# Initialisation from lattice Boltzmann to multi-step Finite Difference methods: modified equations and discrete observability

Thomas Bellotti

# Initialisation from lattice Boltzmann to multi-step Finite Difference methods: modified equations and discrete observability

**Thomas Bellotti** (thomas.bellotti@polytechnique.edu)
CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau

August 4, 2023

**Abstract**

Latitude on the choice of initialisation is a shared feature between one-step extended state-space and multi-step methods. The paper focuses on lattice Boltzmann schemes, which can be interpreted as examples of both previous categories of numerical schemes. We propose a modified equation analysis of the initialisation schemes for lattice Boltzmann methods, determined by the choice of initial data. These modified equations provide guidelines to devise and analyze the initialisation in terms of order of consistency with respect to the target Cauchy problem and time smoothness of the numerical solution. In detail, the larger the number of matched terms between modified equations for initialisation and bulk methods, the smoother the obtained numerical solution. This is particularly manifest for numerical dissipation. Starting from the constraints to achieve time smoothness, which can quickly become prohibitive for they have to take the parasitic modes into consideration, we explain how the distinct lack of observability for certain lattice Boltzmann schemes—seen as dynamical systems on a commutative ring—can yield rather simple conditions and be easily studied as far as their initialisation is concerned. This comes from the reduced number of initialisation schemes at the fully discrete level. These theoretical results are successfully assessed on several lattice Boltzmann methods.

***Keywords—*** lattice Boltzmann; initialisation; Finite Difference; modified equations; observability; magic parameters.
***MSC—*** 65M75, 65M06, 65M15, 93B07

## 1 Introduction

Numerical analysis features two notable frameworks where the knowledge of the initial state for numerical schemes is incomplete: one-step extended state-space methods (*e.g.* relaxation schemes, kinetic schemes, gas-kinetic schemes, *etc.*) and multi-step methods. On the one hand, lattice Boltzmann schemes have historically been considered in the realm of the one-step extended state-space methods [Kuznik et al., 2013]. From this standpoint, they have previously been compared [Graille, 2014, Simonis et al., 2020] to approximations of systems of conservation laws taking the form of relaxation systems *à la* Jin-Xin [Jin and Xin, 1995] and interpreted as peculiar discretisations of these systems when collision and transport terms are split and the relaxation time tends to zero proportionally to the time step. Both in the relaxation systems and the lattice Boltzmann schemes, conserved and non-conserved quantities are present at the same time but only conserved ones appear in the original system of conservation laws at hand. Although the initialisation of the non-conserved quantities remains free in principle, it has important repercussions on the behaviour of the solution —such as the formation of time boundary layers—both for the relaxation systems and the lattice Boltzmann schemes. On the other hand, in recent works [Suga, 2010, Dellacherie, 2014, Fučík and Straka, 2021, Bellotti et al., 2022, Bellotti, 2023], lattice Boltzmann schemes have been thought and recast—as far as the evolution of the conserved quantities of interest is concerned—as multi-step Finite Difference schemes. Unsurprisingly, multi-step schemes both for Ordinary [Ascher and Petzold, 1998, Hairer et al., 2008, Hundsdorfer and Ruuth, 2006, Hundsdorfer et al., 2003] and Partial Differential Equations [Gustafsson et al., 1995, Strikwerda, 2004] need to be properly initialised by some starting procedure with desired features, for example, consistency. When lattice Boltzmann schemes are seen in their original formulation, where conserved and non-conserved moments mingle, the initialisation of the non-conserved moments can be freely devised. Once the lattice Boltzmann schemes are recast as corresponding multi-step Finite Difference schemes [Bellotti et al., 2022] solely on the conserved moments, the choice of initialisation for the conserved and non-conserved moments determines what the initialisation schemes feeding the corresponding bulk Finite Difference scheme at the beginning of the simulation are.

The previous discussion highlights that for numerical methods such as lattice Boltzmann schemes, the information gap between initial conditions for the target system of conservation laws and the numerical method must be filled and thus the issue of providing decision tools to throng this hollow clearly manifests. Furthermore, one must be careful when comparing numerical schemes to the continuous problem they aim at approximating, because: "Finite difference approximations have a more complicated "physics" than the equations they are designed to simulate. The irony is no paradox, however, for

finite differences are used not because the numbers they generate have simple properties, but because those numbers are simple to compute", see [Trefethen, 1996, Chapter 5]. Since the seminal paper of Warming and Hyett [Warming and Hyett, 1974], the method of the modified equation [Gustafsson et al., 1995, Strikwerda, 2004, Carpentier et al., 1997] has proved to be a valuable tool to describe such "complicated physics". Moreover, since lattice Boltzmann schemes (respectively, their corresponding Finite Difference schemes) feature non-physical moments (respectively, parasitic modes/eigenvalues), these terms play a role in the consistency of the initialisation routines—contrarily to what happens in the bulk—creating a rich yet complex dynamics.

In the framework of lattice Boltzmann schemes, previous efforts [Van Leemput et al., 2009] (under acoustic scaling), [Caiazzo, 2005, Junk and Yang, 2015, Huang et al., 2015] (under diffusive scaling) have provided the first guidelines to establish the initial conditions, relying on asymptotic expansions both on the conserved and non-conserved variables. One aim of these studies has been to suppress initial oscillating boundary layers being part of the "more complicated physics" of the discrete numerical method evoked in [Trefethen, 1996], which are however absent in the solution of the target conservation law. Even if the techniques introduced in these works guarantee the elimination of the initial oscillating boundary phenomena, no precise quantitative analysis of their inner structure has been presented. Moreover, since the non-conserved moments do not have an analogue in the continuous problem, these procedures are—despite the fact of providing good indications—intrinsically formal. Finally, these works have only addressed the initialisation of specific lattice Boltzmann schemes, namely the $\mathrm{D}_1\mathrm{Q}_2$ for [Van Leemput et al., 2009], the $\mathrm{D}_1\mathrm{Q}_2$ and $\mathrm{D}_1\mathrm{Q}_3$ for [Junk and Yang, 2015] and the $\mathrm{D}_2\mathrm{Q}_9$ for [Caiazzo, 2005, Huang et al., 2015].

Inspired by the open questions left by previous works in the literature, the present contribution aims at being the first general study on the initialisation of lattice Boltzmann schemes. The pivotal tool that we introduce is a modified equation analysis for the initial conditions/starting schemes and provides explicit constraints for general lattice Boltzmann schemes guaranteeing a sufficient order of consistency of the initialisation schemes to avoid order reduction of the overall method. The modified equations are obtained by considering that the choice of initial data shapes the starting schemes on the conserved variables of interest. Since the non-conserved moments are eliminated, the analyses we perform rely on less formal assumptions than the ones available in the literature. Pushing this tool one order further in the discretisation parameter, we meticulously describe the internal structure of the initial oscillating boundary layers, caused by incompatible numerical features—in particular, dissipation—between initialisation and bulk schemes. Previous works [Van Leemput et al., 2009] have just certified the existence of these oscillations in numerical simulations. Let us insist once again on the fact that the dissipation of the physical (or consistency) mode for the initialisation schemes is driven both by the physical and parasitic eigenvalues of the bulk Finite Difference scheme. Another novelty in our work is the characterisation—by seeing lattice Boltzmann methods as dynamical systems on a commutative ring and exploiting the concept of observability—of a vast well-known class of lattice Boltzmann schemes with a reduced number of initialisation schemes, irrespective of the number of non-conserved moments. The initial motivation to introduce the concept of observability is—for this class of schemes—to successfully determine the constraints needed to eliminate initial oscillating boundary layers due to the dissipation mismatch.

We consider lattice Boltzmann schemes with one conserved moment, for the sake of keeping the discussion and the notations simple and essential. The extension to several conserved moments can be envisioned in the spirit of our previous works [Bellotti et al., 2022, Bellotti, 2023]. We particularly concentrate on the widely adopted acoustic scaling [Dubois, 2021] between time and space steps. The diffusive scaling [Zhao and Yong, 2017, Zhang et al., 2019] is succinctly discussed with the very same techniques at the end of the work. Moreover, we consider linear schemes [Van Leemput et al., 2009], hence the equilibria for the non-conserved moments are linear functions of the conserved one. The lattice Boltzmann schemes we focus on aim at approximating the solution of the following linear Cauchy problem

$$\begin{cases} \partial_t u(t,\boldsymbol{x}) + \boldsymbol{V}\cdot\nabla_{\boldsymbol{x}} u(t,\boldsymbol{x}) = 0, & (t,\boldsymbol{x}) \in \mathbb{R}_+ \times \mathbb{R}^d, \quad (1) \\ u(0,\boldsymbol{x}) = u^\circ(\boldsymbol{x}), & \boldsymbol{x} \in \mathbb{R}^d, \quad (2) \end{cases}$$

with velocity $\boldsymbol{V} \in \mathbb{R}^d$ and initial datum $u^\circ$ which is a smooth function defined everywhere in $\mathbb{R}^d$. In this work, we only consider, contrarily to [Van Leemput et al., 2009], explicit initialisations, to keep the presentation simple. However, the analysis of implicit initialisations can be done with the same techniques.

The paper is structured as follows. At the beginning, Section 2 introduces the general lattice Boltzmann schemes treated in our study and Section 3 briefly recalls the main points about the reformulation of lattice Boltzmann schemes as Finite Difference schemes, away from the initial time. This reformulation has allowed [Bellotti, 2023] to rigorously study the consistency of lattice Boltzmann schemes apart from their initialisation and now characterises the number of needed initialisation schemes. In Section 4, we introduce the modified equation analysis of these starting schemes and find the constraints under which they are consistent with the same equation as the bulk Finite Difference scheme. The examples and numerical simulations of Section 5 are introduced to corroborate the theoretical findings of Section 4 and—pushing the computation of the modified equations of the starting schemes one order further—we describe the internal structure of the initial oscillating boundary layers. One particular scheme also stimulates the discussion of the following Section 6, where we re-evaluate the number of initialisation schemes at the discrete level more closely, thanks to the introduction

of the notion of observability for the lattice Boltzmann schemes. This allows to clearly identify and study a category of schemes for which the study of the initial conditions is greatly simplified and thus the constraints to avoid initial oscillating boundary layers can be easily established. We eventually conclude in Section 7.

## 2 Lattice Boltzmann schemes

To introduce the multiple-relaxation-times lattice Boltzmann schemes under the D'Humières formalism [D'Humières, 1992], which can be used to handle the previous Cauchy problem (1), (2) and the present paper is concerned by, one considers the following building blocks.

- Time and space steps $\Delta t$ and $\Delta x$, which are linked through a scaling. In the paper, since the target problem (1) is hyperbolic, the acoustic scaling $\Delta t = \Delta x/\lambda$ for a given fixed lattice velocity $\lambda > 0$ is taken, unless otherwise indicated.
- Discrete velocities $\boldsymbol{c}_1, \dots, \boldsymbol{c}_q \in \mathbb{Z}^d$, with $q \in \mathbb{N}^*$.
- An invertible moment matrix $\boldsymbol{M} \in \mathrm{GL}_q(\mathbb{R})$.
- A relaxation matrix $\boldsymbol{S} = \mathrm{diag}(s_1, s_2, \dots, s_q)$, where $s_i \in ]0,2]$ for $i \in [\![2, q]\!]$ and $s_1 \in \mathbb{R}$.
- The equilibrium coefficients $\boldsymbol{\epsilon} \in \mathbb{R}^q$ such that $\epsilon_1 = 1$, meaning that the first moment is conserved.

**Algorithm 1** Lattice Boltzmann scheme.

- Given $\mathbf{m}(0, \boldsymbol{x}) \in \mathbb{R}^q$ for every $\boldsymbol{x} \in \Delta x \mathbb{Z}^d$.
- For $n \in \mathbb{N}$
  - **Collision**. Using the collision matrix $\boldsymbol{K} := \boldsymbol{I} - \boldsymbol{S}(\boldsymbol{I} - \boldsymbol{\epsilon} \otimes \boldsymbol{e}_1)$, it reads
  $$\mathbf{m}^\star(n\Delta t, \boldsymbol{x}) = \boldsymbol{K}\mathbf{m}(n\Delta t, \boldsymbol{x}), \qquad \boldsymbol{x} \in \Delta x \mathbb{Z}^d. \tag{3}$$
  The post-collision distribution densities are recovered by $\mathbf{f}^\star(n\Delta t, \boldsymbol{x}) = \boldsymbol{M}^{-1}\mathbf{m}^\star(n\Delta t, \boldsymbol{x})$ on every point $\boldsymbol{x} \in \Delta x \mathbb{Z}^d$ of the lattice.
  - **Transport**, which reads
  $$\mathsf{f}_j((n+1)\Delta t, \boldsymbol{x}) = \mathsf{f}_j^\star(n\Delta t, \boldsymbol{x} - \Delta x \boldsymbol{c}_j), \qquad \boldsymbol{x} \in \Delta x \mathbb{Z}^d, \quad j \in [\![1, q]\!]. \tag{4}$$
  The moments at the new time step are obtained by $\mathbf{m}((n+1)\Delta t, \boldsymbol{x}) = \boldsymbol{M}\mathbf{f}((n+1)\Delta t, \boldsymbol{x})$ on every point $\boldsymbol{x} \in \Delta x \mathbb{Z}^d$.

The lattice Boltzmann scheme then reads as in Algorithm 1. The way of devising this algorithm—*i.e.* choosing the discrete velocities $\boldsymbol{c}_1, \dots, \boldsymbol{c}_q$, the moment matrix $\boldsymbol{M}$, the collision parameters $\boldsymbol{S}$ and the equilibrium coefficients $\boldsymbol{\epsilon}$—in order to obtain consistency with (1) has been formally studied with different strategies in a myriad of papers [Lallemand and Luo, 2000, Junk et al., 2005, Dubois, 2008, Yong et al., 2016, Dubois, 2021, Chai and Shi, 2020] to cite a few, and rigorously in our recent contribution [Bellotti, 2023], relying on an exact algebraic elimination of the non-conserved moments which are not present in (1). The present work particularly focuses on the choice of $\mathbf{m}(0, \cdot)$. For future use, we introduce the number $Q$ of non-conserved moments which do not relax to their equilibrium value during the collision phase (3):

$$Q := \#\{s_i \neq 1 \ : \ i \in [\![2, q]\!]\} \in [\![0, q[\![. \tag{5}$$

Roughly speaking, the larger $Q$, the stronger the "entanglement" between moments in the scheme. Remark that, since the corresponding column in $\boldsymbol{K}$ is zero, there is even no need to specify the initial value $\mathsf{m}_i(0, \cdot)$ when $s_i = 1$, for $i \in [\![2, q]\!]$. This comes from the fact that the post-collisional values of these moments are entirely determined by their values at equilibrium.

