P}\mathcal{G}_k + \mathcal{K}\big) + \mu\mathbf{a}_0 \otimes (\mathbb{M}(\mathbb{P}\mathbf{F}_k^{n,0}) - \mathbf{F}_k^{n,0}).$$

Now, if $\mathrm{Id}_{q\times q} + \mu A$ is invertible, we can compute only once and store its inverse to obtain an easy solution of the whole problem, i.e.,

$$\mathcal{F}_k = \big(\mathrm{Id}_{q\times q} + \mu A\big)^{-1}\left(\mathcal{G}_k + \mathcal{F}_k^{(0)} - \frac{\Delta t}{\Delta x}C\Lambda\delta_k\mathcal{F}^{(0)}\right) + \big(\mathrm{Id}_{q\times q} + \mu A\big)^{-1}\mu A\mathbb{M}\big(\mathbb{P}\mathcal{G}_k + \mathcal{K}\big)$$

$$+\left(\mathrm{Id}_{q\times q}+\mu A\right)^{-1}\mu\mathbf{a}_0\otimes(\mathbb{M}(\mathbb{P}\mathbf{F}_k^{n,0})-\mathbf{F}_k^{n,0}). \tag{2.13a}$$

In this way, we are able to find a solution of the system $\mathcal{L}^1(\mathcal{F})=\mathcal{G}$ in a computationally explicit way: the source term is split into the linearly implicit part and the Maxwellian evaluated in the previously computed $\mathbb{P}\mathcal{F}$. The remaining terms are the explicit right-hand side $\mathcal{G}$, the explicit convection term $\delta_k\mathcal{F}^{(0)}$ and the explicit part of the high order time integration of the source. There have been many other works using similar techniques in order to explicitly solve the implicit discretization of the kinetic system (1.1b), inter alia [2, 3, 8, 18]. This DeC approach combining the two operators, has the advantage of being arbitrarily high order without building complicated structures. Indeed, we can use this computationally explicit solution to solve (2.12).

Setting

$$\mathcal{G}_k=\mathcal{L}_1(\mathcal{F}^{(p)})_k-\mathcal{L}_2(\mathcal{F}^{(p)})_k,$$

we can apply (2.13) directly. However, since the source term discretisation is the same, we have simplifications. Indeed, after little algebra, (2.12) is rewritten as:

$$\begin{aligned}\mathcal{F}^{(p+1)}-\mu A\Big(\mathbb{M}(\mathbb{P}\mathcal{F}^{(p+1)})-\mathcal{F}^{(p+1)}\Big)=&\mathcal{F}^{(0)}-\frac{\Delta t}{\Delta x}\Lambda A\delta\mathcal{F}^{(p)}-\frac{\Delta t}{\Delta x}\mathbf{a}_0\otimes\Lambda\delta\mathbf{F}^{n,0}\\&+\mu\mathbf{a}_0\otimes(\mathbb{M}(\mathbb{P}\mathbf{F}^{n,0})-\mathbf{F}^{n,0})\end{aligned} \tag{2.14}$$

and we can apply the same technique. We first apply $\mathbb{P}$,

$$\mathbb{P}\mathcal{F}^{(p+1)}=\mathbb{P}\mathcal{F}^{(0)}-\frac{\Delta t}{\Delta x}\mathbb{P}\Lambda A\delta\mathcal{F}^{(p)}-\frac{\Delta t}{\Delta x}\mathbf{a}_0\otimes\mathbb{P}\Lambda\delta\mathbf{F}^{n,0}, \tag{2.15}$$

so we know explicitly $\mathbb{P}\mathcal{F}^{(p+1)}$, and then we can solve

$$\begin{aligned}\left(\mathrm{Id}_{q\times q}+\mu A\right)\mathcal{F}^{(p+1)}=&\mu A\mathbb{M}(\mathbb{P}\mathcal{F}^{(p+1)})+\mathcal{F}^{(0)}-\frac{\Delta t}{\Delta x}\Lambda A\delta\mathcal{F}^{(p)}-\frac{\Delta t}{\Delta x}\mathbf{a}_0\otimes\Lambda\delta\mathbf{F}^{n,0}\\&+\mu\mathbf{a}_0\otimes(\mathbb{M}(\mathbb{P}\mathbf{F}^{n,0})-\mathbf{F}^{n,0}),\end{aligned}$$

and then

$$\begin{aligned}\mathcal{F}^{(p+1)}=\left(\mathrm{Id}_{q\times q}+\mu A\right)^{-1}\Bigg(&\mu A\mathbb{M}(\mathbb{P}\mathcal{F}^{(p+1)})+\mathcal{F}^{(0)}-\frac{\Delta t}{\Delta x}\Lambda A\delta\mathcal{F}^{(p)}\\&-\frac{\Delta t}{\Delta x}\mathbf{a}_0\otimes\Lambda\delta\mathbf{F}^{n,0}+\mu\mathbf{a}_0\otimes(\mathbb{M}(\mathbb{P}\mathbf{F}^{n,0})-\mathbf{F}^{n,0})\Bigg).\end{aligned} \tag{2.16}$$

Again, we see that the method is computationally explicit.

Next, we show the error estimate, and then we comment more on (2.16), in particular when $\varepsilon\to 0$.

2.1.3. Error estimate. If $\varphi:\mathbb{R}\to\mathbb{R}^p$ is $C^1(\mathbb{R})^p$ and has a compact support, we can consider the discrete version of its L^2 and H^1 norms:

$$\|\varphi\|_{L^2}^2=\sum_{j\in\mathbb{Z}}\Delta x\|\varphi_i\|^2,\quad \|\varphi\|_{H^1}^2=\|\varphi\|_{L^2}^2+\sum_{j\in\mathbb{Z}}\Delta x\|D_i\varphi\|^2$$

where $D_i\varphi=\frac{\varphi_{i+1}-\varphi_i}{\Delta x}$.

We will establish error estimates that are valid in a given but arbitrary compact $I=[a,b]$ with discrete equivalent of L^2_{loc} and H^{-1}_{loc} estimates:

$$\|\mathcal{F}\|_{2,I}=\sup_{\varphi\in C^1_0([a,b])^p}\frac{\sum_j\Delta x\langle\varphi_i,\mathbf{F}_i\rangle}{\|\varphi\|_{L^2}}\text{ and }\|\mathcal{F}\|_{-1,I}=\sup_{\varphi\in C^1_0([a,b])^p}\frac{\sum_j\Delta x\langle\varphi_i,\mathbf{F}_i\rangle}{\|\varphi\|_{H^1}}$$

and we note that for $\varphi\in C^1_0([a,b])^p$, we have a Poincaré-like inequality

$$\|\varphi\|_{2,I}\le(b-a)\|D\varphi\|_{2,I}.$$

We first show that

LEMMA 2.1. *If* $\widehat{\mathcal{F}}_{k+1/2}=\sum_{l=-p}^{q}\alpha_l\mathcal{F}_{k+l}$ *and letting* $C=\max_{-p\le l\le q}|\alpha_l|\times\max_i|\lambda_i|$, *we have*

$$\|\mathcal{L}^2(\mathcal{F})-\mathcal{L}^1(\mathcal{F})\|_{-1,I}\le C\|\mathcal{F}\|_{2,I}\,\Delta t. \tag{2.17}$$

Proof. We have, from (2.11) and since $\delta_k\mathcal{F}=\widehat{\mathcal{F}}_{k+1/2}-\widehat{\mathcal{F}}_{k-1/2}$, we have, using that φ has a compact support,

$$\begin{aligned}\left|\sum_k\Delta x\langle\varphi_k,\mathcal{L}^2_k(\mathcal{F})-\mathcal{L}^1_k(\mathcal{F})\rangle\right|&=\left|\sum_k\Delta t\langle\varphi_k,A\Lambda\big(\widehat{\mathcal{F}}_{k+1/2}-\widehat{\mathcal{F}}_{k-1/2}\big)\rangle\right|\\&=\left|\sum_k\Delta t\Delta x\langle D_{k+1/2}\varphi,A\Lambda\widehat{\mathcal{F}}_{k+1/2}\rangle\right|\\&\le\|A\|\,\|\Lambda\|\,\Delta t\sqrt{\sum_k\Delta x\|D_{k+1/2}\varphi\|^2}\sqrt{\sum_k\Delta x\|\widehat{\mathcal{F}}_{k+1/2}\|^2}\\&\le C\Delta t\|\varphi\|_{H^1}\|\mathcal{F}\|_{2,I}.\end{aligned}$$

We remark that the norm of the coefficients A is smaller or equal to 1. ∎

We also have the following lemma on $\mathcal{L}_1$:

LEMMA 2.2. *We assume that the Maxwellian is Lipschitz continuous and that there exists* $C,C'>0$ *such that for all* $\varepsilon>0$,

$$\|\big(Id_{(q-1)\times(q-1)}+\mu A\big)^{-1}\|\le C,\quad\mu\|\big(Id_{(q-1)\times(q-1)}+\mu A\big)^{-1}A\|\le C'.$$

Let us consider $\mathcal{F},\mathcal{F}'$ *and* $\mathcal{G},\mathcal{G}'$ *such that*

$$\mathcal{L}_1(\mathcal{F})=\mathcal{G}\text{ and }\mathcal{L}_1(\mathcal{F}')=\mathcal{G}'.$$

Then, there exists $\alpha>0$, *independent of* $\mathcal{F},\mathcal{F}'$, ε *and* I *such that*

$$\|\mathcal{F}-\mathcal{F}'\|_{2,I}\le\alpha\|\mathcal{G}-\mathcal{G}'\|_{2,I}$$

and

$$\|\mathcal{F}-\mathcal{F}'\|_{-1,I}\le\alpha\|\mathcal{G}-\mathcal{G}'\|_{-1,I}$$

Proof. We have the explicit solution $\mathcal{F}$ and $\mathcal{F}'$ from (2.13), and we see that

$$\big(\mathrm{Id}_{(q-1)\times(q-1)}+\mu A\big)\Big(\mathcal{F}_k-\mathcal{F}'_k\Big)=\mathcal{G}_k-\mathcal{G}'_k+\mu A\mathbb{M}\big(\mathbb{P}\mathcal{G}_k-\mathbb{P}\mathcal{G}'_k\big),$$

so that if $\|\ .\ \|$ is any of the two norms, we have

$$\|\mathcal{F}_k - \mathcal{F}'_k\| \leq \alpha \|\mathcal{G}_k - \mathcal{G}'_k\|.$$

The constant α depends on C, C', Λ and the Lipschitz constant of the Maxwellian. Remember also that in (2.13), $\mathcal{K}$ depends only on $\mathcal{F}_0$, $\frac{\Delta t}{\Delta x}$ and Λ. It is independent of ε. ∎

Then, wrapping all together, we have the following proposition:

PROPOSITION 2.1. *Under the assumptions of Lemmas 2.1 and 2.2, if $\mathcal{F}^\star$ is the unique solution of $\mathcal{L}^2(\mathcal{F}) = 0$, there exists θ independent of ε such that we have, for all $p \in \mathbb{N}$*

$$\|\mathcal{F}^{(p+1)} - \mathcal{F}^\star\|_{L^2} \leq \left(\theta \Delta t\right)^{p+1} \|\mathcal{F}^{(0)} - \mathcal{F}^\star\|_{L^2}. \tag{2.18}$$

Proof. We first have, since $\mathcal{L}^2(\mathcal{F}^\star) = 0$

$$\begin{aligned}\mathcal{L}^1(\mathcal{F}^{(p+1)}) - \mathcal{L}^1(\mathcal{F}^\star) &= \left(\mathcal{L}^1(\mathcal{F}^{(p)}) - \mathcal{L}^1(\mathcal{F}^\star)\right) - \mathcal{L}^2(\mathcal{F}^{(p)}) \\ &= \left(\mathcal{L}^1(\mathcal{F}^{(p)}) - \mathcal{L}^1(\mathcal{F}^\star)\right) - \left(\mathcal{L}^2(\mathcal{F}^{(p)}) - \mathcal{L}^2(\mathcal{F}^\star)\right),\end{aligned}$$

so that combining the inequalities of Lemmas 2.1, 2.2 and the Poincaré-like inequality, we obtain the result. ∎

REMARK 2.1 (Comments about inequality (2.18)). This result shows that after $p+1$ iteration, the error is $O(\Delta x^{p+1}) = O(\Delta t^{p+1})$ if a CFL-like condition is available. Of course it needs to be better that $\theta \Delta x < 1$ for the inequality to be effective, so we may experience a reduction of the CFL number. This reduction needs to be studied case by case, however this also shows that the overall cost of the method is of the order of an explicit one. This is why we name this computationally explicit.