## 3 Corresponding Finite Difference schemes and initialisation schemes

Recasting the lattice Boltzmann scheme as a multi-step Finite Difference scheme [Bellotti et al., 2022] has allowed to rigorously define the notions of stability and consistency [Bellotti, 2023], without a precise consideration on the role of the initial conditions, which is indeed the aim of the present paper. In Section 3, we briefly recall the essential concepts on the way of rewriting lattice Boltzmann schemes as Finite Difference ones, showing that the choice of initial data determines what the initialisation schemes, used to start up the multi-step Finite Difference scheme, are.

## 3.1 Finite Difference scheme in the bulk

The transport phase (4) can be rewritten on the moments $\mathbf{m}$ by introducing the non-diagonal matrix of operators $\mathsf{T} := \boldsymbol{M}\mathrm{diag}(\mathbf{x}^{\boldsymbol{c}_1},\dots,\mathbf{x}^{\boldsymbol{c}_q})\boldsymbol{M}^{-1}$, using the multi-index notation, where $\mathbf{x} = (\mathsf{x}_1,\dots,\mathsf{x}_d)$ and thus $\mathbf{x}^{\boldsymbol{c}} = \mathsf{x}_1^{c_1}\cdots\mathsf{x}_d^{c_d}$ for any $\boldsymbol{c} \in \mathbb{Z}^d$. In this expression, we have considered the upwind space shift operators $\mathsf{x}_\ell$ for $\ell \in [\![1,d]\!]$ such that

$$(\mathsf{x}_\ell\phi)(\boldsymbol{x}) = \phi(\boldsymbol{x} - \Delta x\boldsymbol{e}_\ell), \qquad \boldsymbol{x} \in \mathbb{R}^d,$$

with $\boldsymbol{e}_\ell$ the $\ell$-th vector of the canonical basis. Considering also the forward time shift operator $\mathsf{z}$ such that

$$(\mathsf{z}\phi)(t) = \phi(t + \Delta t), \qquad t \in \mathbb{R},$$

the whole lattice Boltzmann scheme on the moments reads

$$(\mathsf{z}\mathbf{m})(t,\boldsymbol{x}) = (\mathbf{E}\mathbf{m})(t,\boldsymbol{x}), \qquad (t,\boldsymbol{x}) \in \Delta t\mathbb{N} \times \Delta x\mathbb{Z}^d, \tag{6}$$

where the evolution matrix of the scheme $\mathbf{E} := \mathsf{T}\boldsymbol{K} \in \mathcal{M}_q(\mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}])$ has entries in the ring of Laurent polynomials in the indeterminates $\mathsf{x}_1,\dots,\mathsf{x}_d$. In what follows, we shall remove the parenthesis to indicate the action of operators on functions, for the sake of notation. For any space operator $\mathsf{d} = \mathsf{d}(\mathbf{x}) \in \mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}]$, we define its conjugate operator by $\overline{\mathsf{d}} = \overline{\mathsf{d}}(\mathbf{x}) := \mathsf{d}(\mathbf{x}^{-1})$, which allows to introduce symmetric and anti-symmetric parts as $\mathsf{S}(\mathsf{d}) := (\mathsf{d} + \overline{\mathsf{d}})/2$ and $\mathsf{A}(\mathsf{d}) := (\mathsf{d} - \overline{\mathsf{d}})/2$. In [Bellotti et al., 2022], we have shown that—by means of the Cayley-Hamilton theorem on the commutative ring $\mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}]$ —the discrete dynamics of the conserved moment $\mathsf{m}_1$ computed by Algorithm 1 or by (6)—away from the initial time—is the one of the corresponding Finite Difference scheme under the form

$$\mathsf{z}^{Q+1-q}\det(\mathsf{z}\boldsymbol{I} - \mathbf{E})\mathsf{m}_1(t,\boldsymbol{x}) = \sum_{n=0}^{q} \mathsf{c}_n\mathsf{z}^{n+Q+1-q}\mathsf{m}_1(t,\boldsymbol{x}) = 0, \qquad (t,\boldsymbol{x}) \in \Delta t\mathbb{N} \times \Delta x\mathbb{Z}^d, \tag{7}$$

where $(\mathsf{c}_n)_{n=0}^{n=q} \subset \mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}]$ are the coefficients of the characteristic polynomial $\det(\mathsf{z}\boldsymbol{I} - \mathbf{E}) = \sum_{n=0}^{n=q}\mathsf{c}_n\mathsf{z}^n$ of $\mathbf{E}$, which is indeed the amplification polynomial of the scheme. One can easily see that $\mathsf{c}_n = 0$ for $n \in [\![0, q-Q-2]\!]$, whence the important role played by $Q$. Furthermore, since the characteristic polynomial is monic, *i.e.* $\mathsf{c}_q = 1$, the scheme is explicit, thus can be recast into

$$\mathsf{z}\mathsf{m}_1(t,\boldsymbol{x}) = -\sum_{n=q-Q-1}^{q-1} \mathsf{c}_n\mathsf{z}^{n+1-q}\mathsf{m}_1(t,\boldsymbol{x}), \qquad (t,\boldsymbol{x}) \in \Delta t[\![Q,+\infty[\![\times\Delta x\mathbb{Z}^d.$$

We call this scheme corresponding bulk Finite Difference scheme acting on the bulk time steps $[\![Q,+\infty[\![$, which is a multi-step scheme with $Q+2$ stages (or indeed $Q+1$ steps). We remark the need for initialisation data through $Q$ initialisation schemes, that we shall analyze in what follows.

## 3.2 Initialisation schemes

The initialisation schemes—the outcome of which eventually "nourishes" the bulk Finite Difference scheme—are determined by the choice of initial datum $\mathbf{m}(0,\cdot)$. They are:

$$\mathsf{m}_1(n\Delta t,\boldsymbol{x}) = (\mathbf{E}^n\mathbf{m})_1(0,\boldsymbol{x}), \qquad n \in [\![1,Q]\!], \quad \boldsymbol{x} \in \Delta x\mathbb{Z}^d.$$

The formulation we have proposed is provided in an abstract yet general form. In order to make the link with well-known lattice Boltzmann schemes and illustrate our purpose, let us introduce the following example. More of them are provided in Section 5 and Section 6.

*Example* 1 ($\mathrm{D_1Q_2}$). Consider the $\mathrm{D_1Q_2}$ as [Van Leemput et al., 2009, Dellacherie, 2014, Graille, 2014, Caetano et al., 2019], where $d = 1$, $q = 2$, $c_1 = 1$, $c_2 = -1$ and the moment matrix is

$$\boldsymbol{M} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \quad \text{Therefore} \quad \mathsf{T} = \begin{bmatrix} \mathsf{S}(\mathsf{x}_1) & \mathsf{A}(\mathsf{x}_1) \\ \mathsf{A}(\mathsf{x}_1) & \mathsf{S}(\mathsf{x}_1) \end{bmatrix}, \quad \text{and} \quad \boldsymbol{K} = \begin{bmatrix} 1 & 0 \\ s_2\epsilon_2 & 1-s_2 \end{bmatrix}.$$

The bulk Finite Difference scheme comes from $\det(\mathsf{z}\boldsymbol{I} - \mathbf{E}) = \mathsf{z}^2 + ((s_2-2)\mathsf{S}(\mathsf{x}_1) - s_2\epsilon_2\mathsf{A}(\mathsf{x}_1))\mathsf{z} + (1-s_2)$, encompassing the result from [Dellacherie, 2014], and hence reads

$$\mathsf{m}_1((n+1)\Delta t,x) = ((2-s_2)\mathsf{S}(\mathsf{x}_1) + s_2\epsilon_2\mathsf{A}(\mathsf{x}_1))\mathsf{m}_1(n\Delta t,x) + (s_2-1)\mathsf{m}_1((n-1)\Delta t,x), \tag{8}$$

for $n \in [\![Q,+\infty[\![$ and $x \in \Delta x\mathbb{Z}$. This is a Lax-Friedrichs scheme when $s_2 = 1$—which is first-order consistent with the transport equation at velocity $\lambda\epsilon_2$—and a leap-frog scheme when $s_2 = 2$, which is second-order consistent. Thus, to approximate the

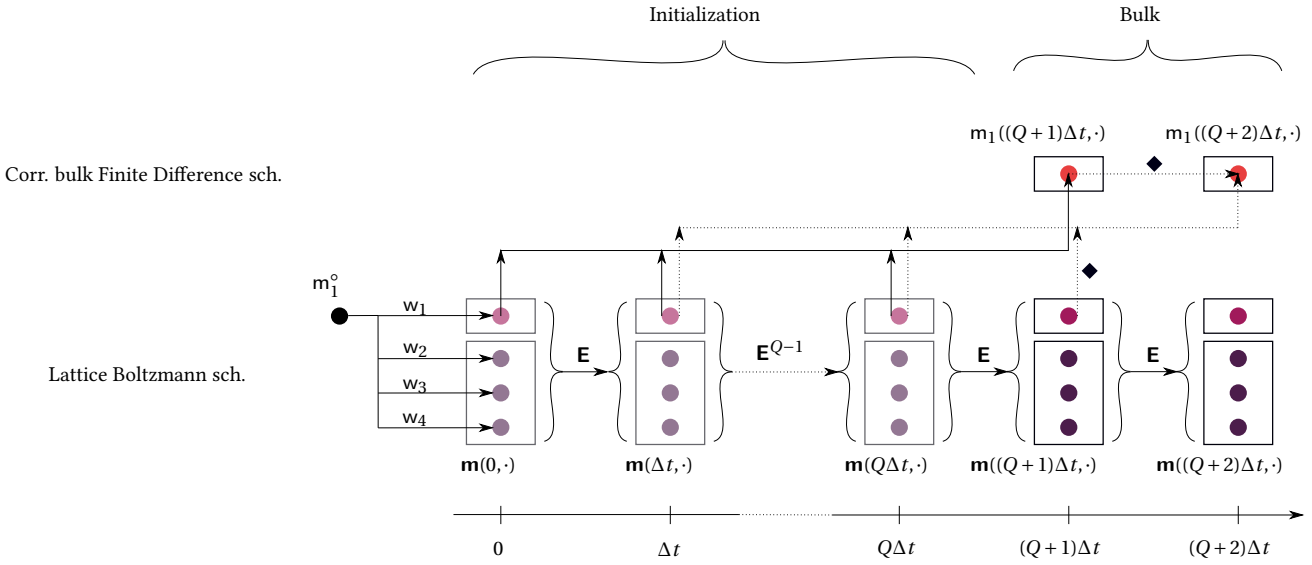


Figure 1: Illustration of the way of working of the lattice Boltzmann scheme (bottom) and the bulk Finite Difference scheme (top). The former acts both on the conserved (light violet) and the non-conserved (dark violet) moments. The latter implies only the conserved moment, drawn in light violet in the initialisation layer and in red in the bulk. Remark that to compute the conserved moment for the bulk Finite Difference scheme at time $(Q+2)\Delta t$, one can either rely on the information at time $(Q+1)\Delta t$ in light violet (from the lattice Boltzmann scheme) or on the one in red (from the Finite Difference scheme), as highlighted by the symbol ◆. This holds because these quantities are equal for any time step in the bulk for they stem from a common initialisation process. Partial transparency is used to denote the initialisation steps.

solution of (1) by $\mathsf{m}_1 \approx u$, the choice of equilibrium is $\epsilon_2 = V/\lambda$. The bulk Finite Difference scheme (8) is multi-step with $Q = 1$ when $s_2 \neq 1$: in this case, one needs to specify one initialisation scheme, which is

$$\mathsf{m}_1(\Delta t, x) = (\mathsf{S}(\mathsf{x}_1) + s_2\epsilon_2\mathsf{A}(\mathsf{x}_1))\mathsf{m}_1(0,x) + (1-s_2)\mathsf{A}(\mathsf{x}_1)\mathsf{m}_2(0,x), \qquad x \in \Delta x\mathbb{Z}.$$

We see that both the choice of the conserved moment $\mathsf{m}_1(0,\cdot)$ and the non-conserved moment $\mathsf{m}_2(0,\cdot)$ with respect to $u^\circ$ determine the initial scheme. Unsurprisingly, this scheme coincides with the bulk scheme when $s_2 = 1$.

## 3.3 Overall scheme

The bulk Finite Difference scheme supplemented by the initialisation schemes reads as in Algorithm 2. We stress that

**Algorithm 2** Corresponding Finite Difference scheme.

- Given $\mathbf{m}(0,\boldsymbol{x})$ for every $\boldsymbol{x} \in \Delta x\mathbb{Z}^d$.
- **Initialisation schemes**. For $n \in [\![1, Q]\!]$

$$\mathsf{m}_1(n\Delta t, \boldsymbol{x}) = (\mathbf{E}^n\mathbf{m})_1(0,\boldsymbol{x}), \qquad \boldsymbol{x} \in \Delta x\mathbb{Z}^d. \tag{9}$$

- **Corresponding bulk Finite Difference scheme**. For $n \in [\![Q, +\infty[\![$

$$\mathsf{m}_1((n+1)\Delta t, \boldsymbol{x}) = -\sum_{\ell=q-Q-1}^{q-1} \mathsf{c}_\ell \mathsf{m}_1((n+\ell+1-q)\Delta t, \boldsymbol{x}), \qquad \boldsymbol{x} \in \Delta x\mathbb{Z}^d. \tag{10}$$

Algorithm 2 is the corresponding scheme of Algorithm 1 in the sense that they issue the same discrete dynamics of the conserved moment $\mathsf{m}_1$ approximating $u$, see Figure 1. Of course, the non-conserved moments $\mathsf{m}_2, \dots, \mathsf{m}_q$ have been eliminated, at the price of handling a multi-step Finite Difference scheme. They still remain in the initialisation (*cf.* Example 1), giving a first intuition of why we claimed that non-physical modes—associated with non-conserved moments—play a role in this topic.

*Remark* 1. It is worthwhile observing that even if the initialisation schemes (9) are considered here close to the initial time, *i.e.* for $n \in [\![1, Q]\!]$, they also represent the action of the lattice Boltzmann scheme through its evolution operator $\mathbf{E}$ away from the initial time, that is, when $n > Q$. In the sequel, we shall employ the following nomenclature:

- "**initialisation schemes**", to indicate (9) for $n \in [\![1, Q]\!]$;
- "**starting schemes**", to indicate (9) for any $n \in \mathbb{N}^*$.

Hence, the initialisation schemes are a proper subset of the starting schemes. Indeed, in Section 4 and Section 5, we shall also consider the behaviour of (9) for $n > Q$, aiming at analysing the agreement between the behaviour of the numerical schemes inside the initial layer and the one purely in the bulk. This idea of matching is reminiscent of the singularly perturbed dynamical systems, see [O'Malley, 1991, Bender et al., 1999].

# 4 Modified equation analysis of the initial conditions under acoustic scaling

The study of the consistency of the initialisation schemes is crucial—especially when one wants to reach high-order accuracy. For the overall method, [Strikwerda, 2004, Theorem 10.6.2] states that, under acoustic scaling, if the initialisation of a stable multi-step scheme is obtained using schemes of accuracy $H-1$ in $\Delta x$, where $H$ is the accuracy of the multi-step scheme without accounting for the initialisation, then for smooth initial data, the order of accuracy of the multi-step scheme accounting for the initialisation remains $H$.

In what follows, we shall make use of the notion of asymptotic equivalence [Bellotti, 2023] between discrete operators in the ring $\mathbb{R}[\mathsf{z}, \mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$ and formal series of continuous differential operators.

**Definition 1** (Asymptotic equivalence)**.** Considering a scaling between $\Delta t$ and $\Delta x$, so that we take $\Delta x$ as driving discretisation parameter, for any discrete time-space operator $\mathsf{d} \in \mathbb{R}[\mathsf{z}, \mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$, we indicate $\mathsf{d} \asymp \delta$ where $\delta \in (\mathbb{R}[\partial_t, \partial_{x_1}, \dots, \partial_{x_d}])[\![\Delta x]\!]$ is a formal series in $\Delta x$ with coefficients in the ring of time-space differential operators $\mathbb{R}[\partial_t, \partial_{x_1}, \dots, \partial_{x_d}]$, if for every smooth $\phi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$

$$\mathsf{d}\phi(t, \boldsymbol{x}) = \sum_{h=0}^{+\infty} \Delta x^h \delta^{(h)} \phi(t, \boldsymbol{x}), \qquad (t, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^d,$$

in the limit $\Delta x \to 0$.

To perform the consistency analysis of the schemes *via* the modified equation [Warming and Hyett, 1974, Strikwerda, 2004, Gustafsson et al., 1995], one practical way of proceeding is to deploy the scheme on smooth functions over $\mathbb{R} \times \mathbb{R}^d$ instead of on grid functions defined over $\Delta t \mathbb{N} \times \Delta x \mathbb{Z}^d$, and use truncated asymptotic equivalents according to Definition 1. The scaling assumptions [Bellotti, 2023] the whole work will rely on are—unless further notice—that $\boldsymbol{M}$, $\boldsymbol{S}$ and $\boldsymbol{\epsilon}$ are independent of $\Delta x$ as $\Delta x \to 0$. Following [Dubois, 2021], we have introduced [Bellotti, 2023] the matrix of first-order space differential operators

$$\boldsymbol{\mathcal{G}} = \boldsymbol{M} \sum_{|\mathbf{n}|=1} \operatorname{diag}(\boldsymbol{c}_1^{\mathbf{n}}, \dots, \boldsymbol{c}_q^{\mathbf{n}}) \partial_{\boldsymbol{x}}^{\mathbf{n}} \boldsymbol{M}^{-1} \in \mathcal{M}_q(\mathbb{R}[\partial_t, \partial_{x_1}, \dots, \partial_{x_d}]),$$

influenced both by the choice of discrete velocities and the moment matrix at hand. The entries of this matrix shall be used to write the modified equations for general lattice Boltzmann schemes.

*Example* 2. Coming back to the context of Example 1, we have that

$$\boldsymbol{\mathcal{G}} = \begin{bmatrix} 0 & \partial_{x_1} \\ \partial_{x_1} & 0 \end{bmatrix}.$$

## 4.1 Review on the modified equation in the bulk

The consistency of the bulk Finite Difference scheme (10) is described in the following result.

**Theorem 1** ([Bellotti, 2023] Modified equation of the bulk scheme)**.** *Under acoustic scaling, that is, when $\lambda > 0$ is fixed as $\Delta x \to 0$, the modified equation for the bulk Finite Difference scheme* (10) *is given by*

$$\partial_t \phi(t, \boldsymbol{x}) + \lambda \Big( \mathcal{G}_{11} + \sum_{r=2}^{q} \mathcal{G}_{1r} \epsilon_r \Big) \phi(t, \boldsymbol{x}) - \lambda \Delta x \sum_{i=2}^{q} \Big( \frac{1}{s_i} - \frac{1}{2} \Big) \mathcal{G}_{1i} \Big( \mathcal{G}_{i1} + \sum_{r=2}^{q} \mathcal{G}_{ir} \epsilon_r - \Big( \mathcal{G}_{11} + \sum_{r=2}^{q} \mathcal{G}_{1r} \epsilon_r \Big) \epsilon_i \Big) \phi(t, \boldsymbol{x}) = O(\Delta x^2),$$

*for $(t, \boldsymbol{x}) \in \mathbb{R}_+ \times \mathbb{R}^d$.*

Comparing (11) and (1), the consistency with the equation of the Cauchy problem shall be enforced selecting the components of the lattice Boltzmann scheme such that $\lambda(\mathcal{G}_{11}+\sum_{r=2}^{r=q}\mathcal{G}_{1r}\epsilon_r)=\boldsymbol{V}\cdot\nabla_{\boldsymbol{x}}$. Since we shall employ the expression "at order $O(\Delta x^h)$" in the following discussion, let us specify what we mean, by taking advantage of the claim from Theorem 1. The terms $\partial_t$ and $\lambda(\mathcal{G}_{11}+\sum_{r=2}^{r=q}\mathcal{G}_{1r}\epsilon_r)$ appear at order $O(\Delta x)$ when the actual proof of Theorem 1 is done, thus we call them "$O(\Delta x)$ terms". Then, these terms appear at leading order in (11) because all the $O(1)$ terms simplify on both sides of the equation. The remaining term $O(\Delta x)$ in (11) originally shows at order $O(\Delta x^2)$ and is made up of numerical diffusion.

## 4.2 Linking the discrete initial datum with the one of the Cauchy problem

We now adapt the same techniques to concentrate on the role of the initial data. From the initial datum of the Cauchy problem $u^\circ$, we consider its point-wise discretisation with a lattice function $\mathsf{m}_1^\circ$ such that $\mathsf{m}_1^\circ(\boldsymbol{x})=u^\circ(\boldsymbol{x})$ for $\boldsymbol{x}\in\Delta x\mathbb{Z}^d$. Coherently with the fact of considering a linear problem and because the equilibria of the non-conserved moments are linear functions of the conserved one through $\boldsymbol{\epsilon}$, a linear initialisation reads

$$\mathbf{m}(0,\boldsymbol{x})=\mathbf{w}\mathsf{m}_1^\circ(\boldsymbol{x}),\qquad \boldsymbol{x}\in\Delta x\mathbb{Z}^d, \tag{11}$$

where $\mathbf{w}$ can be chosen in two different fashions.

- If $\mathbf{w}\in\mathbb{R}^q$ is considered, we obtain what we call "**local initialisation**". However, in order to gain more freedom on the initialisation and achieve desired numerical properties, another choice is possible.
- If $\mathbf{w}\in(\mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}])^q$ is considered, we obtain the "**prepared initialisation**", where we allow for an initial rearrangement of the information issued from the initial datum of the Cauchy problem between neighboring sites of the lattice.

It can be observed that the local initialisation is only a particular case of prepared initialisation using constant polynomials, since $\mathbb{R}$ is a sub-ring of $\mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}]$. By allowing $\mathsf{w}_1\in\mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}]$, we also permit to perform a preliminary modification of the point-wise discretisation of the initial datum (2) of the Cauchy problem, which can also be interpreted as an initial filtering of the datum, before assigning it to $\mathsf{m}_1$. For example, when $d=1$, considering $\mathsf{w}_1=\mathsf{S}(\mathsf{x}_1)$ yields $\mathsf{m}_1(0,x)=(u^\circ(x-\Delta x)+u^\circ(x+\Delta x))/2$ for every $x\in\Delta x\mathbb{Z}$. Observe that the following developments can be easily adapted to deal with implicit initialisations [Van Leemput et al., 2009] of the form $\mathsf{w}_i\mathsf{m}_i(0,\boldsymbol{x})=\mathsf{b}_i\mathsf{m}_1^\circ(\boldsymbol{x})$ with $\mathsf{b}_i\in\mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}]$ for $i\in[\![1,q]\!]$.

## 4.3 Modified equations for the initialisation schemes: local initialisation

Let us now compute the modified equations for the starting schemes when a local initialisation is considered. In the general framework, we shall stop at order $O(\Delta x)$ for two reasons. The first one is that we are not aware of any stable lattice Boltzmann scheme which—under acoustic scaling—would be third-order consistent in the bulk with the target equation (1) and therefore would call for second-order accurate initialisation schemes. Second, the expressions for higher order terms are excessively involved to be written down in a convenient form as functions of $n\in[\![1,Q]\!]$ for general schemes. Again, this is due to the role played by the non-physical eigenvalues of $\mathbf{E}$. Still, one more order in the expansion shall be needed to analyse the smooth initialisation proposed by [Van Leemput et al., 2009, Junk and Yang, 2015], as we shall do in Section 5 for some particularly simple yet instructive examples and for a more general class of schemes in Section 6.

**Proposition 1** (Modified equation of the starting schemes with local initialisation)**.** *Under acoustic scaling, that is, when $\lambda>0$ is fixed as $\Delta x\to 0$, considering a local initialisation,* i.e. $\mathbf{w}\in\mathbb{R}^q$*, the modified equations for the starting schemes are, for any $n\in\mathbb{N}^*$*

$$\begin{aligned}&\phi(0,\boldsymbol{x})+n\frac{\Delta x}{\lambda}\partial_t\phi(0,\boldsymbol{x})+O(\Delta x^2)\\ &=\mathsf{w}_1\phi(0,\boldsymbol{x})-n\Delta x\Big(\mathcal{G}_{11}\mathsf{w}_1+\sum_{r=2}^{q}\mathcal{G}_{1r}\mathsf{w}_r+\frac{1}{n}\sum_{r=2}^{q}\mathcal{G}_{1r}(\epsilon_r\mathsf{w}_1-\mathsf{w}_r)\sum_{\ell=0}^{n-1}\pi_{n-\ell}(s_r)\Big)\phi(0,\boldsymbol{x})+O(\Delta x^2),\quad \boldsymbol{x}\in\mathbb{R}^d,\end{aligned} \tag{12}$$

*where $\pi_\ell(X)=1-(1-X)^\ell$ for $\ell\in\mathbb{N}$.*

*Proof.* We start by describing the particular structure of the powers of collision matrix $\boldsymbol{K}$. It is straightforward to see that

we obtain an upper-triangular matrix with

$$\boldsymbol{K}^\ell = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ \pi_\ell(s_2)\epsilon_2 & (1-s_2)^\ell & 0 & & & \vdots \\ \pi_\ell(s_3)\epsilon_3 & 0 & (1-s_3)^\ell & \ddots & & \vdots \\ \pi_\ell(s_4)\epsilon_4 & 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ \pi_\ell(s_q)\epsilon_q & 0 & 0 & \cdots & 0 & (1-s_q)^\ell \end{bmatrix}, \qquad \ell \in \mathbb{N}^*, \tag{13}$$

where the polynomials $\pi_\ell$ are defined recursively as $\pi_0(X) := 0$ and $\pi_{\ell+1}(X) := X + (1-X)\pi_\ell(X)$ for $\ell \in \mathbb{N}$. Therefore $\pi_\ell(X) = 1 - (1-X)^\ell$ for $\ell \in \mathbb{N}$. The starting schemes read

$$\mathsf{z}^n \mathsf{m}_1(0, \boldsymbol{x}) = (\mathbf{E}^n \mathbf{w})_1 \mathsf{m}_1^\circ(\boldsymbol{x}), \qquad n \in \mathbb{N}^*, \quad \boldsymbol{x} \in \Delta x \mathbb{Z}^d. \tag{14}$$

Concerning the time shifts on the left hand side of (14), we have $\mathsf{z}^n \asymp \exp(n\frac{\Delta x}{\lambda}\partial_t) = 1 + n\frac{\Delta x}{\lambda}\partial_t + O(\Delta x^2)$ for $n \in \mathbb{N}$. For the right hand side of (14), we have that $\mathbf{E} \asymp \boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{T}}\boldsymbol{K}$ where $\mathbf{T} \asymp \boldsymbol{\mathcal{T}} = \exp(-\Delta x \boldsymbol{\mathcal{G}}) = \boldsymbol{I} - \Delta x \boldsymbol{\mathcal{G}} + O(\Delta x^2)$, see [Bellotti, 2023], and for $n \in \mathbb{N}^*$

$$\begin{aligned} \boldsymbol{\mathcal{E}}^n &= (\boldsymbol{\mathcal{E}}^{(0)} + \Delta x \boldsymbol{\mathcal{E}}^{(1)} + O(\Delta x^2))^n \\ &= (\boldsymbol{\mathcal{E}}^{(0)})^n + \Delta x \sum \{\text{permutations of } \boldsymbol{\mathcal{E}}^{(0)} \ (n-1 \text{ times}) \text{ and } \boldsymbol{\mathcal{E}}^{(1)} \ (\text{once})\} + O(\Delta x^2) \\ &= (\boldsymbol{\mathcal{E}}^{(0)})^n + \Delta x \sum_{\ell=0}^{\ell=n-1} (\boldsymbol{\mathcal{E}}^{(0)})^\ell \boldsymbol{\mathcal{E}}^{(1)} (\boldsymbol{\mathcal{E}}^{(0)})^{n-1-\ell} + O(\Delta x^2) = \boldsymbol{K}^n - \Delta x \sum_{\ell=0}^{\ell=n-1} \boldsymbol{K}^\ell \boldsymbol{\mathcal{G}} \boldsymbol{K}^{n-\ell} + O(\Delta x^2), \end{aligned} \tag{15}$$

where we use the fact that $\boldsymbol{\mathcal{E}}^{(h)} = \boldsymbol{\mathcal{T}}^{(h)} \boldsymbol{K}$ for $h \in \mathbb{N}$. Plugging into (14), employing a smooth function $\phi$ instead of $\mathsf{m}_1$ and $\mathsf{m}_1^\circ$ and using the fact that the initialisation is local, we have for $n \in \mathbb{N}^*$

$$\phi(0, \boldsymbol{x}) + n\frac{\Delta x}{\lambda}\partial_t \phi(0, \boldsymbol{x}) + O(\Delta x^2) = (\boldsymbol{K}^n \mathbf{w})_1 \phi(0, \boldsymbol{x}) - \Delta x \Big( \sum_{\ell=0}^{n-1} \boldsymbol{K}^\ell \boldsymbol{\mathcal{G}} \boldsymbol{K}^{n-\ell} \mathbf{w} \Big)_1 \phi(0, \boldsymbol{x}) + O(\Delta x^2), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