2.1.4. Asymptotic preservation. We can show that the presented method is asymptotic preserving (AP), starting from the Chapman–Enskog expansion of the model (1.2a). Let us define $\mathbf{u}^\varepsilon := \mathbb{P}\mathbf{F}$, we obtain that

$$\begin{aligned}&\mathbf{F} = \mathbb{M}(\mathbf{u}^\varepsilon) + \mathcal{O}(\varepsilon) \\ &\frac{\partial \mathbf{u}^\varepsilon}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{u}^\varepsilon)}{\partial x} = \mathcal{O}(\varepsilon).\end{aligned} \tag{2.19}$$

PROPOSITION 2.2. *The discretisation given by (2.16) is consistent with the limit model (2.19) up to an $\mathcal{O}(\varepsilon)$.*

Proof. Now, using first (2.16) and then (2.15), defining $\mathbf{u}^{(p),\varepsilon} = \mathbb{P}\mathcal{F}^{(p)}$ and recalling that $\mu = \frac{\Delta t}{\varepsilon}$, $\mathbb{M}(\mathbb{P}\mathbf{u}) = \mathbf{u}$, $\mathbb{P}\Lambda\mathbb{M}(\mathbf{u}) = \mathbf{f}(\mathbf{u})$, by induction on the subtimesteps p

$$\mathbf{u}^{(p),\varepsilon} = \mathbb{P}\mathcal{F}^{(0)} - \frac{\Delta t}{\Delta x}\mathbb{P}\Lambda A \delta \mathcal{F}^{(p)} - \frac{\Delta t}{\Delta x}\mathbb{P}\mathbf{a}_0 \otimes \delta \mathcal{F}^{(p)} + \mathcal{O}(\varepsilon),$$

we can extend the formal expansion also in the discrete case, i.e.,

$$\begin{aligned}\mathcal{F}^{(p+1)} &= \mathbb{M}\left(\mathbb{P}\mathcal{F}^{(0)} - \frac{\Delta t}{\Delta x}\mathbb{P}\Lambda A \delta \mathcal{F}^{(p)} - \frac{\Delta t}{\Delta x}\mathbb{P}\Lambda \mathbf{a}_0 \otimes \delta \mathcal{F}^{(0)}\right) + \mathcal{O}(\varepsilon) = \mathbb{M}\left(\mathbf{u}^{(p),\varepsilon}\right) + \mathcal{O}(\varepsilon) \\ \mathbf{u}^{(p+1),\varepsilon} &= \mathbf{u}^{(0)} - \frac{\Delta t}{\Delta x} A \delta \mathbb{P}\Lambda \mathcal{F}^{(p)} - \frac{\Delta t}{\Delta x}\mathbf{a}_0 \otimes \mathbb{P}\Lambda \delta \mathcal{F}^{(0)} + \mathcal{O}(\varepsilon) \\ &= \mathbf{u}^{(0)} - \frac{\Delta t}{\Delta x} A \delta \mathbf{f}(\mathbf{u}^{(p),\varepsilon}) - \frac{\Delta t}{\Delta x}\mathbf{a}_0 \otimes \delta \mathbf{f}(\mathbf{u}^{(0),\varepsilon}) + \mathcal{O}(\varepsilon).\end{aligned} \tag{2.20}$$

The final result is a discretisation in space and time of the asymptotic model given by (2.19), if the the spatial discretisation is consistent with the space derivative. ∎

REMARK 2.2. One can proceed further and prove that, both in the discrete and the continuous case, the next term of the Chapman Enskog expansion is a diffusive term under Whitham's subcharacteristic condition of $\Lambda^2 - \partial_u \mathbf{f}(u)$ being positive definite. We can also prove that the discretisation is consistent also with that term up to an $\mathcal{O}(\varepsilon^2) + \mathcal{O}(\Delta t^2)$ if the spatial discretisation is at least consistent. We refer to [2] for the details of such computations for the sake of brevity.

2.1.5. Examples of $\mathcal{L}^2$ time discretisation. Here, we will consider second and fourth order approximation in time in the $\mathcal{L}^2$ operator, namely the Crank-Nicholson method and the fourth order one that uses the points t_n, $t_n + \frac{\Delta t}{2}$ and t_{n+1}. They are described by their matrices A,

- Second order

$$A_2 = \left(\tfrac{1}{2}\right), \mathbf{a}_0 = \left(\tfrac{1}{2}\right), \quad \mathcal{F} = \begin{pmatrix} \mathbf{F}^{n,1} \\ \mathbf{F}^{n,0} \end{pmatrix}$$

with $\mathbf{F}^{n,0} = \mathbf{F}^n$ and $\mathbf{F}^{n,1} \approx \mathbf{F}(t^{n+1})$. Writing the $\mathcal{L}^2$ discretisation of the time derivative applied to $\mathbf{F}$, we would have

$$\mathbf{F}^{n,1} - \mathbf{F}^{n,0} + \Delta t A \Lambda \delta \mathcal{F} = 0,$$

i.e.,

$$\mathbf{F}^{n,1} - \mathbf{F}^{n,0} + \Delta t \left(\frac{1}{2} \Lambda \delta \mathbf{F}^{n,1} + \frac{1}{2} \Lambda \delta \mathbf{F}^{n,0}\right) = 0.$$

This is Crank-Nicholson. We see that

$$\left(\mathrm{Id}_{1,1} + \mu A_2\right)^{-1} = \frac{2\varepsilon}{2\varepsilon + \Delta t}, \qquad \mu A_2 \left(\mathrm{Id}_{1,1} + \mu A_2\right)^{-1} = \frac{2\Delta t}{2\varepsilon + \Delta t}.$$

are uniformly bounded in $[0,2]$ for any $(\Delta t, \varepsilon)$.

- The fourth order scheme is obtained by

$$A_3 = \begin{pmatrix} \frac{1}{3} & \frac{-1}{24} \\ \frac{2}{3} & \frac{1}{6} \end{pmatrix}, \mathbf{a}_0 = \begin{pmatrix} \frac{5}{24} \\ \frac{1}{6} \end{pmatrix}, \mathcal{F} = \begin{pmatrix} \mathbf{F}^{n,2} \\ \mathbf{F}^{n,1} \\ \mathbf{F}^{n,0} \end{pmatrix}$$

where $\mathbf{F}^{n,0} = \mathbf{F}^n$, $\mathbf{F}^{n,1} \approx \mathbf{F}(t_n + \frac{\Delta t}{2})$ and $\mathbf{F}^{n,2} \approx \mathbf{F}(t^{n+1})$. We see that

$$\det\left(\mathrm{Id}_{2\times 2} + \mu A_3\right) = \left(1 + \frac{\mu}{3}\right)\left(1 + \frac{\mu}{6}\right) + \frac{1}{36} > 0$$

so the matrix is invertible. It is also easy to see that the matrices

$$\left(\mathrm{Id}_{2\times 2} + \mu A_3\right)^{-1} \text{ and } \mu \left(\mathrm{Id}_{2\times 2} + \mu A_3\right)^{-1} A$$

are uniformly bounded.
In fact, in this case, the operator $\mathcal{L}^2$ corresponds to the scheme Lobatto III, which is fourth order accurate [11]. For that reason, we will use this temporal scheme in conjunction with a fourth order spatial approximation.

2.2. Space discretisation: Definition of the δ operator. The only question left is how to define a stable scheme. As we have seen in Section 2.1.4, the scheme is asymptotic preserving. Under Whitham's subcharacteristic conditions the relaxation terms introduces diffusion which further stabilize the scheme. Hence, we focus only on the stability of the convection scheme, which will also guarantee the stability of the full scheme. The answer for the fully nonlinear convection problem is out of reach, at least for this paper, so we will rely on a classical linear stability analysis. The stability of the convection schemes splits into two sub-questions: is the convection scheme defined by $\mathcal{L}^2=0$ conditionally or unconditionally stable, and then, is the convection scheme defined by the DeC iteration (2.13) stable, and under which conditions. In the next section, we will provide 3 examples with increasing accuracy, and sketch a general method.

The matrix Λ is diagonal. In [12], the author considers the transport equation

$$u_t+au_x=0$$

and shows that if $a<0$ and

$$u_x(x_i)\approx\frac{1}{\Delta x}\sum_{j=-r}^{s}\alpha_j u_{i+j},$$

then the order is at most $2\min(r+1,s)$ and in addition the only stable methods are those defined for $r=s$ or $s=r+1$ or $s=r+2$. If $a>0$, we set

$$u_x(x_i)\approx\frac{1}{\Delta x}\sum_{j=-s}^{r}\alpha_j u_{i+j},$$

while in that case $r=s$ or $r=s+1$ or $r=s+2$. We will only consider these approximations. Following [12], we have

$$\alpha_j=\frac{(-1)^{j+1}}{j}\frac{r!s!}{(r+j)!(s-j)!},\qquad -r\le j\le s, j\neq 0,$$
$$\alpha_0=-\sum_{j=-r,j\neq 0}^{s}\alpha_j$$

and

$$\frac{\delta_k u}{\Delta x}-\frac{\partial u}{\partial x}(x_k)=c\Delta x^q\frac{\partial^{q+1}u}{\partial x^{q+1}}(x_k)+O(\Delta x^{q+1}),\qquad q=r+s,$$
$$c=\frac{(-1)^{s-1}r!s!}{(r+s+1)!}.$$

Remark 2.3 (Conservation). We note that we can always write

$$\delta_i u=\hat{\mathbf{f}}_{i+1/2}-\hat{\mathbf{f}}_{i-1/2}\tag{2.21}$$

with

$$\hat{\mathbf{f}}_{i+1/2}=\sum_{j=-r+1}^{s}\beta_j u_{i+j},\qquad \beta_j=\sum_{l\ge j+1}\alpha_l.\tag{2.22}$$

Proof. Assuming that $\hat{\mathbf{f}}_{i+1/2} = \sum_{j=-r}^{s-1} \beta_j u_{i+j}$ for any i, we write

$$\begin{aligned}\alpha_{-r}u_{i-r}+\ldots+\alpha_s u_{i+s} &= \Big(\beta_{-r+1}u_{i-r+1}+\ldots\beta_s u_{i+s}\Big)-\Big(\beta_{-r+1}u_{i-r}+\ldots\beta_s u_{i+s-1}\Big)\\ &= -\beta_{-r+1}u_{i-r}+(\beta_{-r+1}-\beta_{-r})u_{i-1-r}+\ldots\\ &\quad +(\beta_l-\beta_{l-1})u_{i+l-1}+\ldots+\beta_s u_{i+s}\end{aligned}$$

so that $\beta_j = -\sum_{l\geq j+1}\alpha_l$, using that $\sum_{l=-s}^{r}\alpha_l = 0$. ∎

This means that the approximations (2.13) and (2.16), in the limit $\varepsilon\to 0$, are always conservative since Λ is diagonal, and thanks to (2.10).

We list some possible choices for δ:

- First order approximation: this is the upwind scheme. If $a>0$, we take $\delta_1 u_j = u_j - u_{j-1}$, while if $a<0$, $\delta_1 u_j = u_{j+1}-u_j$. If $a=0$, of course $\delta_1 u_j = 0$. The flux is

$$\hat{\mathbf{f}}_{j+1/2} = \frac{1}{2}\big(u_j + u_{j+1} + \text{sign }(a)(u_{j+1}-u_j)\big), \qquad \text{sign}(a) = \frac{a}{|a|}.$$

- Second order: for $a<0$,

$$\delta_2 u_j = -\frac{u_{j-1}}{3} - \frac{u_j}{2} + u_{j+1} - \frac{u_{j+2}}{6}$$

so that

$$u_x = \frac{1}{\Delta x}\left(-\frac{u_{j-1}}{3} - \frac{u_j}{2} + u_{j+1} - \frac{u_{j+2}}{6}\right) + c\Delta x^3 \frac{\partial^4 u}{\partial x^4} + O(\Delta x^4)$$

with

$$c = -\frac{1}{12}.$$

This corresponds to the $[r, r+2]$ approximation with $r=-1$. In terms of flux, we have (for $a<0$):

$$f_{j+1/2} = \frac{1}{6}\big(2u_j + 5u_{j+1} - u_{j+2}\big).$$

For $a>0$, we have

$$f_{j+1/2} = \frac{1}{6}\big(2u_{j+1} + 5u_j - u_{j-1}\big),$$

so all in all

$$\hat{\mathbf{f}}_{j+1/2} = \frac{1-\text{ sign }(a)}{12}\big(2u_j + 5u_{j+1} - u_{j+2}\big) + \frac{1+\text{ sign }(a)}{12}\big(2u_{j+1} + 5u_j - u_{j-1}\big).$$

- Fourth order: if $r=s=2$, and for any a

$$\delta_4^1 u_j = \frac{u_{j+2}-u_{j-2}}{12} + 2\frac{u_{j+1}-u_{j-1}}{3}$$

hence

$$\frac{\partial u}{\partial x}-\frac{\delta_4^1 u}{\Delta x}=c\Delta x^4\frac{\partial^5 u}{\partial x^5}+O(\Delta x^5)$$

and if $r=1$, $s=3$ and $a<0$,

$$\delta_4^2 u=-\frac{u_{j-1}}{4}-\frac{5}{6}u_j+\frac{3}{2}u_{j+1}-\frac{u_{j+2}}{2}+\frac{u_{j+3}}{12}.$$

In terms of flux, we have:

– for δ_4^1,

$$\hat{\mathbf{f}}_{j+1/2}=a\left(\frac{u_{j+2}}{12}+\frac{3}{4}u_{j+1}+\frac{3}{4}u_j+\frac{u_{j-1}}{12}\right)$$

– for δ_4^2,

$$\hat{\mathbf{f}}_{j+1/2}=\frac{1-\text{ sign }(a)}{2}\left(\frac{u_{j+3}}{12}-\frac{5}{12}u_{j+2}+\frac{13}{12}u_{j+1}+\frac{u_j}{4}\right)+\frac{1+\text{ sign }(a)}{2}\left(\frac{u_{j+1}}{4}+\frac{13}{12}u_j-\frac{5}{12}u_{j-1}+\frac{u_{j-2}}{12}\right).$$

3. Stability analysis

We study the stability of the discretisation of the homogeneous problem. Since Λ is diagonal, it is enough to look at the scalar conservation problem. We first look at the implicit method defined by $\mathcal{L}^2=0$, and then at the DeC iteration that is constructed on top of it. This is done by Fourier analysis, we can assume that $a>0$ and the Fourier symbol of δ is g. The Table 3.1 displays the symbols of the operators.

Operator	Symbol g
δ_1	$1-e^{-i\theta}$
δ_2	$\frac{1}{3}e^{i\theta}+\frac{1}{2}-e^{-i\theta}+\frac{1}{6}e^{-2i\theta}$
δ_4^1	$i\left(\frac{\sin(2\theta)}{6}+\frac{4}{3}\sin\theta\right)$
δ_4^2	$\frac{e^{i\theta}}{4}+\frac{5}{6}-\frac{3}{2}e^{-i\theta}+\frac{1}{2}e^{-2i\theta}-\frac{e^{-3i\theta}}{12}$

TABLE 3.1. *List of Fourier symbols.*

The next step is to evaluate the amplification factors of the method, first without DeC iteration, then with DeC iteration.

3.1. First order in time. For a first order scheme the $\mathcal{L}^2$ operator can be written as an implicit Euler method, though being computationally explicit, while the DeC iteration, which consists of one step, resembles the explicit Euler method with CFL constrained $0\le\lambda\le 1$, where $\lambda=a\Delta t/\Delta x$. For the $\mathcal{L}^2=0$ operator, by Fourier transform, we have $\hat{u}^{n+1}-\hat{u}^n+\lambda g\hat{u}^{n+1}=0$, so that the amplification factor is $G=\dfrac{1}{1+\lambda g}$ which is of modulus ≤ 1 if

$$2\lambda\Re(g)+\lambda^2|g|^2\ge 0.$$

If $\lambda \to 0^+$, we see that $\Re(g) \geq 0$ is a necessary condition, while if $\lambda \to 0^-$, $\Re(g) \leq 0$. In all cases, $\lambda \Re(g) \geq 0$ is a necessary condition. Writing $g = a + ib$, and assuming that $\lambda \neq 0$, we see that this condition reads:

$$2\lambda a + \lambda^2(a^2 + b^2) = (\lambda a + 1)^2 + \lambda^2 b^2 - 1 \geq 0.$$

We also see that

$$(\lambda a + 1)^2 + \lambda^2 b^2 \geq (\lambda a + 1)^2 \geq 1$$

so that $\lambda \Re(g) \geq 0$ is a necessary and sufficient condition for stability. The Table 3.2 provides the stability condition for the first, second and fourth order schemes. For the rest of the discussion we consider $a = 1$ and in case it is different, one has to rescale λ accordingly, as classically done for CFL conditions.

3.2. Second order in time. In that case the $\mathcal{L}^2 = 0$ scheme reads:

$$u_i^{n+1} - u_i^n + \frac{\lambda}{2}\left(\delta u_i^n + \delta u_i^{n+1}\right) = 0,$$

for which the amplification factor is simply

$$G = \frac{1 - \frac{\lambda}{2} g}{1 + \frac{\lambda}{2} g}.$$

We have $|G| \leq 1$ if and only if

$$\lambda \Re(g) \geq 0.$$

Again, the Table 3.2 provides the stability condition for the first, second and fourth order schemes.

The DeC iteration is

$$u_i^{(p+1)} = u_i^n - \frac{\lambda}{2}\left(\delta u_i^n + \delta u_i^{(p)}\right),$$

so that

$$G_0 = 1$$
$$G_{p+1} = 1 - \frac{\lambda}{2}\left(g + g G_p\right)$$

and we see that

$$G_{p+1} - G = -\frac{\lambda g}{2}\left(G_p - G\right) = \left(-\frac{\lambda g}{2}\right)^{p+1}\left(1 - G\right).$$

3.3. Fourth order in time. Here, the $\mathcal{L}^2 = 0$ scheme reads:

$$u_i^{n+1/2} - u_i^n + \lambda\left(\frac{5}{24}\delta u_i^n + \frac{1}{3}\delta u_i^{n+1/2} - \frac{1}{24}\delta u_i^{n+1}\right) = 0$$
$$u_i^{n+1} - u_i^n + \lambda\left(\frac{1}{6}\delta u_i^n + \frac{2}{3}\delta u_i^{n+1/2} + \frac{1}{6}\delta u_i^{n+1}\right) = 0$$

so that the Fourier transform gives

$$\begin{pmatrix}\hat{u}^{n+1/2}\\ \hat{u}^{n+1}\end{pmatrix}=G\begin{pmatrix}\hat{u}^{n}\\ \hat{u}^{n}\end{pmatrix}$$

with

$$G=\begin{pmatrix}1+\frac{\lambda g}{3} & -\frac{\lambda g}{24}\\ \frac{2\lambda g}{3} & 1+\frac{\lambda g}{6}\end{pmatrix}^{-1}\begin{pmatrix}1-\frac{5\lambda}{24}g\\ 1-\frac{\lambda g}{6}\end{pmatrix}=\begin{pmatrix}G_1\\ G_2\end{pmatrix}$$

and we have to look at $\max\{|G_1|,|G_2|\}\le 1$ for the calculation of $\hat{u}^{n+1/2}$ and $\hat{u}^{n+1}$ to be stable. We have

$$G=\begin{pmatrix}\frac{-g^2\lambda^2+24}{2g^2\lambda^2+12\lambda g+24}\\ \frac{g^2\lambda^2-6\lambda g+12}{g^2\lambda^2+6\lambda g+12}\end{pmatrix}.$$

Then with obvious notations, the DeC iteration is

$$\begin{aligned}v_1^{(p+1)}-u_i^n+\lambda\left(\frac{5}{24}\delta u_i^n+\frac{1}{3}\delta v_1^{(p)}-\frac{1}{24}\delta v_2^{(p)}\right)=0\\ v_2^{(p+1)}-u_i^n+\lambda\left(\frac{1}{6}\delta u_i^n+\frac{2}{3}\delta v_1^{(p)}+\frac{1}{6}\delta v_2^{(p)}\right)=0\end{aligned}$$

The Fourier analysis gives:

$$\hat{v}^{(p+1)}=\begin{pmatrix}1-\lambda\theta_0^1 g\\ 1-\lambda\theta_0^2 g\end{pmatrix}\hat{u}^n-\lambda g\begin{pmatrix}\theta_1^1 & \theta_2^1\\ \theta_1^2 & \theta_2^2\end{pmatrix}\hat{v}^{(p)},\qquad \begin{pmatrix}\theta_0^1 & \theta_1^1 & \theta_2^1\\ \theta_0^1 & \theta_1^2 & \theta_2^2\end{pmatrix}=\begin{pmatrix}\frac{5}{24} & \frac{1}{3} & \frac{-1}{24}\\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6}\end{pmatrix}.$$

The amplification vector, G_p after the p-th iteration is defined by

$$\begin{aligned}G_0&=\begin{pmatrix}1\\ 1\end{pmatrix}\\ G_{p+1}&=\begin{pmatrix}1-\lambda\theta_0^1 g\\ 1-\lambda\theta_0^2 g\end{pmatrix}-\lambda g\begin{pmatrix}\theta_1^1 & \theta_2^1\\ \theta_1^2 & \theta_2^2\end{pmatrix}G_p.\end{aligned}\tag{3.1}$$

We note that, setting $\theta=\begin{pmatrix}\theta_1^1 & \theta_2^1\\ \theta_1^2 & \theta_2^2\end{pmatrix}$,

$$G_{p+1}-G=(-\lambda g)^p\theta^p\left(\begin{pmatrix}1\\ 1\end{pmatrix}-G\right)$$

and $\rho(\theta)=\dfrac{1}{2\sqrt{3}}$. So using the spectra decomposition of θ which has two complex and distinct eigenvalues, we have that

$$\rho(\theta^p)\le\mu_p=\sqrt{\frac{17}{16}+\frac{\sqrt{241}}{16}}\left(\frac{1}{2\sqrt{3}}\right)^p.$$

We get finally $\mu_1=0.4115783562$, $\mu_2=0.1188124373$, $\mu_3=0.03429819635$, $\mu_4=0.009901036444$, hence the convergence is very quick.