We have that $(\boldsymbol{K}^n \mathbf{w})_1 \phi(0, \boldsymbol{x}) = \mathsf{w}_1 \phi(0, \boldsymbol{x})$ thanks to (13) and for $j \in [\![1, q]\!]$

$$\begin{aligned} (\boldsymbol{K}^\ell \boldsymbol{\mathcal{G}} \boldsymbol{K}^{n-\ell})_{1j} = \sum_{p=1}^q \sum_{r=1}^q (\boldsymbol{K}^\ell)_{1p} \mathcal{G}_{pr} (\boldsymbol{K}^{n-\ell})_{rj} &= \sum_{r=1}^q \mathcal{G}_{1r} (\boldsymbol{K}^{n-\ell})_{rj} \\ &= \mathcal{G}_{11}\delta_{1j} + \sum_{r=2}^q \mathcal{G}_{1r} (\pi_{n-\ell}(s_r)\epsilon_r \delta_{1j} + (1-s_r)^{n-\ell}\delta_{rj}). \end{aligned}$$

Therefore for $n \in \mathbb{N}^*$

$$\begin{aligned} \sum_{\ell=0}^{n-1} (\boldsymbol{K}^\ell \boldsymbol{\mathcal{G}} \boldsymbol{K}^{n-\ell} \mathbf{w})_1 &= n\mathcal{G}_{11}\mathsf{w}_1 + \sum_{r=2}^q \mathcal{G}_{1r} \sum_{\ell=0}^{n-1} (\pi_{n-\ell}(s_r)\epsilon_r + (1-s_r)^{n-\ell}\mathsf{w}_r) \\ &= n\Big( \mathcal{G}_{11}\mathsf{w}_1 + \sum_{r=2}^q \mathcal{G}_{1r}\mathsf{w}_r + \frac{1}{n}\sum_{r=2}^q \mathcal{G}_{1r}(\epsilon_r \mathsf{w}_1 - \mathsf{w}_r) \sum_{\ell=0}^{n-1} \pi_{n-\ell}(s_r) \Big), \end{aligned}$$

where we have used that by the definition of $\pi_\ell$, $(n - \sum_{\ell=0}^{\ell=n-1} \pi_{n-\ell}(s_r)) / \sum_{\ell=0}^{\ell=n-1} (1-s_r)^{n-\ell} = 1$ for every $n \in \mathbb{N}^*$, yielding the claim. □

With Proposition 1, we can now compare the modified equation for the bulk Finite Difference scheme and the modified equations of the starting schemes, so respectively the dashed lines and the dots in Figure 2 at order $O(1)$ and $O(\Delta x)$.

## 4.4 Consistency of the initialisation schemes: local initialisation

The agreement between the terms at these two orders takes place under the following conditions.

**Corollary 1** (Consistency of the starting schemes with local initialisation)**.** *Under acoustic scaling, that is, when $\lambda > 0$ is fixed as $\Delta x \to 0$, considering a local initialisation,* i.e. $\mathbf{w} \in \mathbb{R}^q$*, under the conditions*

$$\mathsf{w}_1 = 1, \tag{16}$$

$$\text{for} \quad r \in [\![2, q]\!], \quad \text{if} \quad \mathcal{G}_{1r} \neq 0, \quad \text{then} \quad \mathsf{w}_r = \epsilon_r, \tag{17}$$

*where $\boldsymbol{\epsilon}$ are the equilibrium coefficients, the starting schemes are consistent with the modified equation* (11) *of the bulk Finite Difference scheme at order $O(\Delta x)$. Moreover, the initial datum feeding the bulk Finite Difference scheme and the starting schemes is consistent with the initial datum* (2) *of the Cauchy problem.*

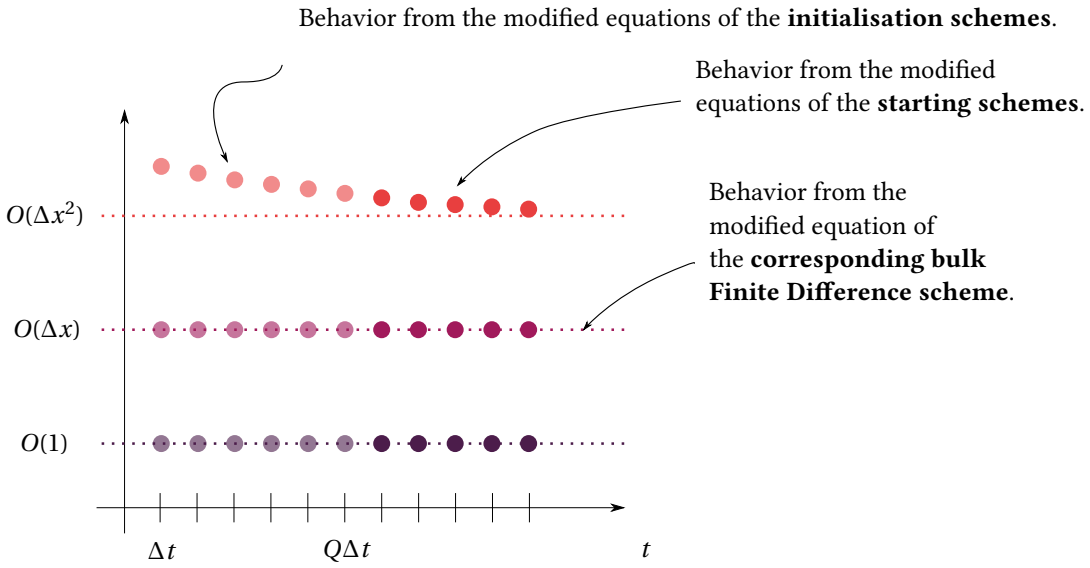


Figure 2: Example of behaviour of the inner expansion (dots, concerning the starting schemes) and the outer expansion (dashed lines, relative to the bulk Finite Difference scheme) at different orders in $\Delta x$ for $\Delta x \to 0$.

The first condition (16) implies that the initial datum for $\mathsf{m}_1$, used both by the starting schemes and the bulk Finite Difference scheme, is left untouched compared to the one of the Cauchy problem. The second condition (17) is expected: for the non-conserved moments involved in the modified equation (11) at leading order, we need to consider the initial datum at equilibrium. It is also to observe that this requirement does not *a priori* fix all the initialisation parameters, contrarily to Example 1, because some of them can affect only higher orders in the developments, *i.e.* $\mathcal{G}_{1r} = 0$ for some $r \in [\![1, q]\!]$.

*Proof of Corollary 1.* The proof proceeds order-by-order in $\Delta x$.

- $O(1)$. This order indicates that the initial datum for the conserved moment has to be consistent with the one of the Cauchy problem (2). From Proposition 1, it reads

$$\phi(0, \boldsymbol{x}) = \mathsf{w}_1 \phi(0, \boldsymbol{x}) + O(\Delta x), \qquad n \in \mathbb{N}^*, \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{18}$$

  hence we enforce $\mathsf{w}_1 = 1$. Remark that (18) is satisfied both for $n \in [\![1, Q]\!]$ and for $n > Q$, that is, both for initialisation schemes and starting schemes. This condition being fulfilled, the next order to check is

$$\partial_t \phi(0, \boldsymbol{x}) + \lambda \Big( \mathcal{G}_{11} + \sum_{r=2}^{q} \mathcal{G}_{1r} \mathsf{w}_r + \frac{1}{n} \sum_{r=2}^{q} \mathcal{G}_{1r} (\epsilon_r - \mathsf{w}_r) \sum_{\ell=0}^{n-1} \pi_{n-\ell}(s_r) \Big) = O(\Delta x), \qquad n \in \mathbb{N}^*, \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{19}$$

- $O(\Delta x)$. Evaluating the bulk modified equation (11) at time $t = 0$ gives

$$\partial_t \phi(0, \boldsymbol{x}) + \lambda \Big( \mathcal{G}_{11} + \sum_{r=2}^{q} \mathcal{G}_{1r} \epsilon_r \Big) \phi(0, \boldsymbol{x}) = O(\Delta x), \qquad \boldsymbol{x} \in \mathbb{R}^d, \tag{20}$$

  and trying to match each term with (19) yields the condition

$$\text{for} \quad r \in [\![2, q]\!], \quad \text{if} \quad \mathcal{G}_{1r} \neq 0, \quad \text{then} \quad \mathsf{w}_r = \epsilon_r.$$

□

## 4.5 Modified equations for the initialisation schemes: prepared initialisation

Now that the principles concerning the computation of modified equations for the starting schemes and the way of matching terms with the modified equation of the bulk Finite Difference scheme are clarified, we can tackle the case of prepared initialisations.

**Proposition 2** (Modified equation of the starting schemes with prepared initialisation)**.** *Under acoustic scaling, that is, when $\lambda > 0$ is fixed as $\Delta x \to 0$, considering a prepared initialisation,* i.e. $\mathbf{w} \in (\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}])^q$*, which can be put under the form*

$$\mathsf{w}_i = \sum_{\mathfrak{e}} w_{i,\mathfrak{e}} \mathbf{x}^{\mathfrak{e}}, \qquad i \in [\![1, q]\!], \tag{21}$$

*where the sequences of coefficients $(w_{i,\mathfrak{e}})_{\mathfrak{e}} \subset \mathbb{R}$ are compactly supported, the modified equations for the starting schemes are, for any $n \in \mathbb{N}^*$ and $\boldsymbol{x} \in \mathbb{R}^d$*

$$\begin{aligned}\phi(0,\boldsymbol{x}) + n\frac{\Delta x}{\lambda}\partial_t\phi(0,\boldsymbol{x}) + O(\Delta x^2) = \omega_1^{(0)}\phi(0,\boldsymbol{x}) \\ - n\Delta x\Big(\mathcal{G}_{11}\omega_1^{(0)} + \sum_{r=2}^{q}\mathcal{G}_{1r}\omega_r^{(0)} + \frac{1}{n}\sum_{r=2}^{q}\mathcal{G}_{1r}(\epsilon_r\omega_1^{(0)} - \omega_r^{(0)})\sum_{\ell=0}^{n-1}\pi_{n-\ell}(s_r) - \frac{1}{n}\omega_1^{(1)}\Big)\phi(0,\boldsymbol{x}) + O(\Delta x^2),\end{aligned}$$

*where*

$$\omega_i^{(0)} = \sum_{\mathfrak{e}} w_{i,\mathfrak{e}}, \qquad \omega_i^{(1)} = -\sum_{|\mathfrak{n}|=1}\Big(\sum_{\mathfrak{e}} w_{i,\mathfrak{e}}\mathfrak{e}^{\mathfrak{n}}\Big)\partial_{\boldsymbol{x}}^{\mathfrak{n}}, \qquad i \in [\![1, q]\!],$$

*and such that $\mathsf{w}_i \asymp \omega_i^{(0)} + \Delta x\omega_i^{(1)} + O(\Delta x)$ and $\pi_\ell(X) = 1 - (1-X)^\ell$ for $\ell \in \mathbb{N}$.*

*Proof.* The asymptotic equivalent of the initialisation $\mathbf{w}$ reads

$$\mathsf{w}_i \asymp \omega_i = \sum_{\mathfrak{e}} w_{i,\mathfrak{e}} - \Delta x\sum_{|\mathfrak{n}|=1}\Big(\sum_{\mathfrak{e}} w_{i,\mathfrak{e}}\mathfrak{e}^{\mathfrak{n}}\Big)\partial_{\boldsymbol{x}}^{\mathfrak{n}} + O(\Delta x^2), \qquad i \in [\![1, q]\!].$$

Using the Cauchy product between formal series, we have $\mathbf{E}^n\mathbf{w} \asymp \boldsymbol{\mathcal{E}}^n\boldsymbol{\omega} = (\boldsymbol{\mathcal{E}}^n)^{(0)}\boldsymbol{\omega}^{(0)} + \Delta x((\boldsymbol{\mathcal{E}}^n)^{(1)}\boldsymbol{\omega}^{(0)} + (\boldsymbol{\mathcal{E}}^n)^{(0)}\boldsymbol{\omega}^{(1)}) + O(\Delta x^2)$ for $n \in \mathbb{N}^*$. The $O(\Delta x)$ term in the previous expansion is made up of two contributions. The first one is $(\boldsymbol{\mathcal{E}}^n)^{(1)}\boldsymbol{\omega}^{(0)}$ and is not influenced by the "prepared" character of the initialisation, because it was also present for the local initialisation. The second one is inherent to the prepared initialisation. The result comes from the very same computations as Proposition 1. □

## 4.6 Consistency of the initialisation schemes: prepared initialisation

**Corollary 2** (Consistency of the starting schemes with prepared initialisation)**.** *Under acoustic scaling, that is, when $\lambda > 0$ is fixed as $\Delta x \to 0$, considering a prepared initialisation,* i.e. $\mathbf{w} \in (\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}])^q$*, with* (21)*, under the conditions*

$$\sum_{\mathfrak{e}} w_{1,\mathfrak{e}} = 1, \tag{22}$$

$$\text{for every} \quad |\mathfrak{n}| = 1, \quad \sum_{\mathfrak{e}} w_{1,\mathfrak{e}}\mathfrak{e}^{\mathfrak{n}} = 0, \tag{23}$$

$$\text{for} \quad r \in [\![2, q]\!], \quad \text{if} \quad \mathcal{G}_{1r} \neq 0, \quad \text{then} \quad \sum_{\mathfrak{e}} w_{r,\mathfrak{e}} = \epsilon_r, \tag{24}$$

*the starting schemes are consistent with the modified equation* (11) *of the bulk Finite Difference scheme at order $O(\Delta x)$. Moreover, the initial datum feeding the bulk Finite Difference scheme and the starting schemes is consistent with the initial datum* (2) *of the Cauchy problem up to order $O(\Delta x^2)$.*

Condition (22) is the analogue of (16). However, since the initialisation of the conserved moment can also be prepared, an additional condition (23) has to be taken into account. This guarantees, in particular, that the initial datum of the Cauchy problem used for $\mathsf{m}_1$ is not perturbed by some drift term at order $O(\Delta x)$. This is useful because of the multi-step nature of the bulk Finite Difference scheme (10), which shall also be fed with (11). Finally, (24) has to be compared with (17). This condition maintains that the non-conserved moments participating to the consistency at leading order have to be chosen—at leading order—at equilibrium.