	First order	second order	Fourth order
δ_1	✓	✓	✓
δ_2	✓	✓	✓ if $\lambda \le 4.5$
δ_4^1	✓	✓ ($\lvert G\rvert = 1$)	✓ ($\lvert G\rvert = 1$)
δ_4^2	✓	✓	✓ if $\lambda \le \frac{9}{4}$
Analytical condition	$\lambda \Re(g) \ge 0$	$\lambda \Re(g) \ge 0$	$\lambda \Re\left(g - \frac{\lambda g^2}{6}\right) > 0$

TABLE 3.2. *Stability conditions for the original scheme.*

3.4. Summary of the stability analysis. Combining these expressions with the actual form of the Fourier symbol of δ, we get the results of Table 3.2.

Now, we turn our attention on the DeC iteration. For the second-order-in-time approximation, we first have

$$G_p = (1-\theta_p)G + \theta_p \qquad \text{with } \theta_p = (-1)^p \left(\frac{\lambda}{2} g\right)^p.$$

So we get

$$|G_p|^2 - 1 = |1-\theta_p|^2 \left(|G|^2 - 1\right) + 2\Re\left(\overline{\theta_p}(1-\theta_p)(G-1)\right)$$

hence if $|G| \le 1$, a sufficient condition is that

$$\Re\left(\overline{\theta_p}(1-\theta_p)(G-1)\right) \le 0.$$

For the fourth order scheme, we have similarly

$$G_p = \left(\mathrm{Id} - (-\lambda g \theta)^p\right) G + (-\lambda g \theta)^p e, \qquad e = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

but it is more complicated to get an analytical condition. So we rely on Maple.

The stability conditions are summarised in Table 3.3.

Scheme		# iterations					
Order	δ	1	2	3	4	5	6
2	δ_1	1	1	1	1	1	1
2	δ_2	0	≥ 0.85	≥ 1.22	≥ 1.02	≥ 1.08	≥ 1.23
2	δ_4^1	0	0	≥ 1.45	≥ 1.45	≥ 0.002	≥ 0.01
2	δ_4^2	0	≥ 0.5	≥ 0.69	0.71	0.73	0.73
3	δ_1	6	≥ 1.5	≥ 1.87	≥ 2	≥ 2.23	≥ 2.48
3	δ_2	0	0	1	≥ 2.0447	≥ 2.17120	≥ 2.568
3	δ_4^1	0	0	0	≥ 1.6171	≥ 2.4727	≥ 2.9162
3	δ_4^2	0	0	≥ 0.1	≥ 1.3096	≥ 1.3955	≥ 1.8282

TABLE 3.3. *CFL number for stability of the DeC iterations.* 0 *means that the scheme is unconditionally unstable. If a real number x is given, it means that the scheme is stable up to CFL x, if $\ge x$ is written, this means that the scheme is stable for at least CFL x (and slightly above indeed).*

4. Wave model

We have to specify the diagonal matrix Λ and the Maxwellians $\mathbb{M}$. We will use two kinds of wave models:

- A two waves model. In that case,

$$\Lambda = \begin{pmatrix} a & 0 \\ 0 & -a \end{pmatrix}$$

with $a \geq \max_i \rho(\mathbf{f}'(u_i))$. Setting $\mathbf{u}^\varepsilon = \mathbb{P}\mathbf{F}$, we have $\mathbf{f}(\mathbf{u}^\varepsilon) = \mathbb{P}\Lambda\mathbf{F}$ and we know explicitly $\mathbb{M} = (\mathbb{M}_1, \mathbb{M}_2)$:

$$\mathbb{M}_1(\mathbb{P}\mathbf{F}) = \frac{1}{2}\left(\mathbf{u}^\varepsilon + \frac{\mathbf{f}}{a}\right), \qquad \mathbb{M}_2(\mathbb{P}\mathbf{F}) = \frac{1}{2}\left(\mathbf{u}^\varepsilon - \frac{\mathbf{f}}{a}\right). \tag{4.1}$$

- A three waves model, where

$$\Lambda = \begin{pmatrix} a & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -a \end{pmatrix}.$$

In that case, the Maxwellian is $\mathbb{M} = (\mathbb{M}_1, \mathbb{M}_2, \mathbb{M}_3)$ and we have

$$\begin{aligned} \mathbf{u}^\varepsilon &= \mathbb{M}_1 + \mathbb{M}_2 + \mathbb{M}_3 \\ \mathbf{f}(\mathbf{u}^\varepsilon) &= a\mathbb{M}_1 - a\mathbb{M}_3 \end{aligned}$$

so we need to specify $\mathbb{M}_2$.

For the scalar problems, we will use the two waves model that reveals itself sufficient. For the fluid problems, we will show that the two waves model is not perfect, and hence the three waves model needs to be considered.

In the case of 3 waves, let us specify $\mathbb{M}_2$. Following [6], we know that the sub-characteristic condition is equivalent to the monotonicity of the Maxwellians: they need to be differentiable and have only positive eigenvalues. In [6, 17], it is proposed to use

$$\begin{aligned} \mathbb{M}_1(\mathbf{u}^\varepsilon) &= \frac{1}{a}\mathbf{f}_+(\mathbf{u}^\varepsilon) \\ \mathbb{M}_2(\mathbf{u}^\varepsilon) &= \mathbf{u} - \frac{\mathbf{f}_+(\mathbf{u}^\varepsilon) - \mathbf{f}_-(\mathbf{u}^\varepsilon)}{a} \\ \mathbb{M}_3(\mathbf{u}^\varepsilon) &= \frac{1}{a}\mathbf{f}_-(\mathbf{u}^\varepsilon) \end{aligned} \tag{4.2}$$

where $\mathbf{f}(\mathbf{u}^\varepsilon) = \mathbf{f}_+(\mathbf{u}^\varepsilon) + \mathbf{f}_-(\mathbf{u}^\varepsilon)$, $\mathbf{f}_\pm$ are differentiable, $\nabla_{\mathbf{u}}\mathbf{f}_+(\mathbf{u})$ has only positive eigenvalues, while $\nabla_{\mathbf{u}}\mathbf{f}_-(\mathbf{u}^\varepsilon)$ has only negative eigenvalues. A possible choice, inspired by the Enquist-Osher-Solomon flux, is

$$\mathbb{M}_2(\mathbf{u}^\varepsilon) = \frac{\int_0^{\mathbf{u}^\varepsilon} |\mathbf{f}'(s)|\, ds}{|a|},$$

but the integral (or the path integral for system) must be evaluated. In the case of the Euler equations, we give a second one that does not necessitate the evaluation of an integral. In the case of the Euler equations,

$$\mathbf{u}^\varepsilon = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}, \quad \mathbf{f}(\mathbf{u}^\varepsilon) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(E+p) \end{pmatrix}, \quad p = (\gamma - 1)\big(E - \frac{1}{2}\rho u^2\big), \tag{4.3}$$

we propose to use a Maxwellian that relies on the van Leer flux splitting [21]. It is purely algebraic and defined by:

(1) if $M=\dfrac{u}{c}\le -1$, with $c^2=\gamma\frac{p}{\rho}$, then $\mathbf{f}_-(\mathbf{u}^\varepsilon)=\mathbf{f}(\mathbf{u}^\varepsilon)$, $\mathbf{f}_+(\mathbf{u}^\varepsilon)=0$,

(2) if $M\ge 1$, then $\mathbf{f}_+(\mathbf{u}^\varepsilon)=\mathbf{f}(\mathbf{u}^\varepsilon)$, $\mathbf{f}_-(\mathbf{u}^\varepsilon)=0$,

(3) if $-1\le M\le 1$, then

$$\mathbf{f}_-(\mathbf{u}^\varepsilon)=\begin{pmatrix} Q \\ \frac{QR}{\gamma} \\ \frac{QR^2}{2(\gamma^2-1)} \end{pmatrix},\ Q=-\frac{\rho}{4c}(u-c)^2,\ R=(\gamma-1)u-2c,$$

and $\mathbf{f}_+=\mathbf{f}-\mathbf{f}_-$.

The eigenvalues of $\mathbf{f}_\pm$ are bounded by

$$a=\begin{cases} (|u|+c)\dfrac{\gamma+3}{2\gamma+|M|(3-\gamma)} & \text{if } |M|\le 1 \\ |u|+c & \text{else.} \end{cases}$$

Note that $\frac{\gamma+3}{2\gamma+|M|(3-\gamma)}\le\frac{\gamma+3}{2\gamma}$ for $|M|\le 1$. For $\gamma=1.4$, $\frac{\gamma+3}{2\gamma}=\frac{11}{7}\approx 1.57$.

5. Nonlinear stabilisation

If the solution is expected to be nonsmooth, then one can expect the occurrence of spurious oscillations. Sometimes, oscillations are acceptable, provided they do not lead to the crash of the simulation. In order to get rid of them, or to control them, we have adopted the MOOD technique initially designed in [7] with some improvements described in [22]. We have adapted it our way in order to get results that are formally of order $p+1$ in space and time, here $p=1,2,3$.

MOOD is an *a posteriori* corrector of high order numerical methods [7,22]. MOOD requires a sequence of schemes ordered from the most accurate/less stable one to the low order/more reliable one. It also requires a series of criteria that the solution should fulfill, e.g. physical admissibility, discrete minimum principle or numerical errors. After having performed a step of the most accurate scheme, it checks the criteria on each cell/degree of freedom, and detects the areas where the criteria are not met. There, we switch to the next scheme in a *cascade* style, which is supposed to be more stable and reliable. We proceed iteratively until either the criteria are met or the most reliable/less accurate *parachute* scheme is used. The *parachute* scheme should analytically guarantee all the criteria.

In the following we describe how the criteria must be verified on the described spatial discretisation, while the list of the schemes that we use consists always of 2 schemes (the considered one and the upwind discretisation δ_1 as *parachute* scheme) and we specify directly in the numerical simulations which criteria will be considered, as they are problem dependent.