*Proof of Corollary 2.* Proceeding order-by-order in $\Delta x$, we obtain:

- $O(1)$. The dominant order in the analogous of (12). Hence the consistency with the datum of the Cauchy problem reads $\omega_1^{(0)} = \sum_{\mathfrak{e}} w_{1,\mathfrak{e}} = 1$.
- $O(\Delta x)$. We see that now, there is the additional term associated with $\omega_1^{(1)}$ corresponding to a drift term in the initialisation of the conserved moment. In general, we now have wider possibilities in terms of how initialise, still remaining consistent with the modified equation of the bulk Finite Difference scheme at the desired order, at least for the initialisation schemes (*i.e.* $n \in [\![1, Q]\!]$). Indeed, it is sufficient to enforce that

$$\mathcal{G}_{11} + \sum_{r=2}^{q}\mathcal{G}_{1r}\omega_r^{(0)} + \frac{1}{n}\sum_{r=2}^{q}\mathcal{G}_{1r}(\epsilon_r - \omega_r^{(0)})\sum_{\ell=0}^{n-1}\pi_{n-\ell}(s_r) - \frac{1}{n}\omega_1^{(1)} = \mathcal{G}_{11} + \sum_{r=2}^{q}\mathcal{G}_{1r}\epsilon_r.$$

Occasionally, for some $n \in [\![1, Q]\!]$, the previous inequality can be satisfied even if $\omega_1^{(1)} \neq 0$, see examples in Section 5. However, we are interested in enforcing it for every for every $n \in \mathbb{N}^*$—that is—for all starting schemes. This comes, as previously claimed, from the multi-step nature of the bulk Finite Difference scheme: we have to ensure that the order of consistency with the initial datum (2) of the Cauchy problem is high enough not to lower the overall order of the method. Hence, suppressing the drift term for the conserved moment, thus enforcing $\omega_1^{(1)} = 0$, we have

$$\text{for every} \quad |\mathfrak{n}| = 1, \quad \sum_{\mathfrak{e}} w_{1,\mathfrak{e}} \mathfrak{e}^{\mathfrak{n}} = 0, \quad \text{and} \quad \text{for} \quad r \in [\![2, q]\!], \quad \text{if} \quad \mathcal{G}_{1r} \neq 0, \quad \text{then} \quad \sum_{\mathfrak{e}} w_{r,\mathfrak{e}} = \epsilon_r.$$

□

## 4.7 Initialisation schemes *versus* starting schemes

Before proceeding to some numerical illustrations, we point out important facts concerning the match of terms between the bulk Finite Difference scheme and the initialisation schemes/starting schemes.

**Proposition 3** (Control on the initialisation schemes leads control on the starting schemes)**.** *Let $H \in \mathbb{N}^*$. Assume that*

- $\omega_1^{(0)} = 1$ *and* $\omega_1^{(h)} = 0$ *for* $h \in [\![1, H]\!]$.
- *The modified equations of the initialisation schemes (*(9) *for $n \in [\![1, Q]\!]$) match the one of the bulk Finite Difference scheme* (10) *at any order $h \in [\![1, H]\!]$.*

*Then, the modified equations of the starting schemes (*(9) *for $n > Q$) match the one of the bulk Finite Difference scheme* (10) *at any order $h \in [\![1, H]\!]$.*

Proposition 3 does not provide indications on how to equate the order at $h \in [\![1, H]\!]$—*i.e.* how to fulfill its assumptions—contrarily to what Corollary 1 and Corollary 2 do for $H = 1$. Again, this is due to the fact that the general expression of the asymptotic expansion of $(\mathbf{E}^n \mathbf{w})_1$ can quickly become messy as the considered order increases. Still, Proposition 3 claims that if one is able to match the modified equation of the initialisation schemes with the one of the bulk Finite Difference scheme until a given order $H$ (as we shall do for specific schemes in Section 5 with $H = 2$), then this guarantees the same property on the starting schemes. Otherwise said—referring to Figure 2—if one is able to ensure that the terms represented by the dots lie on the corresponding dashed line for $n \in [\![1, Q]\!]$, then one will be sure that these dots will lie on the very same line for any $n \in \mathbb{N}^*$. This result seems intuitively reasonable by virtue of the Cayley-Hamilton theorem, which allows to recast any power $\mathbf{E}^n$ for $n \geq Q + 1$ as combination of $\boldsymbol{I}, \mathbf{E}, \ldots, \mathbf{E}^Q$.

*Proof of Proposition 3.* Let us consider $d = 1$ for the sake of notation: for $d > 1$, the multi-index notation would suffice. Consider a one-step linear Finite Difference scheme on the lattice function u, under the form $\mathsf{z}\mathsf{u}(t, x) = \mathsf{g}_1 \mathsf{u}(t, x)$ for $(t, x) \in \Delta t\mathbb{N} \times \Delta x\mathbb{Z}$, where $\mathsf{g}_1 \in \mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}]$. This can rewritten using the Fourier transform in space, that is

$$\mathsf{z}\hat{\mathsf{u}}(t, \xi\Delta x) = \hat{\mathsf{g}}_1(\xi\Delta x)\hat{\mathsf{u}}(t, \xi\Delta x), \qquad (t, \xi) \in \Delta t\mathbb{N} \times [-\pi/\Delta x, \pi/\Delta x]. \tag{25}$$

The frequency-dependent eigenvalue $\hat{\mathsf{g}}_1(\xi\Delta x) \in \mathbb{C}$ shall be a Laurent polynomial in the indeterminate $e^{i\xi\Delta x}$ and encodes both the stability features of the method, for every $\xi \in [-\pi/\Delta x, \pi/\Delta x]$ and the consistency features, in the low-frequency limit $|\xi\Delta x| \ll 1$. In particular, to be consistent with an equation of the form (1) with a first-order derivative in time, one can easily see that

$$\hat{\mathsf{g}}_1(\xi\Delta x) = 1 + O(|\xi\Delta x|) \qquad \text{in the limit} \quad |\xi\Delta x| \ll 1. \tag{26}$$

Applying the scheme (25) $n \in \mathbb{N}^*$ times provides a sort of multi-step scheme which we shall compare to the starting schemes (9)

$$\mathsf{z}^n \hat{\mathsf{u}}(t, \xi\Delta x) = \hat{\mathsf{g}}_1(\xi\Delta x)^n \hat{\mathsf{u}}(t, \xi\Delta x), \qquad (t, \xi) \in \Delta t\mathbb{N} \times [-\pi/\Delta x, \pi/\Delta x], \tag{27}$$

with associated amplification polynomial

$$\hat{\tilde{\Phi}}(\mathsf{z}, \xi\Delta x) = \mathsf{z}^n - \hat{\mathsf{g}}_1(\xi\Delta x)^n = (\mathsf{z} - \hat{\mathsf{g}}_1(\xi\Delta x)) \sum_{\ell=0}^{n-1} \hat{\mathsf{g}}_1(\xi\Delta x)^\ell \mathsf{z}^{n-1-\ell} = (\mathsf{z} - \hat{\mathsf{g}}_1(\xi\Delta x)) \prod_{\ell=2}^{n} (\mathsf{z} - \hat{\tilde{\mathsf{g}}}_\ell(\xi\Delta x)), \tag{28}$$

having roots $\hat{\tilde{\mathsf{g}}}_1 = \hat{\mathsf{g}}_1, \hat{\tilde{\mathsf{g}}}_2, \ldots, \hat{\tilde{\mathsf{g}}}_n$ (recall that $\mathbb{C}$ is an algebraically closed field). By differentiating th amplification polynomial (28) using the rule for the derivative of a product, we get

$$\frac{\mathrm{d}\hat{\tilde{\Phi}}(\mathsf{z}, \xi\Delta x)}{\mathrm{d}\mathsf{z}} = n\mathsf{z}^{n-1} = \prod_{\ell=2}^{n} (\mathsf{z} - \hat{\tilde{\mathsf{g}}}_\ell(\xi\Delta x)) + (\mathsf{z} - \hat{\mathsf{g}}_1(\xi\Delta x)) + \sum_{p=2}^{n} \prod_{\substack{\ell=2 \\ \ell \neq p}}^{n} (\mathsf{z} - \hat{\tilde{\mathsf{g}}}_\ell(\xi\Delta x)).$$

Taking $\mathsf{z}=1$ in the limit $|\xi\Delta x|\ll 1$ gives $0\neq n=\prod_{\ell=2}^{\ell=n}(1-\hat{\tilde{g}}_\ell^{(0)})$, where $\hat{\tilde{g}}_\ell(\xi\Delta x)=\hat{\tilde{g}}_\ell^{(0)}+O(|\xi\Delta x|)$, thanks to (26), thus all the other eigenvalues $\hat{\tilde{g}}_2,\dots,\hat{\tilde{g}}_n$ are not equal to one for small frequencies and thus are not linked with consistency. The only which matters is $\hat{\tilde{g}}_1=\hat{g}_1$, thus the scheme (27) with amplification polynomial (28) has the same modified equations as (25). An alternative way of seeing this is to use the approach from the proof of [Carpentier et al., 1997, Proposition 1], which aims at automatically handling the "reinjection" of the previous orders in the expansions to eliminate time derivatives above first order. Inserting the asymptotic equivalent $\exp(n\frac{\Delta x}{\lambda}\partial_t)\asymp\mathsf{z}^n$ into (27) using a smooth "test" function $\hat{\phi}$ gives $\exp(n\frac{\Delta x}{\lambda}\partial_t)\hat{\phi}(t,\xi)=\hat{g}_1(\xi\Delta x)^n\hat{\phi}(t,\xi)$ for $(t,\xi)\in\mathbb{R}_+\times\mathbb{R}$, which means that if we do not want $\hat{\phi}$ to trivially vanish, we must enforce the formal identity $\exp(n\frac{\Delta x}{\lambda}\partial_t)=\hat{g}_1(\xi\Delta x)^n$. Since the exponential is bijective close to zero (here we are considering the limit $|\xi\Delta x|\ll 1$), we can take the logarithm to yield:

$$\partial_t=\frac{n}{n}\frac{\lambda}{\Delta x}\log(\hat{g}_1(\xi\Delta x)),$$

which is thus independent of $n$.

Differently, a $(Q+2)$-stage Finite Difference scheme, like the bulk Finite Difference scheme (10), has associated amplification polynomial

$$\hat{\Phi}(\mathsf{z},\xi\Delta x):=\frac{1}{\mathsf{z}^{q-Q-1}}\det(\mathsf{z}\boldsymbol{I}-\hat{\mathbf{E}}(\xi\Delta x))=\mathsf{z}^{Q+1}+\sum_{n=0}^{Q}\hat{\mathsf{c}}_{n+q-Q-1}(\xi\Delta x)\mathsf{z}^n=\prod_{\ell=1}^{Q+1}(\mathsf{z}-\hat{g}_\ell(\xi\Delta x)).\tag{29}$$

Out of the roots in (29), we shall number the (unique) eigenvalue providing the modified equation (11), *i.e.* such that (26) holds, by $\hat{g}_1$. This is the amplification factor of the so-called "pseudo-scheme" [Strikwerda, 2004]. Furthermore, the higher-order terms in the modified equation of the bulk Finite Difference scheme stem from $\hat{g}_1(\xi\Delta x)=1+\sum_{h=1}^{h=H}(\xi\Delta x)^h\hat{g}_1^{(h)}+O(|\xi\Delta x|^{H+1})$ in the limit $|\xi\Delta x|\ll 1$. The initialisation schemes read

$$\mathsf{z}^n\hat{\mathsf{m}}_1(0,\xi\Delta x)=\underbrace{(\hat{\mathbf{E}}(\xi\Delta x)^n\hat{\mathbf{w}}(\xi\Delta x))_1}_{=:\hat{\mathsf{g}}^{[n]}(\xi\Delta x)}\hat{\mathsf{m}}_1^\circ(\xi\Delta x),\qquad n\in[\![1,Q]\!],\tag{30}$$

with $\xi\in[-\pi/\Delta x,\pi/\Delta x]$. Using the assumption that $\omega_1^{(0)}=1$, the proof of Proposition 2 naturally entails that $\hat{\mathsf{g}}^{[n]}(\xi\Delta x)=1+O(|\xi\Delta x|)$ for $|\xi\Delta x|\ll 1$. Comparing (27) and (30), we cannot employ the same trick without a deeper discussion. We have

$$\partial_t=\frac{\lambda}{\Delta x}\log(\hat{g}_1(\xi\Delta x))\qquad\text{and}\qquad\partial_t=\frac{1}{n}\frac{\lambda}{\Delta x}\log(\hat{\mathsf{g}}^{[n]}(\xi\Delta x)),\qquad n\in[\![1,Q]\!],$$

where the first equation comes from the modified equation of the bulk Finite Difference scheme and the second one from (30). Since the initialisation schemes and the bulk Finite Difference scheme have the same modified equations up to order $H$, then we have, in the limit $|\xi\Delta x|\ll 1$

$$n\log(\hat{g}_1(\xi\Delta x))=\log((\hat{g}_1(\xi\Delta x))^n)=\log(\hat{\mathsf{g}}^{[n]}(\xi\Delta x))+O(|\xi\Delta x|^{H+1}),\qquad n\in[\![1,Q]\!],$$

hence we deduce that $\hat{\mathsf{g}}^{[n]}(\xi\Delta x)=\hat{g}_1(\xi\Delta x)^n+O(|\xi\Delta x|^{H+1})$ for $n\in[\![1,Q]\!]$. To finish the proof, we now consider (9) for $n=Q+1$

$$\mathsf{z}^{Q+1}\hat{\mathsf{m}}_1(0,\xi\Delta x)=-\sum_{n=0}^{Q}\hat{\mathsf{c}}_{n+q-Q-1}(\xi\Delta x)\mathsf{z}^n\hat{\mathsf{m}}_1(0,\xi\Delta x)\tag{31}$$

$$=\underbrace{(\hat{\mathbf{E}}(\xi\Delta x)^{Q+1}\hat{\mathbf{w}}(\xi\Delta x))_1}_{=:\hat{\mathsf{g}}^{[Q+1]}(\xi\Delta x)}\hat{\mathsf{m}}_1^\circ(\xi\Delta x).\tag{32}$$

We compute the modified equation of (32), yielding the thesis, by using (31). We have

$$\begin{aligned}\mathsf{z}^{Q+1}\hat{\mathsf{m}}_1(0,\xi\Delta x)&=-\sum_{n=1}^{Q}\hat{\mathsf{c}}_{n+q-Q-1}(\xi\Delta x)\mathsf{z}^n\hat{\mathsf{m}}_1(0,\xi\Delta x)-\hat{\mathsf{c}}_{q-Q-1}(\xi\Delta x)\hat{\mathsf{w}}_1(\xi\Delta x)\hat{\mathsf{m}}_1^\circ(\xi\Delta x)\\&=-\sum_{n=1}^{Q}\hat{\mathsf{c}}_{n+q-Q-1}(\xi\Delta x)\hat{\mathsf{g}}^{[n]}(\xi\Delta x)\hat{\mathsf{m}}_1^\circ(\xi\Delta x)-\hat{\mathsf{c}}_{q-Q-1}(\xi\Delta x)\hat{\mathsf{w}}_1(\xi\Delta x)\hat{\mathsf{m}}_1^\circ(\xi\Delta x).\end{aligned}$$