We proceed as follows: at the time step t_n, we have the values $(\mathbf{F}_k^n)_k$. From now on, we drop the superscript n, since there is no ambiguity. In the DeC iteration (2.5), with the spatial scheme defined by δ_p, writes (with the convention that $\mathbf{F}^{(l),0}=\mathbf{F}^0$ for $l=0,\dots,q-1$)

$$\mathbf{F}_k^{(p+1),j}-\mathbf{F}_k^0+\frac{\Delta t}{\Delta x}\left(\sum_{l=0}^{q}a_{jl}\Lambda\delta_k\mathbf{F}^{(p),l}-\mu\sum_{l=0}^{q}a_{jl}\left(\mathbb{MP}\mathbf{F}_k^{(p+1),l}-\mathbf{F}_k^{(p+1),l}\right)\right),\ p=0,\dots,q-1,$$

from which we get

$$\mathbb{P}\mathbf{F}_k^{(p+1),j} - \mathbb{P}\mathbf{F}_k^0 + \frac{\Delta t}{\Delta x}\left(\sum_{l=0}^{q} a_{jl}\mathbb{P}\Lambda\delta_k\mathbf{F}^{(p),l}\right) = 0, \quad p = 0,\dots,q-1. \tag{5.1}$$

The increment $\delta_k\mathbf{F}^l$ is the difference of two terms, and we write

$$\delta_k\mathbf{F}^l = \Lambda\left(\hat{\mathbf{F}}^l_{k+1/2} - \hat{\mathbf{F}}^l_{k-1/2}\right) = \Phi_k^{[k,k+1],l} + \Phi_k^{[k-1,k],l}$$

with

$$\Phi_k^{[k,k+1],l} = \Lambda\hat{\mathbf{F}}^l_{k+1/2} - \Lambda\mathbf{F}^l_k, \quad \Phi_{k+1}^{[k,k+1],l} = \Lambda\mathbf{F}^l_{k+1} - \Lambda\hat{\mathbf{F}}^l_{k+1/2}.$$

One equivalent way to rephrase the conservation is

$$\Phi_k^{[k,k+1],l} + \Phi_{k+1}^{[k,k+1],l} = \Lambda\left(\mathbf{F}^l_{k+1} - \mathbf{F}^l_k\right) \tag{5.2}$$

and the right-hand side of this relation is independent of the order p. It is equivalent because we see that

$$\Lambda\hat{\mathbf{F}}^l_{k+1/2} = \frac{1}{2}\left(\Lambda\mathbf{F}^l_{k+1} + \Lambda\mathbf{F}^l_k - \left(\Phi_k^{[k,k+1],l} - \Phi_{k+1}^{[k,k+1],l}\right)\right).$$

Using this we rewrite (5.1) as:

$$\mathbb{P}\mathbf{F}_k^{(p+1),j} = \frac{1}{2}\left(\left(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}}\right)_{k-1/2} + \left(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}}\right)_{k+1/2}\right) \tag{5.3a}$$

with

$$\begin{aligned}\left(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}}\right)_{k-1/2} &= \mathbb{P}\mathbf{F}_k^0 - \frac{\Delta t}{\Delta x}\Phi_k^{[k-1,k],(p),j}\\ \left(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}}\right)_{k+1/2} &= \mathbb{P}\mathbf{F}_k^0 - \frac{\Delta t}{\Delta x}\Phi_k^{[k,k+1],(p),j}\end{aligned} \tag{5.3b}$$

In practice, we compute for each interval $[k,k+1]$

$$\begin{aligned}\left(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}}\right)_{k+1/2} &= \mathbb{P}\mathbf{F}_k^0 - \frac{\Delta t}{\Delta x}\Phi_k^{[k,k+1],(p),j}\\ \left(\widetilde{\mathbb{P}\mathbf{F}_{k+1}^{(p+1),j}}\right)_{k+1/2} &= \mathbb{P}\mathbf{F}_{k+1}^0 - \frac{\Delta t}{\Delta x}\Phi_{k+1}^{[k,k+1],(p),j}\end{aligned} \tag{5.4}$$

and then apply (5.3a).

In the simplified version of the MOOD algorithm we use, we consider only two spatial approximations, namely the first order one defined by δ_1, and the high order one defined by δ_p, $p=2$ or 3 in this paper. The idea is to use as often as possible the highest order scheme, and to use the low order one to correct potential problems. Knowing the $\{\mathbf{F}_k^{(p),j}\}_k$, we first compute for each interval $[k,k+1]$ the quantities defined by (5.4) with the high order residuals. Then we test the values of the results using a set of criteria, applied on $(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}})_{k+1/2}$ and $(\widetilde{\mathbb{P}\mathbf{F}_{k+1}^{(p+1),j}})_{k+1/2}$. This set of criteria is explained in the next paragraph. If both $(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}})_{k+1/2}$ and $(\widetilde{\mathbb{P}\mathbf{F}_{k+1}^{(p+1),j}})_{k+1/2}$ pass the tests, this element is declared sane, else un-sane. This enables to identify a set $\mathcal{I}$

of un-sane elements $[k,k+1]$ where the criteria are not met, and we store the residual $\{\Phi_k^{[k,k+1],(p),j},\Phi_{k+1}^{[k,k+1],(p),j}\}$ for the sane elements. We then repeat the procedure for the un-sane elements with the lowest order scheme. At the end of the procedure, we have evaluated residuals, that we still denote by $\{\Phi_k^{[k,k+1],(p),j},\Phi_{k+1}^{[k,k+1],(p),j}\}$, even though they are potentially evaluated by different schemes. We then compute $\mathbb{P}\mathbf{F}_k^{(p+1),j}$ by (5.1). There is no problem of conservation since (5.2) holds true.

Now we describe the criteria we apply to $(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}})_{k+1/2}$ and $(\widetilde{\mathbb{P}\mathbf{F}_{k+1}^{(p+1),j}})_{k+1/2}$, following the ideas of [7,22] with some small adaptation to the context. When specific tests are done on a variable, we denote this variable by ξ. For a scalar problem, ξ is simply the conserved variable. In the case of the Euler equations, we test this on some primitive variables: the density and the energy, and for some severe problems, the velocity. We can add as many criteria as needed.

(1) We first check if $(\widetilde{\mathbb{P}\mathbf{F}_k^{(p+1),j}})_{k+1/2}$ and $(\widetilde{\mathbb{P}\mathbf{F}_{k+1}^{(p+1),j}})_{k+1/2}$ lie in the invariance domain; if relevant: in the case of the Euler equation, we check if the density and the internal energy are both positive. If not, we set the criteria to .FALSE. on this element. In that case we jump to the next element, else we look for the next criterion.

(2) We check if the solution is not locally constant. Taking $\nu=\Delta x^3$ and $\mathcal{S}$ the stencil defined by the operator δ, we check if

$$\left|\max_{l\in\mathcal{S}}\xi_{i+l}-\min_{l\in\mathcal{S}}\xi_{i+l}\right|\le\nu \text{ and } \left|\max_{l\in\mathcal{S}}\xi_{i+1+l}-\min_{l\in\mathcal{S}}\xi_{i+1+l}\right|\le\nu.$$

If this is true, the criteria is kept to .TRUE., else it is set to .FALSE. and we jump to the next element.

(3) We check if a new extrema is created or not, by comparing with the solution at the previous time step, in a neighbourhood extended to the right and the left by one cell: we are running at CFL 1.

(a) We first test if $\xi_k^{n+1},\xi_{k+1}^{n+1}\in[\min_{l\in\mathcal{S}}\xi_{k+l}+\epsilon,\max_{l\in\mathcal{S}}\xi_{k+l}-\epsilon]\cap[\min_{l\in\mathcal{S}}\xi_{k+1+l}+\epsilon,\max_{l\in\mathcal{S}}\xi_{k+1+l}-\epsilon]$. If this is true, we jump to the next element,

(b) else, denoting by P_j the Lagrange interpolation polynomial that interpolates $\{\xi_{j+l}\}_{l\in\mathcal{S}}$

- we compute $\xi'=P_k'(x_k)$, $\xi_L'=P_k'(x_k-\frac{\Delta x}{2})$, $\xi_{min/max}'^{,k-1/2}=\min/\max\big(P_k'(x_k-\frac{\Delta x}{2}),P_{k-1}'(x_k-\frac{\Delta x}{2})\big)$ then
  - if $\xi_L'<\xi'$, $\alpha_L=\min(1,\frac{\xi_{max}'^{,k-1/2}-\xi'}{\xi_L'-\xi'})$
  - if $\xi_L'=\xi'$, $\alpha_L=1$
  - if $\xi_L'<\xi'$, $\alpha_L=\min(1,\frac{\xi_{min}'^{,k-1/2}-\xi'}{\xi_L'-\xi'})$
- we compute $\xi'=P_k'(x_k)$, $\xi_R'=P_k'(x_k+\frac{\Delta x}{2})$, $\xi_{min/max}'^{,k+1/2}=\min/\max\big(P_{k+1}'(x_k+\frac{\Delta x}{2}),P_k'(x_k+\frac{\Delta x}{2})\big)$ then
  - if $\xi_R'<\xi'$, $\alpha_R=\min(1,\frac{\xi_{max}'^{,k+1/2}-\xi'}{\xi_R'-\xi'})$
  - if $\xi_R'=\xi'$, $\alpha_R=1$

– if $\xi'_R < \xi'$, $\alpha_R = \min(1, \dfrac{\xi'^{,k+1/2}_{min} - \xi'}{\xi'_R - \xi'})$

- we set $\alpha = \min(\alpha_L, \alpha_R)$
- if $\alpha = 1$, then we have a true extrema, keep the criteria to .TRUE. and jump to the next element. Else, we set the criteria to .FALSE. and jump to the next element.

The idea behind the step 3 is described in [22] and is also related to [14]: we try to check if the gradient of the interpolation ξ lies in the interval $\left[\min\left(\xi'^{,k-1/2}_{min}, \xi'^{,k+1/2}_{min}\right), \max\left(\xi'^{,k-1/2}_{max}, \xi'^{,k+1/2}_{max}\right)\right]$.

REMARK 5.1 (Stability). The von Neumann stability study of Section 3 does not hold directly for the MOOD algorithm but it is clear that we are dealing with a combination of the stabilities of the schemes used in the MOOD *cascade*. Hence, if we choose CFL conditions that guarantee the stability of both the high order schemes and of the *parachute* scheme (upwind CFL=1 in our case), then we know that the global MOOD scheme will be von Neumann stable.

6. Numerical examples

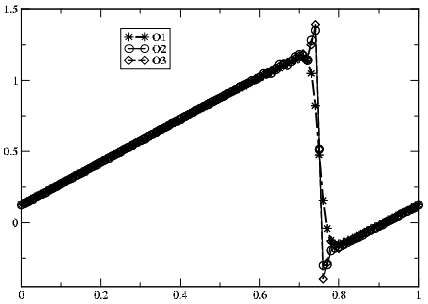
6.1. Scalar problems. The first problem is the transport equation with periodic boundary conditions

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

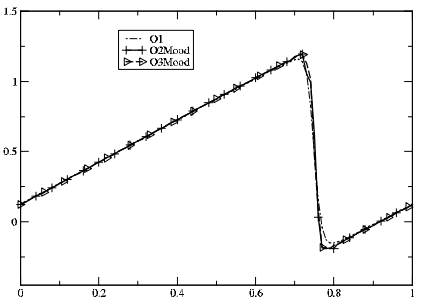
where the initial condition is

$$u_0(x) = \sin(2\pi x) + 0.5. \tag{6.1}$$

A two waves model is used with $a = 1.01$ (so a little larger that the actual maximum speed. We always proceed as such for scalar and system cases. We make a convergence test for short and long final times, namely $T = 0.5$ and $T = 10$. The CFL number, with respect to the wave model maximum speed, is always set to 1. In both cases, we see that the expected order of accuracy is obtained, see Table 6.1 for $T = 0.5$ and Table 6.2 for $T = 10$.



(a) First order, second and fourth order solution, with no stabilisation.



(b) First order, second and fourth order with MOOD.

FIG. 6.1. *Burgers equation T=0.5. Initial condition :* $u_0(x) = \sin(2\pi x) + 0.5$

Note that for the second and fourth order schemes, the nonlinear stabilisation does not detect any troubled point, giving exactly the same error as in the non-stabilised case. We show the results for the fourth order schemes. All the calculations are done with the two waves model. Note that the first order scheme, with the initial condition given by the Maxwellian, is nothing more than the Lax-Friedrichs scheme, for second order in time approximation. For fourth order in time, since the equilibrium relaxation is more complex, we get a different scheme. Note that the nonlinear stabilisation procedure of Section 5 does not flag any cell.

h	L^1	r	L^2	r	L^∞	r
First order						
50	$2.75963992\,10^{-2}$	-	$3.89822088\,10^{-2}$	-	$2.49972343\,10^{-2}$	-
100	$1.31966826\,10^{-2}$	1.43	$1.8655914\,10^{-2}$	1.43	$1.18893785\,10^{-2}$	1.43
200	$6.94037229\,10^{-3}$	1.33	$9.8105203\,10^{-3}$	1.33	$6.26028096\,10^{-3}$	1.33
400	$3.47535103\,10^{-3}$	1.38	$4.9137393\,10^{-3}$	1.38	$3.13191721\,10^{-3}$	1.38
800	$1.73895701\,10^{-3}$	1.38	$2.4590014\,10^{-3}$	1.38	$1.56634545\,10^{-3}$	1.38
Second order						
50	$4.83627617\,10^{-3}$	-	$6.70434069\,10^{-3}$	-	$4.40502120\,10^{-3}$	-
100	$1.21754361\,10^{-3}$	2.07	$1.70489808\,10^{-3}$	2.06	$1.10206485\,10^{-3}$	2.08
200	$3.05118738\,10^{-4}$	2.07	$4.2937923\,10^{-4}$	2.07	$2.75470491\,10^{-4}$	2.08
400	$7.60697367\,10^{-5}$	2.08	$1.07314205\,10^{-4}$	2.08	$6.85840860\,10^{-5}$	2.08
800	$1.89899602\,10^{-5}$	2.08	$2.6821670\,10^{-5}$	2.08	$1.71091069\,10^{-5}$	2.08
Fourth order						
50	$0.20142\,10^{-4}$	-	$0.27897\,10^{-4}$	-	$0.18321\,10^{-4}$	-
100	$0.12237\,10^{-5}$	4.04	$0.17133\,10^{-5}$	4.02	$0.11081\,10^{-5}$	4.04
200	$0.75854\,10^{-7}$	4.01	$0.10674\,10^{-6}$	4.00	$0.68485\,10^{-7}$	4.01
400	$0.47247\,10^{-8}$	4.00	$0.66651\,10^{-8}$	4.00	$0.42597\,10^{-8}$	4.00
800	$0.29483\,10^{-9}$	4.00	$0.41643\,10^{-9}$	4.00	$0.26563\,10^{-9}$	4.00
Fourth order+MOOD						
50	$0.20142\,10^{-4}$	-	$0.27897\,10^{-4}$	-	$0.18321\,10^{-4}$	-
100	$0.12237\,10^{-5}$	4.04	$0.17133\,10^{-5}$	4.02	$0.11081\,10^{-5}$	4.04
200	$0.75854\,10^{-7}$	4.01	$0.10674\,10^{-6}$	4.00	$0.68485\,10^{-7}$	4.01
400	$0.47247\,10^{-8}$	4.00	$0.66651\,10^{-8}$	4.00	$0.42597\,10^{-8}$	4.00
800	$0.29483\,10^{-9}$	4.00	$0.41643\,10^{-9}$	4.00	$0.26563\,10^{-9}$	4.00

TABLE 6.1. *Order of convergence for the convection problem and two waves model for order 1, 2 and 4, and 4th order with MOOD. The final time is* $T=0.5$. *One can see that the two fourth order results are* identical *as expected.*

The convergence tables are obtained against the solution of the asymptotic model, proving moreover that the model is asymptotic preserving as expected.

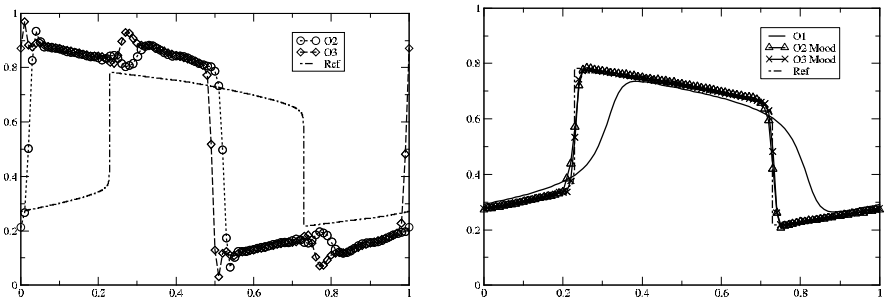
The Figure 6.1 shows some results for the Burgers equation

$$\frac{\partial u}{\partial t}+\frac{1}{2}\frac{\partial u^2}{\partial x}=0 \tag{6.2}$$

with the initial condition (6.1). This generates an unsteady shock wave, so a priori more challenging than a steady one. The nonlinear stabilisation performs correctly.

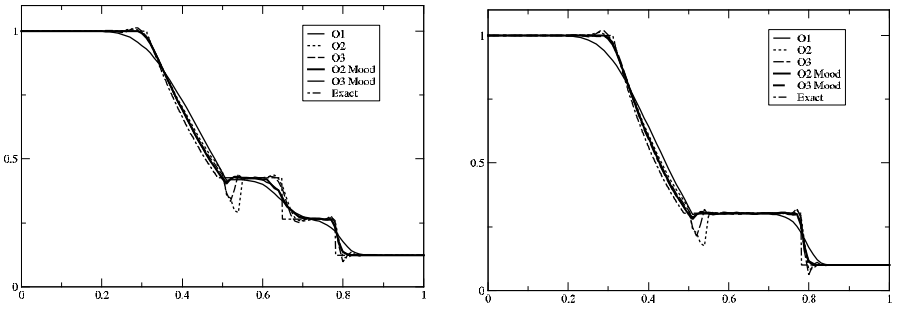
The last scalar example is the Buckley-Leverett equation

$$\frac{\partial u}{\partial t}+\frac{\partial f(u)}{\partial x}=0, \qquad f(u)=\frac{u^2}{u^2+(1-u)^2} \tag{6.3}$$

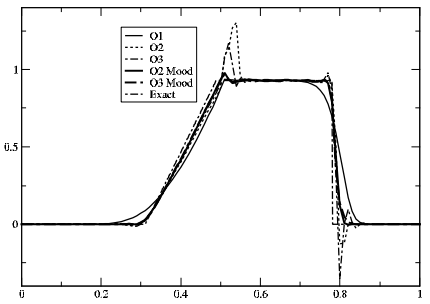


(a) unlimited O2-O4 solutions (b) first order and O2-O4 limited solutions

FIG. 6.2. *Buckley-Leverett problem with 100 points. A reference solution (first order with 10000 points) is also indicated.*



(a) Density (b) Pressure



(c) Velocity

FIG. 6.3. *Sod problem with 3-waves model: plot of the density, velocity and the pressure. Displayed solutions for order 1, 2 and 4 schemes with 100 points with and without MOOD. Also the exact solution is plotted.*

again with the same initial condition (6.1). The flux is nonconvex, so the problem is a bit more challenging.

We have run the simulations with 100 spatial points, until time $T=1$, with the 2 waves model we have considered. The first order (O1), second order (O2), fourth order (O4), second order with nonlinear stabilisation (O2M), and fourth order with

h	L^1	r	L^2	r	L^∞	r
First order						
50	0.374521524	-	0.529551387	-	0.337824076	-
100	0.222015068	1.22	0.313846916	1.22	0.200003594	1.21
200	0.121430904	1.23	0.171709701	1.30	0.109344706	1.3
400	$6.35740533\,10^{-2}$	1.34	$8.99051651\,10^{-2}$	1.34	$5.72395548\,10^{-2}$	1.34
800	$3.25359367\,10^{-2}$	1.36	$4.60121371\,10^{-2}$	1.36	$2.92929020\,10^{-2}$	1.36
Second order						
50	$9.76886451\,10^{-2}$	-	0.135240585	-	$8.88576061\,10^{-2}$	-
100	$2.43498404\,10^{-2}$	2.08	$3.40924263\,10^{-2}$	2.07	$2.20487341\,10^{-2}$	2.09
200	$6.06694631\,10^{-3}$	2.08	$8.53730459\,10^{-3}$	2.08	$5.47759095\,10^{-3}$	2.08
400	$1.51354610\,10^{-3}$	2.08	$2.13514664\,10^{-3}$	2.08	$1.36459176\,10^{-3}$	2.08
800	$3.77953198\,10^{-4}$	2.08	$5.33837010\,10^{-4}$	2.08	$3.40518804\,10^{-4}$	2.08
Fourth order						
50	$0.39627\,10^{-3}$	-	$0.554329\,10^{-3}$	-	$0.363964\,10^{-3}$	-
100	$0.24452\,7\,10^{-4}$	4.03	$0.342394\,10^{-4}$	4.01	$0.221427\,10^{-4}$	4.03
200	$0.151613\,10^{-5}$	4.01	$0.213344\,10^{-5}$	4.00	$0.136893\,10^{-5}$	4.01
400	$0.944521\,10^{-7}$	4.00	$0.133241\,10^{-6}$	4.00	$0.851587\,10^{-7}$	4.00
800	$0.589189\,10^{-9}$	4.00	$0.832214\,10^{-8}$	4.00	$0.530836\,10^{-8}$	4.00
Fourth order with Mood						
50	$0.39627\,10^{-3}$	-	$0.554329\,10^{-3}$	-	$0.363964\,10^{-3}$	-
100	$0.24452\,7\,10^{-4}$	4.03	$0.342394\,10^{-4}$	4.01	$0.221427\,10^{-4}$	4.03
200	$0.151613\,10^{-5}$	4.01	$0.213344\,10^{-5}$	4.00	$0.136893\,10^{-5}$	4.01
400	$0.944521\,10^{-7}$	4.00	$0.133241\,10^{-6}$	4.00	$0.851587\,10^{-7}$	4.00
800	$0.589189\,10^{-9}$	4.00	$0.832214\,10^{-8}$	4.00	$0.530836\,10^{-8}$	4.00

TABLE 6.2. *Order of convergence for the convection problem and two waves model for order 1, 2 and 4 with MOOD. The final time is* $T=10$. *The fourth order results with and without stabilisation are* identical *as expected.*

ε	0		10^{-6}		10^{-4}		10^{-3}		10^{-2}	
$\log\Delta x$	$\log L^2$	slope	$\log L^2$	slope	$\log L^2$	slope	$\log L^2$	slope	$\log L^2$	slope
-2.995	-1.332	-	-1.332	-	-1.332	-	-1.400	-	-2.049	-
-3.688	-2.655	1.908	-2.655	1.908	-2.655	1.907	-2.710	1.889	-3.314	1.825
-4.382	-4.030	1.984	-4.031	1.984	-4.031	1.984	-4.063	1.951	-4.647	1.922
-5.075	-5.415	1.997	-5.415	1.997	-5.415	1.996	-5.410	1.943	-5.999	1.951
-5.768	-6.801	1.999	-6.801	1.999	-6.800	1.998	-6.749	1.930	-7.365	1.971

TABLE 6.3. *Convection problem: error for the second order scheme with different* ε.