In the limit $|\xi\Delta x|\ll 1$, we have $\hat{\mathsf{w}}_1(\xi\Delta x)=1+O(|\xi\Delta x|^{H+1})$ and $\hat{\mathsf{g}}^{[n]}(\xi\Delta x)=\hat{\mathsf{g}}_1(\xi\Delta x)^n+O(|\xi\Delta x|^{H+1})$ for $n\in[\![1,Q]\!]$, thanks to the assumption on $\mathsf{w}_1$ and to the previous computations. In the limit $|\xi\Delta x|\ll 1$, we have to consider the amplification polynomial

$$\hat{\tilde{\Phi}}(\mathsf{z},\xi\Delta x)=\mathsf{z}^{Q+1}+\sum_{n=0}^{Q}\hat{\mathsf{c}}_{n+q-Q-1}(\xi\Delta x)\hat{\mathsf{g}}_1(\xi\Delta x)^n+O(|\xi\Delta x|^{H+1})=\mathsf{z}^{Q+1}-\hat{\mathsf{g}}_1(\xi\Delta x)^{Q+1}+O(|\xi\Delta x|^{H+1}),$$

using the fact that $\hat{\mathsf{g}}_1$ is a root of (29). We are therefore, up to terms $O(|\xi\Delta x|^{H+1})$, in the same setting as (27) and (28), hence with the usual trick, we gain

$$\partial_t = \frac{Q+1}{Q+1}\frac{\lambda}{\Delta x}\log(\hat{\mathsf{g}}_1(\xi\Delta x)) + O(|\xi\Delta x|^{H+1}),$$

hence also that $\hat{\mathsf{g}}^{[Q+1]}(\xi\Delta x) = \hat{\mathsf{g}}_1(\xi\Delta x)^{Q+1} + O(|\xi\Delta x|^{H+1})$. This concludes the proof. The case $n > Q+1$ is done analogously. □

In order to understand why it is difficult to study the starting schemes above first-order in full generality, we consider the Fourier's setting introduced in the previous proof in the case $d = 1$. To isolate the role of each initialisation scheme, we introduce the $\ell$-th time Green functions $\hat{\mathsf{G}}_\ell^{[n]} = \hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)$, for $\ell \in [\![0, Q]\!]$, see [Cheng and Lu, 1999], defined by

$$\begin{cases} \hat{\mathsf{G}}_\ell^{[n+1]} = -\sum_{p=q-Q-1}^{p=q-1} \hat{\mathsf{c}}_p \hat{\mathsf{G}}_\ell^{[n+p-q+1]} = \boldsymbol{e}_1^{\mathsf{t}} \hat{\mathbf{Q}} \left[\hat{\mathsf{G}}_\ell^{[n]}, \dots, \hat{\mathsf{G}}_\ell^{[n-Q]}\right]^{\mathsf{t}}, & \text{for} \quad n \geq Q, \\ \hat{\mathsf{G}}_\ell^{[p]} = \delta_{\ell,p}, & \text{for} \quad p \in [\![0, Q]\!], \end{cases}$$

where $\mathbf{Q}$ is the companion matrix of size $Q$ associated with the monic polynomial $\mathsf{z}^{Q+1-q}\det(\mathsf{z}\boldsymbol{I} - \mathbf{E})$ of degree $Q+1$. We can also associate a moment Green function $\hat{\mathsf{M}}_i^{[n]} = \hat{\mathsf{M}}_i^{[n]}(\xi\Delta x)$ for the $i \in [\![1, q]\!]$ moment, defined by $\hat{\mathsf{M}}_i^{[n]} = \boldsymbol{e}_1^{\mathsf{t}} \hat{\mathbf{E}}^n \boldsymbol{e}_i$. The numerical solution can then be written—for $n \geq Q+1$—using the following decompositions:

$$\hat{\mathsf{m}}_1(n\Delta t, \xi\Delta x) = \overbrace{\hat{\mathsf{g}}^{[n]}(\xi\Delta x)\hat{\mathsf{m}}_1^\circ(\xi\Delta x)}^{\text{(I) starting schemes}} = \overbrace{\sum_{\ell=1}^{Q+1} \hat{\mathsf{W}}_\ell(\xi\Delta x)\hat{\mathsf{g}}_\ell(\xi\Delta x)^n}^{\text{(II) spectrum of } \mathbf{E}} = \overbrace{\sum_{\ell=0}^{Q} \hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)\hat{\mathsf{g}}^{[\ell]}(\xi\Delta x)}^{\text{(III) time Green functions}}$$

$$= \underbrace{\left(\hat{\mathsf{M}}_1^{[n]}(\xi\Delta x)\hat{\mathsf{w}}_1(\xi\Delta x) + \sum_{\substack{i=2 \\ s_i \neq 1}}^{q} \hat{\mathsf{M}}_i^{[n]}(\xi\Delta x)\hat{\mathsf{w}}_i(\xi\Delta x)\right)\hat{\mathsf{m}}_1^\circ(\xi\Delta x)}_{\text{(IV) moment Green functions}},$$

where the decomposition (II) holds under the assumption that the roots of $\mathsf{z}^{Q+1-q}\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}})$ are simple for any wavenumber. The decomposition (II) makes it clear that the consistency for the initial conditions also depends on the non-physical eigenvalues, and the role of the choice of initial datum influences the weights $\hat{\mathsf{W}}_\ell$ for $\ell \in [\![1, Q+1]\!]$ in a non-trivial way. For the same reasons, it is difficult to analyse (III) and (IV), for the involved Green functions can be difficult to compute, even in the small wavenumber limit, due to the influence of all the modes of the system. Finally, let us insist on the fact that the Cayley-Hamilton theorem entails that the amplification factors of the starting schemes which are not initialisation schemes can be computed in two ways, which read

$$\hat{\mathsf{g}}^{[n]} = \boldsymbol{e}_1^{\mathsf{t}} \hat{\mathbf{E}}^n \hat{\mathbf{w}} = [\hat{\mathsf{M}}_1^{[n]}, \dots, \hat{\mathsf{M}}_q^{[n]}]\hat{\mathbf{w}} = \boldsymbol{e}_1^{\mathsf{t}} \hat{\mathbf{Q}}^{n-Q} [\hat{\mathsf{g}}^{[Q]}, \dots, \hat{\mathsf{g}}^{[1]}, \hat{\mathsf{w}}_1]^{\mathsf{t}},$$

for $n \geq Q+1$.

A second result states that there is little interest in considering the formal limit $n \to +\infty$ in the modified equations of the starting schemes.

**Proposition 4** (Long-time behavior: limits for $n \to +\infty$)**.** *Assume that the scheme is $L^2$ stable, meaning that the roots of the amplification polynomial $\mathsf{z}^{Q+1-q}\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}})$ are inside or on the unit circle and those on the unit circle are simple. Also assume that $\omega_1^{(0)} = 1$. Then:*

- *If $|1 - s_i| < 1$ for $i \in [\![2, q]\!]$, or equivalently $s_i \in ]0, 2[$ for $i \in [\![2, q]\!]$, then the modified equations of the starting schemes in the formal long-time limit $n \to +\infty$ coincide at any order with the one of the bulk Finite Difference scheme.*
- *If it exists $\tilde{i} \in [\![2, q]\!]$ such that $|1 - s_{\tilde{i}}| = 1$, thus equivalently $s_{\tilde{i}} = 2$. Let $H \in \mathbb{N}$. Provided that $\omega_1^{(h)} = 0$ for $h \in [\![1, H]\!]$ and the $Q$ modified equations of the initialisation schemes coincide with the one of the bulk Finite Difference scheme at any order $h \in [\![1, H]\!]$, then the modified equations of the starting schemes in the formal long-time limit $n \to +\infty$ coincide at any order $h \in [\![1, H+1]\!]$ with the one of the bulk Finite Difference scheme.*

The first case in Proposition 4 ensures that all the parasitic modes are damped. The second one deals with the case of undamped parasitic modes (*i.e.* leap–frog-like schemes). Proposition 4 means that the effect of the initialisation decays, provided that the initial filtering on the datum of the Cauchy problem (2) preserves it at leading order and that either the parasitic modes damp in time, or if the parasitic modes are oscillatory, the initialisation schemes are accurate enough. The situation is the one depicted in Figure 2, where the dots asymptotically reach the dashed lines. Let us point out that the assumption concerning stability may not be optimal, in the sense that we can find unstable scheme (for example, violating the Courant-Friedrichs-Lewy condition, but not having relaxation parameters exceeding 2) for which the modified equations of the starting schemes asymptotically reach those of the bulk Finite Difference scheme. However, this schemes are practically useless.

*Example* 3. In order to illustrate Proposition 4, consider Example 1. For this scheme, we consider $\mathsf{w}_1 = 1$ and $\mathsf{w}_2 = 0$. This choice does not satisfy Corollary 1 and the resulting initialization scheme is not consistent with the target equation, thus $H = 0$. According to the choice of $s_2$, we obtain the following modified equations up to order two:

| $n$ | Mod. eq. starting schemes for $s_2 = 3/2$ | Mod. eq. starting schemes for $s_2 = 2$ |
|---|---|---|
| 2 | $\partial_t\phi = -\lambda\frac{9}{8}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{9}{64}\epsilon_2^2 - \frac{1}{4}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 3 | $\partial_t\phi = -\lambda\frac{9}{8}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{51}{128}\epsilon_2^2 - \frac{1}{4}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{4}{3}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{4}{3}\epsilon_2^2 - \frac{1}{6}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 4 | $\partial_t\phi = -\lambda\frac{69}{64}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{441}{2048}\epsilon_2^2 - \frac{7}{32}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 5 | $\partial_t\phi = -\lambda\frac{171}{160}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{549}{2048}\epsilon_2^2 - \frac{17}{80}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{6}{5}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{6}{5}\epsilon_2^2 - \frac{1}{10}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 6 | $\partial_t\phi = -\lambda\frac{135}{128}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{3603}{16384}\epsilon_2^2 - \frac{13}{64}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 7 | $\partial_t\phi = -\lambda\frac{939}{896}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{52281}{229376}\epsilon_2^2 - \frac{89}{448}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{8}{7}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{8}{7}\epsilon_2^2 - \frac{1}{14}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 8 | $\partial_t\phi = -\lambda\frac{2133}{2048}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{222777}{1048576}\epsilon_2^2 - \frac{199}{1024}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 9 | $\partial_t\phi = -\lambda\frac{531}{512}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{110769}{524288}\epsilon_2^2 - \frac{49}{256}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{10}{9}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{10}{9}\epsilon_2^2 - \frac{1}{18}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 10 | $\partial_t\phi = -\lambda\frac{10581}{10240}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{4296249}{20971520}\epsilon_2^2 - \frac{967}{5120}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 11 | $\partial_t\phi = -\lambda\frac{23211}{22528}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{18673209}{92274688}\epsilon_2^2 - \frac{2105}{11264}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{12}{11}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{12}{11}\epsilon_2^2 - \frac{1}{22}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 12 | $\partial_t\phi = -\lambda\frac{16839}{16384}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{26696211}{134217728}\epsilon_2^2 - \frac{1517}{8192}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 13 | $\partial_t\phi = -\lambda\frac{109227}{106496}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{343150137}{1744830464}\epsilon_2^2 - \frac{9785}{53248}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{14}{13}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{14}{13}\epsilon_2^2 - \frac{1}{26}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 14 | $\partial_t\phi = -\lambda\frac{234837}{229376}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{1461161529}{7516192768}\epsilon_2^2 - \frac{20935}{114688}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 15 | $\partial_t\phi = -\lambda\frac{167481}{163840}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{2068175379}{10737418240}\epsilon_2^2 - \frac{14867}{81920}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{16}{15}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{16}{15}\epsilon_2^2 - \frac{1}{30}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 16 | $\partial_t\phi = -\lambda\frac{1070421}{1048576}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{26245566009}{137438953472}\epsilon_2^2 - \frac{94663}{524288}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 17 | $\partial_t\phi = -\lambda\frac{2271915}{2228224}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{110717799993}{584115552256}\epsilon_2^2 - \frac{200249}{1114112}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{18}{17}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{18}{17}\epsilon_2^2 - \frac{1}{34}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| 18 | $\partial_t\phi = -\lambda\frac{533997}{524288}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{51750979761}{274877906944}\epsilon_2^2 - \frac{46927}{262144}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + \lambda\Delta x\epsilon_2^2\partial_{xx}\phi + O(\Delta x^2)$ |
| 19 | $\partial_t\phi = -\lambda\frac{10136235}{9961472}\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{1954701086265}{10445360463872}\epsilon_2^2 - \frac{888377}{4980736}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\frac{20}{19}\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{20}{19}\epsilon_2^2 - \frac{1}{38}\big)\partial_{xx}\phi + O(\Delta x^2)$ |
| Mod. eq. bulk | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi - \lambda\Delta x\big(\frac{1}{6}\epsilon_2^2 - \frac{1}{6}\big)\partial_{xx}\phi + O(\Delta x^2)$ | $\partial_t\phi = -\lambda\epsilon_2\partial_x\phi + O(\Delta x^2)$ |

For $s_2 = 3/2$, we observe convergence at any order, whereas for $s_2 = 2$, we see that convergence takes place until order $H+1 = 1$ (transport) and $H+2 = 2$ (diffusion) is not converging, as claimed by Proposition 4.

*Proof of Proposition 4.* Let us formulate a preliminary remark: we consider formal series in the limit $|\xi\Delta x| \ll 1$. Therefore, the possibility of diagonalise $\hat{\mathbf{Q}}(\xi\Delta x)$ in this limit—or alternatively being obliged to deal with a true Jordan canonical form—is determined by $\hat{\mathbf{Q}}(0)$, *i.e.* the leading-order term in the formal series. We assume—without loss of generality—that all $\hat{\mathsf{g}}_\ell(0)$ for $\ell \in [\![1, Q+1]\!]$ are simple, even those strictly inside the unit circle, so that we can diagonalise the companion matrix in the desired limit. If this does not hold, for example for a SRT (or BGK) scheme, one can easily go through the same proof using the well-known expression for the powers of Jordan blocks.