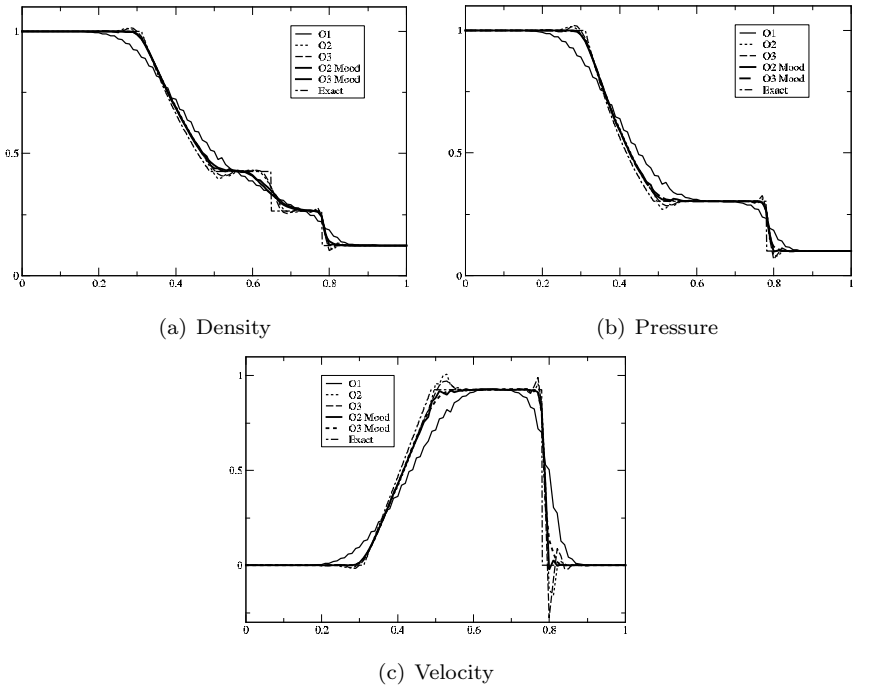
nonlinear stabilisation (O4M) are displayed in Figure 6.2, together with a reference solution computed with 1000 points and the first order scheme: remember that this corresponds to the Lax Friedrichs scheme, and it satisfies all entropy inequalities. This guaranties that the scheme converges. The nonlinearly stabilized solution has a correct behavior.

6.2. Uniform order of accuracy with respect to ε. In order to test the convergence of the scheme also in nonasymptotic regimes, we use again the scalar advection equation

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}=0 \tag{6.4}$$

ε	0		10^{-6}		10^{-4}		10^{-3}		10^{-2}	
$\log\Delta x$	$\log L^2$	slope	$\log L^2$	slope	$\log L^2$	slope	$\log L^2$	slope	$\log L^2$	slope
-2.995	-3.615	-	-3.615	-	-3.701	-	-3.701	-	-4.445	-
-3.688	-6.385	3.995	-6.385	3.996	-6.475	3.996	-6.475	4.001	-7.212	3.992
-4.382	-9.158	4.000	-9.159	4.001	-9.253	4.000	-9.253	4.007	-9.991	4.008
-5.075	-11.93	4.000	-11.93	3.999	-12.03	4.000	-12.03	4.007	-12.76	4.004
-5.768	-14.70	4.000	-14.70	4.000	-14.80	4.000	-14.81	4.006	-15.53	4.001

TABLE 6.4. *Convection problem: error for the fourth order scheme with different ε.*



(a) Density (b) Pressure

(c) Velocity

FIG. 6.4. *Sod problem with 2-waves model: plot of the density, velocity and the pressure. Displayed solutions for order 1, 2 and 4 schemes with 100 points with and without MOOD. Also the exact solution is plotted.*

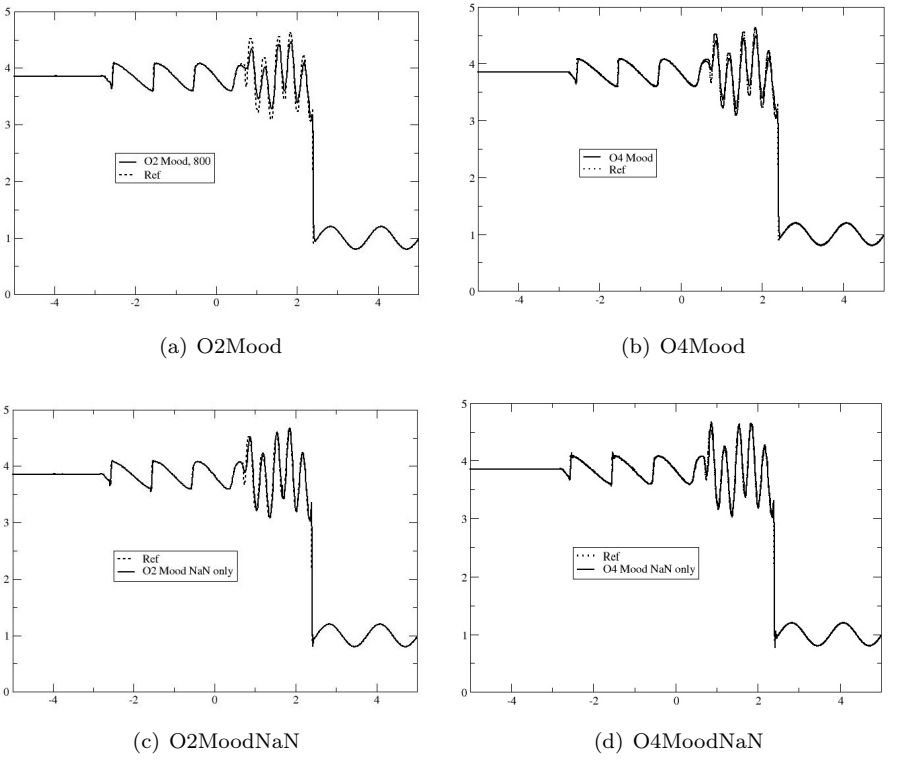
with initial condition

$$u(x,0) = \sin(2\pi x). \tag{6.5}$$

The CFL number is set to 1, the time and space order are set to 2 and 4, the final time is $T=1$, the boundary conditions are periodic. The values of ε are $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. We evaluate the order of convergence with the following standard procedure: if $u_{\Delta x}$, $u_{\Delta x/2}$ are the numerical solutions evaluated for consecutive meshes, the order α is, for the norm $\| \cdot \|$,

$$\alpha = \frac{\log \|u_{\Delta x} - u_{\Delta x/2}\|}{\log \Delta x}.$$

This test does not have an obvious result, as order reduction phenomena are common for IMEX schemes when the space discretization and the relaxation variable are of the same order. Nevertheless, we see on Tables 6.3 and 6.4 that the convergence order does



(a) O2Mood (b) O4Mood

(c) O2MoodNaN (d) O4MoodNaN

FIG. 6.5. *The reference solution is plotted with a dotted line. Comparison of various strategies of MOOD, for 800 points.*

not depend on ε. It is optimal. Also qualitatively, we see that for small enough values of ε, the solution obtained for $\varepsilon = 0$ is almost indistinguishable from $0 < \varepsilon \ll 1$.

6.3. Euler equations. In this section we test our scheme on Euler equations (4.3). We set $\gamma = 1.4$ and we run some standard cases: the Sod case and the Shu-Osher case.

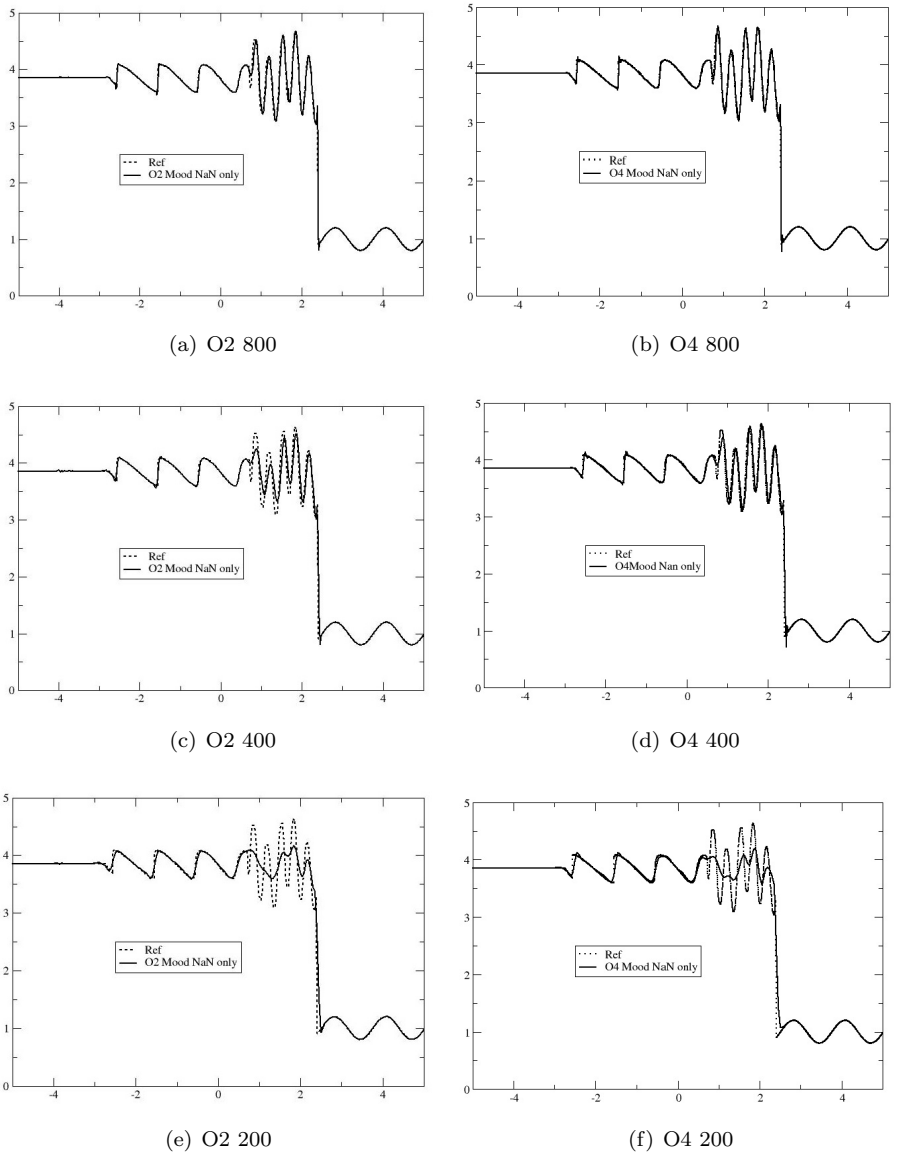
6.3.1. Sod test case. The Sod problem consists of a Riemann problem defined by the following initial conditions:

$$(\rho,u,p)^T = \begin{cases} (1,0,1)^T & \text{for } x < 0.5m \\ (0.125,0,0.1)^T & \text{else.} \end{cases}$$

The final time is $T = 0.16$. We have used the 3-waves model described above. The mesh resolution is of 100 elements, and the CFL is again 1 in all cases. From Figure 6.3, we see that the results are of good quality, at least compared with more standard methods.

For the sake of completeness, we have made the same simulation with the two waves model in Figure 6.4.

We see a stair case solution of the first order in space which is typical for the Lax-Friedrichs scheme. Comparing the solutions, the 2 waves model provides results of lower quality with respect to the 3 waves one. For that reason, we will not consider it anymore for the Euler equations.



(a) O2 800 (b) O4 800

(c) O2 400 (d) O4 400

(e) O2 200 (f) O4 200

FIG. 6.6. *O2 and O4 MOOD solutions with control of NaN only, for 200, 400 and 800 mesh points.*

6.3.2. Shu-Osher test case. The conditions of the Shu-Osher test are

$$(\rho,u,p)=\begin{cases}(3.857143,2.629369,10.3333333) & \text{if } x<-4,\\ (1+0.2\sin(5x),0,1) & \text{else.}\end{cases}$$

on the domain $[-5,5]$ and the final time of the problem is $T=1.8$. The reference solution is obtained with 10.000 points and fourth order limited scheme. It is difficult to see any modification in the solution if we use more grid points, this is why we consider this solution as the reference solution. We display only the solution with the MOOD

stabilisation technique, however, we have tried two different strategies. The figures labeled as OXMood, where X=2 or 4, use the full strategy of Section 5. The physical variables are the density and the pressure, nothing is tested on the velocity. In the figures labeled as OXMoodNaN, with X=2 or 4, we only check if the solution lies in the invariance domain, i.e., density and pressure stay positive and we do not encounter NaN values. In Figure 6.5 we plot the results for 800 points. In Figure 6.6 we compare results for 200, 400 and 800 mesh points only for the MoodNaN strategy.

From Figure 6.6, we see that with 800 points, there is hardly any difference between the O4MoodNaN solution and the reference one.

7. Conclusion

In this paper, simplifying a method described in [2], we show how to construct a class of kinetic numerical methods that can run at least at CFL 1. They can handle in a simple manner hyperbolic problems, and in particular the compressible fluid mechanics ones. These methods are always locally conservative and thus can handle discontinuities correctly. We have described a rather simple stabilisation mechanism which can be further improved or changed: it is not really the core of the proposed method. Our methodology can be arbitrarily high order and can use CFL number larger or equal to unity on regular Cartesian meshes. Extension to the multidimensional case will be the topic of future works. In particular, our implementation of these methods indicates that they can be potentially very fast. The parallelisation should also be straightforward. This, however, has to be confirmed in several spatial dimensions.

Acknowledgment. R.A. would like to thank Prof. Li-Shi Luo (Old Dominion, USA) and Professor Sagaut (Aix-Marseilles University, France) for their encouragement. Finally, he would like to thank Professors Lallemand and D'Humières for the discussion they had during his PhD studies. This paper shows that this has never been forgotten. I take this opportunity to also thank Prof. Chi-Wang Shu for very helpful discussions about time stepping. D.T. has been partially funded by ITN ModCompShock project funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 642768.

Appendix. Another iteration technique for the fourth-order-in-time case. The iteration (2.12) solved by (2.16) reads (when we have no source term or after the application of the projector operator):

$$v_1^{(p+1)} = u_i^n - \lambda\big(\theta_0^1 \delta u_i^n + \theta_1^1 \delta v_1^{(p)} + \theta_2^1 \delta v_2^{(p)}\big)$$
$$v_2^{(p+1)} = u_i^n - \lambda\big(\theta_0^2 \delta u_i^n + \theta_1^2 \delta v_1^{(p)} + \theta_2^2 \delta v_2^{(p)}\big)$$

and after the application of the Fourier transform, we have

$$\hat{v}^{(p+1)} = \hat{u}^n e - \lambda g \theta \mathbf{v}^{(p)}, \quad e = \begin{pmatrix} 1 - \theta_0^1 \lambda g \\ 1 - \theta_0^2 \lambda g \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1^1 & \theta_2^1 \\ \theta_1^2 & \theta_2^2 \end{pmatrix}$$

so that the amplification factor satisfies

$$G^{(p+1)} = e - \lambda g \theta G^{(p)}.$$

This can be seen as the Jacobi iteration for solving the system

$$(\mathrm{Id} + \lambda g \theta) G = e.$$

By analogy, we can define a Gauss-Seidel iteration by

$$v_1^{(p+1)} = u_i^n - \lambda\left(\theta_0^1 \delta u_i^n + \theta_1^1 \delta v_1^{(p)} + \theta_2^1 \delta v_2^{(p)}\right)$$
$$v_2^{(p+1)} = u_i^n - \lambda\left(\theta_0^2 \delta u_i^n + \theta_1^2 \delta v_1^{(p+1)} + \theta_2^2 \delta v_2^{(p)}\right)$$

whose Fourier transform is

$$\hat{v}^{(p+1)} = \hat{u}^n e - \lambda g \Theta_1 \hat{v}^{(p+1)} - \lambda g \Theta_2 \hat{v}^{(p)}, \quad \Theta_1 = \begin{pmatrix} 0 & 0 \\ \theta_1^2 & 0 \end{pmatrix}, \quad \Theta_2 = \begin{pmatrix} \theta_1^1 & \theta_2^1 \\ 0 & \theta_2^2 \end{pmatrix},$$

and hence

$$(\mathrm{Id} + \lambda g \Theta_1) G^{(p+1)} = e - \lambda g \Theta_2 G^{(p)}.$$

In both cases, $G^{(0)} = e$.

We can study the stability of the Gauss–Seidel iteration, and we recall the results of Jacobi's for comparison. Denoting by g_1 (resp. g_2, $g_{4,1}$, $g_{4,2}$) the Fourier symbol of the operators $a\delta_1$ (resp. $a\delta_2$, $a\delta_4^1$, $a\delta_4^2$), we get the results of Table A.1.

Iterations	1	2	3	4	5
Symbol	Gauss Seidel				
g_1	1.5	1.276906714	1.167201858	1.197067146,	1.152628955
g_2	0	≥ 1.65	≥ 1.47	≥ 1.435	≥ 1.55
g_4^1 [2]	0	≥ 0.926	≥ 1.775	0	0
g_4^2	0	≥ 0.917	0.8754013933	≥ 0.89	≥ 0.86
Symbol	Jacobi				
g_1	1	1	1.256372663	1.392646782	1.774161172
g_2	0	≥ 0.87	≥ 1.625	≥ 1.744	≥ 2.06
g_4^1 [3]	0	0	≥ 1.25	≥ 2.06	≥ 2.52
g_4^2	0	0	≥ 0.905	≥ 1.044	≥ 1.321

TABLE A.1. *CFL number for stability of the DeC iterations given by Gauss-Seidel and Jacobi methods. 0 means that the scheme is unconditionally unstable. x means that the scheme is stable up to CFL x, $\geq x$ means that the scheme is stable for at least CFL x (and slightly above).*

REMARK A.1 (A few remarks about Table A.1).

- For g_4^1, the amplification factor is always equal to 1 when $x = \pi$, and strictly below 1 under the condition stated above.
- For g_1 and 3 iterations, the CFL condition can be computed exactly. It is $\frac{1}{2}\sqrt[3]{4+\sqrt{17}} - \frac{1}{2}\frac{1}{\sqrt[3]{4+\sqrt{17}}} + \frac{1}{2} \approx 1.256372663$.

From these results, we see that there is no fundamental reason to prefer Gauss-Seidel iteration to the Jacobi one; the coding of the Gauss-Seidel is also slightly more involved. However, this conclusion holds true only for the schemes we have considered here, and might not be true for others.

[2] always 1 for $x = \pi$
[3] always 1 for $x = \pi$

REFERENCES

[1] R. Abgrall, *High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices*, J. Sci. Comput., 73(2-3):461–494, 2017. 2.1, 2.1.1

[2] R. Abgrall and D. Torlo, *High order asymptotic preserving deferred correction implicit-explicit schemes for kinetic models*, SIAM J. Sci. Comput., 42(3):B816–B845, 2020. 2.1.1, 2.1.2, 2.2, 7

[3] D. Aregba-Driollet and R. Natalini, *Discrete kinetic schemes for multidimensional systems of conservation laws*, SIAM J. Numer. Anal., 37(6):1973–2004, 2000. 1, 2.1.2

[4] M. Banda and M. Sead, *Relaxation WENO schemes for multi-dimensional hyperbolic systems of conservation laws*, Numer. Methods Partial. Differ. Equ., 23(5):1211–1234, 2007. 1

[5] S. Boscarino, L. Pareschi, and G. Russo, *Implicit-explicit Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit*, SIAM J. Sci. Comput., 35(1):A22–A51, 2013. 1

[6] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation laws*, J. Stat. Phys., 95(1/2):113–170, 1999. 1, 4

[7] S. Clain, S. Diot, and R. Loubère, *A high-order finite volume method for systems of conservation laws– multi-dimensional optimal order detection (MOOD)*, J. Comput. Phys., 230(10):4028–4050, 2011. 5, 5

[8] F. Coron and B. Perthame, *Numerical passage from kinetic to fluid equations*, SIAM J. Numer. Anal., 28(1):26–42, 1991. 2.1.2

[9] D. Coulette, E. Franck, P. Helluy, M. Mehrenberger, and L. Navoret, *High-order implicit palindromic discontinuous Galerkin method for kinetic-relaxation approximation*, Comput. Fluids, 190:485–502, 2019. 1

[10] A. Dutt, L. Greengard, and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT Numer. Math., 40(2):241–266, 2000. 2.1.1

[11] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, Springer Series in Comput. Math., Springer-Verlag, Berlin, 14, 2010. 2.1, 2.1.5

[12] A. Iserles, *Order stars and saturation theorem for first-order hyperbolics*, IMA J. Numer. Anal., 2:49–61, 1982. 2.2

[13] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Commun. Pure Appl. Math., 48(3):235–276, 1995. 1, 1

[14] D. Kuzmin, *A vertex-based hierarchical slope limiter for p-adaptive discontinuous Galerkin methods*, J. Comput. Appl. Math., 233(12):3077–3085, 2010. 5

[15] P. Lafitte, W. Melis, and G. Samaey, *A high-order relaxation method with projective integration for solving nonlinear systems of hyperbolic conservation laws*, J. Comput. Phys., 340:1–25, 2017. 1

[16] M.L. Minion, *Semi-implicit spectral deferred correction methods for ordinary differential equations*, Commun. Math. Sci., 1(3):471–500, 2003. 2.1.1

[17] R. Natalini, *A discrete kinetic approximation of entropy solution to multi-dimensional scalar conservation laws*, J. Differ. Equ., 148:292–317, 1998. 1, 4

[18] S. Pieraccini and G. Puppo, *Implicit–explicit schemes for BGK kinetic equations*, J. Sci. Comput., 32(1):1–28, 2007. 2.1.2

[19] H.J. Schroll, *High resolution relaxed upwind schemes in gas dynamics*, J. Sci. Comput., 17:599–607, 2002. 1

[20] R.J. Spiteri and S.J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40(2):469–491, 2002. 2.1

[21] B. van. Leer, *Flux-vector splitting for the Euler equations*, Technical Report NAS1-151810, ICASE, 507–512, 1982. 4

[22] F. Vilar, *A posteriori correction of high-order discontinuous Galerkin scheme through subcell finite volume formulation and flux reconstruction*, J. Comput. Phys., 387:245–279, 2019. 5, 5, 5

[23] T. Xiong, G. Russo, and J.-M. Qiu, *Conservative multi-dimensional semi-Lagrangian finite difference scheme: stability and applications to the kinetic and fluid simulations*, J. Sci. Comput., 79(2):1241–1270, 2019. 1