Let us start the proof. Let $\hat{\mathbf{V}} = \hat{\mathbf{V}}(\xi\Delta x)$ be the Vandermonde matrix associated with $\hat{\mathsf{g}}_1 = \hat{\mathsf{g}}_1(\xi\Delta x), \dots, \hat{\mathsf{g}}_{Q+1} = \hat{\mathsf{g}}_{Q+1}(\xi\Delta x)$. It is well-known that this Vandermonde matrix diagonalises the companion matrix $\hat{\mathbf{Q}}(\xi\Delta x)$, thus for $n \geq Q+1$

$$\begin{aligned}\hat{\mathsf{g}}^{[n]}(\xi\Delta x) &= \boldsymbol{e}_1^{\mathsf{t}}\hat{\mathbf{Q}}(\xi\Delta x)^{n-Q}[\hat{\mathsf{g}}^{[Q]}(\xi\Delta x), \dots, \hat{\mathsf{g}}^{[1]}(\xi\Delta x), \hat{\mathsf{w}}_1(\xi\Delta x)]^{\mathsf{t}} \\ &= \boldsymbol{e}_1^{\mathsf{t}}\hat{\mathbf{V}}(\xi\Delta x)\mathrm{diag}(\hat{\mathsf{g}}_1(\xi\Delta x)^{n-Q}, \dots, \hat{\mathsf{g}}_{Q+1}(\xi\Delta x)^{n-Q})\hat{\mathbf{V}}(\xi\Delta x)^{-1}[\hat{\mathsf{g}}^{[Q]}(\xi\Delta x), \dots, \hat{\mathsf{g}}^{[1]}(\xi\Delta x), \hat{\mathsf{w}}_1(\xi\Delta x)]^{\mathsf{t}}.\end{aligned} \tag{33}$$

The idea of the proof is that the amplification factors associated with the initialisation schemes form an approximation of the eigenvector of $\hat{\mathbf{Q}}(\xi\Delta x)$ relative to the consistency eigenvalue $\hat{\mathsf{g}}_1$, so that the power iteration (33) converges for $n \to +\infty$ up to some order. Up to a re-ordering of the non-conserved moments—in order to start with those which do not relax on the equilibrium, for notational ease—the lower-triangular structure of the collision matrix $\boldsymbol{K}$ entails that $\hat{\mathsf{g}}_\ell(0) = 1 - s_\ell$ for $\ell \in [\![2, Q+1]\!]$. Moreover, we have that $\hat{\mathsf{g}}_1(0) = 1$.

- Using the assumption on the relaxation parameters, we have $|\hat{\mathsf{g}}_\ell(0)| < 1$ for $\ell \in [\![2, Q+1]\!]$. Using the assumption $\omega_1^{(0)} = 1$ (*i.e.* $\hat{\mathsf{w}}_1(\xi\Delta x) = 1 + O(|\xi\Delta x|)$), Proposition 2 provides $\hat{\mathsf{g}}^{[\ell]}(\xi\Delta x) = 1 + O(|\xi\Delta x|)$ for $\ell \in [\![1, Q]\!]$. Therefore

$$[\hat{\mathsf{g}}^{[Q]}(\xi\Delta x), \dots, \hat{\mathsf{g}}^{[1]}(\xi\Delta x), \hat{\mathsf{w}}_1(\xi\Delta x)] = [\hat{\mathsf{g}}_1(\xi\Delta x)^Q, \dots, \hat{\mathsf{g}}_1(\xi\Delta x), 1] + O(|\xi\Delta x|),$$

  which means that the amplification factors of the initialisation schemes are the eigenvector of $\hat{\mathbf{Q}}(\xi\Delta x)$ associated with $\hat{\mathsf{g}}_1(\xi\Delta x)$ at leading order. Back in (33), this gives that

$$\begin{aligned}&\hat{\mathsf{g}}^{[n]}(\xi\Delta x) = \boldsymbol{e}_1^{\mathsf{t}}\hat{\mathbf{V}}(\xi\Delta x)\mathrm{diag}(\hat{\mathsf{g}}_1(\xi\Delta x)^{n-Q}, \dots, \hat{\mathsf{g}}_{Q+1}(\xi\Delta x)^{n-Q})(\boldsymbol{e}_1 + O(|\xi\Delta x|)) \\ &= \boldsymbol{e}_1^{\mathsf{t}}\begin{bmatrix}\hat{\mathsf{g}}_1(\xi\Delta x)^Q & \cdots & \hat{\mathsf{g}}_{Q+1}(\xi\Delta x)^Q \\ \vdots & & \vdots \\ \hat{\mathsf{g}}_1(\xi\Delta x) & \cdots & \hat{\mathsf{g}}_{Q+1}(\xi\Delta x) \\ 1 & \cdots & 1\end{bmatrix}\mathrm{diag}(\hat{\mathsf{g}}_1(\xi\Delta x)^{n-Q}(1+O(|\xi\Delta x|)), \hat{\mathsf{g}}_2(\xi\Delta x)^{n-Q}O(|\xi\Delta x|), \dots, \hat{\mathsf{g}}_{Q+1}(\xi\Delta x)^{n-Q}O(|\xi\Delta x|))\end{aligned}$$

$$= \hat{\mathrm{g}}_1(\xi\Delta x)^n(1+O(|\xi\Delta x|)) + \hat{\mathrm{g}}_2(\xi\Delta x)^n O(|\xi\Delta x|) + \cdots + \hat{\mathrm{g}}_{Q+1}(\xi\Delta x)^n O(|\xi\Delta x|)),$$

where it is important to observe that the $O(|\xi\Delta x|)$-terms are independent of $n$. Considering that $|\hat{\mathrm{g}}_\ell(0)| < 1$ for $\ell \in [\![2, Q+1]\!]$, we deduce that $\lim_{n\to+\infty} \hat{\mathrm{g}}_\ell(\xi\Delta x)^n = 0$ for $\ell \in [\![2, Q+1]\!]$. Of course, convergence can be slow for high orders in the formal series. This entails that we have

$$\hat{\mathrm{g}}^{[n]}(\xi\Delta x) = \hat{\mathrm{g}}_1(\xi\Delta x)^n(1+\hat{\mathrm{r}}^{[n]}(\xi\Delta x)),$$

where the residual $\hat{\mathrm{r}}^{[n]}(\xi\Delta x) = O(|\xi\Delta x|)$ is such that it converges to a fixed formal series for $n \to +\infty$. The usual trick provides

$$\lim_{n\to+\infty} \partial_t = \frac{\lambda}{\Delta x} \lim_{n\to+\infty} \frac{1}{n} \log(\hat{\mathrm{g}}^{[n]}(\xi\Delta x)) = \frac{\lambda}{\Delta x}\Big(\log(\hat{\mathrm{g}}_1(\xi\Delta x)) + \lim_{n\to+\infty} \frac{1}{n} \log(1+\hat{\mathrm{r}}^{[n]}(\xi\Delta x)\Big) = \frac{\lambda}{\Delta x} \log(\hat{\mathrm{g}}_1(\xi\Delta x)).$$

- Observe that thanks to the stability assumption, there can be only one relaxation parameter $s_{\tilde{\imath}} = 2$. Otherwise, there would be a multiple eigenvalue on the unit circle for $\xi\Delta x = 0$, contradicting the stability assumption whilst generating linear instabilities. Up to a rearrangement of the moments, we have $\tilde{\imath} = 2$. By the assumptions on $\mathrm{w}_1$ and the initialisation schemes, we deduce that

$$[\hat{\mathrm{g}}^{[Q]}(\xi\Delta x), \dots, \hat{\mathrm{g}}^{[1]}(\xi\Delta x), \hat{\mathrm{w}}_1(\xi\Delta x)] = [\hat{\mathrm{g}}_1(\xi\Delta x)^Q, \dots, \hat{\mathrm{g}}_1(\xi\Delta x), 1] + O(|\xi\Delta x|^{H+1}),$$

which means that the amplification factors of the initialisation schemes are the eigenvector of $\hat{\mathbf{Q}}(\xi\Delta x)$ associated with $\hat{\mathrm{g}}_1(\xi\Delta x)$ up to order $O(|\xi\Delta x|^{H+1})$. Into (33), this yields

$$\begin{aligned}\hat{\mathrm{g}}^{[n]}(\xi\Delta x) &= \boldsymbol{e}_1^{\mathsf{t}} \hat{\mathbf{V}}(\xi\Delta x) \mathrm{diag}(\hat{\mathrm{g}}_1(\xi\Delta x)^{n-Q}, \hat{\mathrm{g}}_2(\xi\Delta x)^{n-Q}, \dots, \hat{\mathrm{g}}_{Q+1}(\xi\Delta x)^{n-Q})(\boldsymbol{e}_1 + O(|\xi\Delta x|^{H+1})) \\ &= \hat{\mathrm{g}}_1(\xi\Delta x)^n(1+O(|\xi\Delta x|^{H+1})) + \hat{\mathrm{g}}_2(\xi\Delta x)^n O(|\xi\Delta x|^{H+1}) + \cdots + \hat{\mathrm{g}}_{Q+1}(\xi\Delta x)^n O(|\xi\Delta x|^{H+1})),\end{aligned}$$

where all the $O(|\xi\Delta x|^{H+1})$-terms are independent of $n$. Due to the fact that $\hat{\mathrm{g}}_2(0) = 1 - s_2 = -1$, the formal series $\hat{\mathrm{g}}_2(\xi\Delta x)^n$ contains terms that can oscillate by featuring expressions involving $(-1)^n$, and the term at order $h \in [\![0, +\infty[\![$ grows with $n$ at most as a polynomial of degree $h$ in $n$. We indicate this fact using the notation $\hat{\mathrm{g}}_2(\xi\Delta x)^n = \sum_{h=0}^{h=+\infty} O(n^h)(\xi\Delta x)^h$. Therefore

$$\hat{\mathrm{g}}_2(\xi\Delta x)^n O(|\xi\Delta x|^{H+1}) = \sum_{h=H+1}^{+\infty} O(n^{h-H-1})(\xi\Delta x)^h.$$

As previously acknowledged, since $|\hat{\mathrm{g}}_\ell(0)| < 1$ for $\ell \in [\![3, Q+1]\!]$, we deduce that $\lim_{n\to+\infty} \hat{\mathrm{g}}_\ell(\xi\Delta x)^n = 0$ for $\ell \in [\![3, Q+1]\!]$. This ensures that

$$\hat{\mathrm{g}}^{[n]}(\xi\Delta x) = \hat{\mathrm{g}}_1(\xi\Delta x)^n\Big(1 + \sum_{h=H+1}^{+\infty} O(n^{h-H-1})(\xi\Delta x)^h\Big).$$

Utilising the usual trick, we have

$$\begin{aligned}\lim_{n\to+\infty} \partial_t &= \frac{\lambda}{\Delta x} \lim_{n\to+\infty} \frac{1}{n} \log(\hat{\mathrm{g}}^{[n]}(\xi\Delta x)) = \frac{\lambda}{\Delta x}\Big(\log(\hat{\mathrm{g}}_1(\xi\Delta x)) + \lim_{n\to+\infty} \frac{1}{n} \log\Big(1 + \sum_{h=H+1}^{+\infty} O(n^{h-H-1})(\xi\Delta x)^h\Big)\Big) \\ &= \frac{\lambda}{\Delta x}\Big(\log(\hat{\mathrm{g}}_1(\xi\Delta x)) + \lim_{n\to+\infty} \frac{1}{n} \sum_{h=H+1}^{+\infty} O(n^{h-H-1})(\xi\Delta x)^h\Big) = \frac{\lambda}{\Delta x}\Big(\log(\hat{\mathrm{g}}_1(\xi\Delta x)) + \lim_{n\to+\infty} \sum_{h=H+1}^{+\infty} O(n^{h-H-2})(\xi\Delta x)^h\Big) \\ &= \frac{\lambda}{\Delta x}\Big(\log(\hat{\mathrm{g}}_1(\xi\Delta x)) + \lim_{n\to+\infty}\Big(O(n^{-1})(\xi\Delta x)^{H+1} + \sum_{h=H+2}^{+\infty} O(n^{h-H-2})(\xi\Delta x)^h\Big)\Big) \\ &= \frac{\lambda}{\Delta x}\Big(\log(\hat{\mathrm{g}}_1(\xi\Delta x)) + \lim_{n\to+\infty} \sum_{h=H+2}^{+\infty} O(n^{h-H-2})(\xi\Delta x)^h\Big),\end{aligned}$$

achieving the demonstration.

□

## 4.8 Conclusions

In this Section 4, we have proposed a way of linking the initial datum of the lattice Boltzmann scheme $\mathbf{m}(0,\cdot)$ to the initial datum $u^\circ$ of the Cauchy problem (2). This allowed us to propose a modified equation analysis of the initialisation phase— see Proposition 1 and Proposition 2— making the study of the real behaviour of the numerical schemes possible and find the

constraints—see Corollary 1 and Corollary 2— under which the initialisation schemes are consistent with the same equation (1) as the bulk scheme, preventing from having order reductions. We have also stressed that controlling the behaviour of the scheme inside the initialisation layer implies a control on the numerical scheme eventually in time (Proposition 3). The general computations have been done until order $O(\Delta x)$ but can be carried further to $O(\Delta x^2)$ and above for specific schemes, as in Section 5 and Section 6. This provides additional information on other features of the schemes close to the beginning of the simulation, such as dissipation and dispersion.

# 5 Examples and numerical simulations

Section 5 first aims at checking the previously introduced theory concerning consistency on actual numerical simulations on the $D_1Q_2$ (*cf.* Example 1). Moreover, the computations of the modified equation shall be pushed one order further providing the dissipation of the starting schemes, the impact of which is precisely quantified on the numerical experiments for a $D_1Q_2$ and $D_1Q_3$ scheme. Finally, the example of $D_1Q_3$ scheme paves the way for the general discussion of Section 6 concerning a more precise counting of the number of initialisation schemes, with important consequences on the dissipation of the numerical schemes.

## 5.1 Two-velocities $D_1Q_2$ scheme

Consider the scheme from Example 1. The modified equation of the bulk Finite Difference scheme reads as in [Graille, 2014] and Theorem 1

$$\partial_t\phi(t,x)+\lambda\epsilon_2\partial_x\phi(t,x)-\lambda\Delta x\Big(\frac{1}{s_2}-\frac{1}{2}\Big)(1-\epsilon_2^2)\partial_{xx}\phi(t,x)=O(\Delta x^2),\qquad (t,x)\in\mathbb{R}_+\times\mathbb{R},\tag{34}$$

thus to be consistent with the Cauchy problem (1), one takes $\epsilon_2=V/\lambda$. For $s_2<2$, the bulk Finite Difference scheme is first-order accurate, thus initialisation schemes which are non-consistent with the target conservation law—*i.e.* indeed violating (17) or (24)—do not degrade the order of convergence. For $s_2=2$, the bulk Finite Difference scheme is second-order accurate, thus consistent initialisation schemes are needed, *i.e.* verifying (17) or (24). Observe that the scheme is $L^2$ stable according to *von Neumann* (*i.e.* the roots of the amplification polynomial of the bulk Finite Difference scheme are inside the unit disk and those on the unit disk are simple) under the conditions ([Graille, 2014] and Appendix A)

$$s_2\in]0,2],\qquad\text{and}\qquad\begin{cases}|\epsilon_2|\le 1, & \text{if}\quad s_2\in]0,2[,\\ |\epsilon_2|<1, & \text{if}\quad s_2=2.\end{cases}\tag{35}$$

The second conditions are the Courant-Friedrichs-Lewy (CFL) condition, which is known to be strict for the leap-frog scheme.

We consider five different choices of initialisation schemes. They are designed to showcase different facets of the previous theoretical discussion. More precisely, the first initialisation is the one where all data are taken at equilibrium, which is likely the most common way of initializing lattice Boltzmann schemes [Graille, 2014, Caetano et al., 2019]. The second and the third initialisations both render a forward centered scheme as initialisation scheme, which would be unstable if used as bulk scheme. Still, these two initialisations yield different outcomes for the associated numerical simulations and our theory accounts for this phenomenon. The fourth initialisation aims at obtaining a Lax-Wendroff initialisation scheme, which allows to study the effect of a second-order initialisation scheme. Finally, the fifth initialisation is inspired by works from the literature [Van Leemput et al., 2009].

- **Lax-Friedrichs scheme** (LF), a first-order consistent scheme which we shall obtain using the local initialisation

$$\mathsf{w}_1=1,\qquad \mathsf{w}_2=\epsilon_2.\tag{36}$$

  Except when $s_2=1$ (where $Q=0$), the dissipation of the bulk Finite Difference scheme is not matched by the one of the Lax-Friedrichs scheme.

- **Forward centered scheme** (FC). This is a first-order consistent scheme which is unstable even under CFL condition (35) if used as bulk scheme, due to its negative dissipation. Still, it is perfectly suitable for the initialisation of the method (see [Strikwerda, 2004, Chapter 10]). Its diffusivity shall not match the one of the bulk Finite Difference scheme, see (34). This initialisation scheme cannot stem from a local initialisation, *i.e.* $\mathsf{w}_1,\mathsf{w}_2\in\mathbb{R}$, since the only first-order consistent initialisation scheme that can be obtained in this way is the Lax-Friedrichs scheme (36). We could unsuccessfully try to generate it by a local initialisation of the conserved moment, that is $\mathsf{w}_1=1$ and prepared initialisation of the non-conserved one, thus $\mathsf{w}_2\in\mathbb{R}[\mathsf{x}_1,\mathsf{x}_1^{-1},\dots,\mathsf{x}_d,\mathsf{x}_d^{-1}]$. Considering—see Appendix B for the

Table 1: Expected orders of convergence in $\Delta x$ for the $D_1Q_2$ scheme.

| Test | Bulk scheme 1st order ($0 < s_2 < 2$) | Bulk scheme 2nd order ($s_2 = 2$) |
|---|---|---|
| (a) - (41) | order 1/4 | order 1/3 |
| (b) - (42) | order 3/4 | order 1 |
| (c) - (43) | order 1 | order 5/3 |
| (d) - (44) | order 1 | order 2 |

details—a prepared initialisation for both moments, thus $\mathsf{w}_1, \mathsf{w}_2 \in \mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$, several choices are possible to recover this scheme. One is

$$w_{1,\pm1} = \frac{1}{2}, \qquad w_{2,\pm1} = \mp\frac{1 \pm s_2\epsilon_2}{2(1-s_2)}, \qquad w_{2,0} = \frac{\epsilon_2}{1-s_2}, \tag{37}$$

with the notation by (21) and agrees with (22), (23) and (24). Another possible choice to obtain the desired scheme would be

$$w_{1,\pm2} = \pm\frac{\epsilon_2}{2}, \qquad w_{1,\pm1} = \frac{1}{2}, \qquad w_{2,\pm2} = -\frac{\epsilon_2(1 \pm s_2\epsilon_2)}{2(1-s_2)}, \qquad w_{2,\pm1} = \mp\frac{1 \pm s_2\epsilon_2}{2(1-s_2)}. \tag{38}$$

However, this initialisation yields only (22) but does not fulfill either (23) or (24). This means that in this case $\mathsf{m}_1$ is initialized as a first-order perturbation of the datum of the Cauchy problem (2) and that $\mathsf{m}_2$ is not initialized at equilibrium at leading order.

- **Lax-Wendroff scheme** (LW). This is a second-order consistent scheme with no dissipation, thus matches the diffusivity of the bulk Finite Difference scheme when $s_2 = 2$. Remark that since the bulk scheme is at most second-order accurate, it is somehow excessive to initialize with a scheme of the same order. Following an analogous procedure to the centered forward scheme, one possible initialisation is $\mathsf{w}_1, \mathsf{w}_2 \in \mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$ with

$$w_{1,\pm1} = \frac{1-\epsilon_2^2}{2}, \quad w_{1,0} = \epsilon_2^2, \quad w_{2,\pm1} = \mp\frac{(1 \pm s_2\epsilon_2)(1-\epsilon_2^2)}{2(1-s_2)}, \quad w_{2,0} = \frac{\epsilon_2(1-s_2\epsilon_2^2)}{1-s_2}, \tag{39}$$

according to (21), which respects (22), (23) and (24). Again, it is also possible to generate initialisations yielding this scheme which do not fulfill (23) and (24).

- **Smooth initialisation inspired by** [Van Leemput et al., 2009] (RE1). The idea of this initialisation is to make the most of the terms in the modified equation of the initialisation schemes and that of the bulk Finite Difference scheme to match, if possible, without modification the conserved moment, that is $\mathsf{w}_1 \in \mathbb{R}$. In particular, in our case, this allows to match the numerical diffusion coefficient between the two schemes for every $s_2 \in ]0, 2]$, as we shall see. We adapt Equation (13) from [Van Leemput et al., 2009] by discretising the continuous derivative by a second-order centered formula, having

$$\mathsf{w}_1 = 1 \quad \text{and} \quad \mathsf{w}_2 \in \mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}], \quad \text{where} \quad w_{2,\pm1} = \pm\frac{1-\epsilon_2^2}{2s_2}, \qquad w_{2,0} = \epsilon_2, \tag{40}$$

according to (21). This initialisation fulfills (22), (23) and (24).

### 5.1.1 Study of the convergence order

To empirically analyze the preservation of the order of the bulk Finite Difference scheme, we consider the following initial data with different smoothness

$$\text{(a)} \qquad u^\circ(x) = \chi_{|x|\le 1/2}(x) \in H^\sigma, \quad \text{for any} \quad \sigma < \sigma_0 = 1/2. \tag{41}$$

$$\text{(b)} \qquad u^\circ(x) = (1-2|x|)\chi_{|x|\le 1/2}(x) \in H^\sigma, \quad \text{for any} \quad \sigma < \sigma_0 = 3/2. \tag{42}$$

$$\text{(c)} \qquad u^\circ(x) = \cos^2(\pi x)\chi_{|x|\le 1/2}(x) \in H^\sigma, \quad \text{for any} \quad \sigma < \sigma_0 = 5/2. \tag{43}$$

$$\text{(d)} \qquad u^\circ(x) = \exp\left(-1/(1-|2x|^2)\right)\chi_{|x|\le 1/2}(x) \in C_c^\infty, \tag{44}$$

issued from [Bellotti et al., 2022]. As common in the linear framework, we monitor the $L^2$ errors. We simulate for $\lambda = 1$, $\epsilon_2 = V/\lambda = 1/2$ with final time $1/2$ and on a bounded domain $[-1, 1]$ with periodic boundary conditions.

We expect the scheme to be convergent following the orders given in Table 1 [Strikwerda, 2004, Bellotti et al., 2022] and observe orders exceeding one provided that both following conditions are met:

1. the initialisation scheme is at least first-order consistent with the Cauchy problem (1);
2. the initial filter on the initial datum $\mathsf{w}_1$ is such that $\omega_1^{(1)} = 0$, meaning that it perturbs from $O(\Delta x^2)$ or for higher orders.

The results are in agreement with the theory. We just present few of them for the sake of avoiding redundancy, in particular, those concerning the forward centered initialisation schemes (37) and (38) given in Figure 3. As expected, despite the fact that the obtained initialisation scheme is the same, (38) pollutes the initial datum with respect to the one from the Cauchy problem (2) due to a first-order term $\omega_1^{(1)} \neq 0$. Hence, even for $s_2 = 2$, the order of convergence is lowered. We shall reinterpret why (38) yields a poor behaviour.

### 5.1.2 Study of the time smoothness of the numerical solution

We have observed that the only proposed initialisation matching the dissipation of the bulk scheme for every $s_2 \in ]0,2]$ is the one given by (40). To confirm that this is the origin of its good performances in term of time smoothness of the discrete solution close to the initial time, we repeat the numerical experiment found in [Van Leemput et al., 2009]. The simulation is carried on the periodic domain $[0,1]$ discretized with $\Delta x = 1/30$, $s_2 = 1.99$, $\lambda = 1$ and $\epsilon_2 = V/\lambda = 0.66$. The initial datum of the Cauchy problem is $u^\circ(x) = \cos(2\pi x)$.

We initialize using the Lax-Friedrichs initialisation (36) (coinciding with what [Van Leemput et al., 2009] calls RE0 scheme), the forward centered initialisation (37), the Lax-Wendroff initialisation (39), the RE1 initialisation (40) and the implicit initialisations CR0 and CR1 proposed in [Van Leemput et al., 2009], which are not detailed here. We measure the difference between the exact solution and the approximate solution at the eighth cell of the lattice. The results are given in Figure 4 and are in accordance with the previous analysis as well as the computations in [Van Leemput et al., 2009]. Indeed, since the dissipation of the bulk Finite Difference scheme is almost zero for $s_2 = 1.99$, the Lax-Wendroff scheme is supposed to almost match this dissipation. However here, the same phenomenon that took place in Section 5.1.1 at leading order for the forward centered initialisation between (37) and (38), due to the introduction of a first-order perturbation on the conserved moment, now takes place for (39), because it introduces a second-order perturbation on the conserved moment $\mathsf{m}_1$ feeding the multi-step bulk Finite Difference scheme (10), namely $\omega_1^{(2)} \neq 0$. Taking $s_2 = 2$, hence no dissipation from the bulk scheme, we obtain the result in Figure 5, which is not different from the previous one (notice that here the implicit initialisation RE1 cannot be utilized). In this Figure, we have also repeated the simulation using a leap-frog scheme (coinciding with the bulk Finite Difference scheme) initialized with a Lax-Wendroff scheme, which conversely leads the expected smoothness, since we do not have to filter the initial datum of the Cauchy problem.

### 5.1.3 Theoretical analysis using the modified equations

Let us proceed to a more quantitative study of what can be observed in Figure 4 and Figure 5. To this end, we push the computation of the modified equation of the starting schemes for $n \in \mathbb{N}^*$ to order $O(\Delta x^2)$. To complete the previous computations, we are left to consider

$$\begin{aligned}(\boldsymbol{\mathcal{E}}^n)^{(2)} &= \sum\{\text{per. of } \boldsymbol{\mathcal{E}}^{(0)} \ (n-1 \text{ tm.}) \text{ and } \boldsymbol{\mathcal{E}}^{(2)} \text{ (once)}\} + \sum\{\text{per. of } \boldsymbol{\mathcal{E}}^{(0)} \ (n-2 \text{ tm.}) \text{ and } \boldsymbol{\mathcal{E}}^{(1)} \text{ (twice)}\} \\ &= \sum_{\ell=0}^{\ell=n-1} (\boldsymbol{\mathcal{E}}^{(0)})^\ell \boldsymbol{\mathcal{E}}^{(2)} (\boldsymbol{\mathcal{E}}^{(0)})^{n-1-\ell} + \sum_{\ell=0}^{\ell=n-2} \sum_{p=0}^{p=n-2-\ell} (\boldsymbol{\mathcal{E}}^{(0)})^\ell \boldsymbol{\mathcal{E}}^{(1)} (\boldsymbol{\mathcal{E}}^{(0)})^p \boldsymbol{\mathcal{E}}^{(1)} (\boldsymbol{\mathcal{E}}^{(0)})^{n-2-\ell-p}.\end{aligned}$$

Using the matrix from the particular $\mathrm{D_1Q_2}$ scheme, we obtain for every $n \in \mathbb{N}^*$

$$\begin{aligned}(\boldsymbol{\mathcal{E}}^n)_{11}^{(2)} = \Big(&\frac{n}{2} + \sum_{\ell=0}^{\ell=n-2} \sum_{p=1}^{p=n-1-\ell} (1-s_2)^p \\ &+ \epsilon_2^2 \sum_{\ell=0}^{\ell=n-2} \sum_{p=0}^{p=n-2-\ell} \Big( s_2^2 + s_2(1-s_2)\pi_{n-2-\ell-p}(s_2) + (1-s_2)\pi_p(s_2)\pi_{n-1-\ell-p}(s_2) \Big) \Big) \partial_{xx},\end{aligned}$$

and

$$(\boldsymbol{\mathcal{E}}^n)_{12}^{(2)} = \epsilon_2 \sum_{\ell=0}^{n-2} \sum_{p=0}^{n-2-\ell} (1-s_2)^{n-1-\ell-p} \pi_{p+1}(s_2) \partial_{xx}.$$

For all the initialisations we have considered, we have $\omega_1^{(0)} = 1$, corresponding to (22). No other assumption is needed in the following derivation. With the usual procedure, we obtain for $n \in \mathbb{N}^*$ and $x \in \mathbb{R}$

$$\begin{aligned}&\partial_t \phi(0,x) - \frac{\lambda}{n} \Big( (\boldsymbol{\mathcal{E}}^n)_{11}^{(1)} + (\boldsymbol{\mathcal{E}}^n)_{12}^{(1)} \omega_2^{(0)} + \omega_1^{(1)} \Big) \phi(0,x) \\ &+ n \frac{\Delta x}{2\lambda} \partial_{tt} \phi(0,x) - \frac{\lambda \Delta x}{n} \Big( (\boldsymbol{\mathcal{E}}^n)_{11}^{(2)} + (\boldsymbol{\mathcal{E}}^n)_{12}^{(2)} \omega_2^{(0)} + (\boldsymbol{\mathcal{E}}^n)_{11}^{(1)} \omega_1^{(1)} + (\boldsymbol{\mathcal{E}}^n)_{12}^{(1)} \omega_2^{(1)} + \omega_1^{(2)} \Big) \phi(0,x) = O(\Delta x^2).\end{aligned}$$

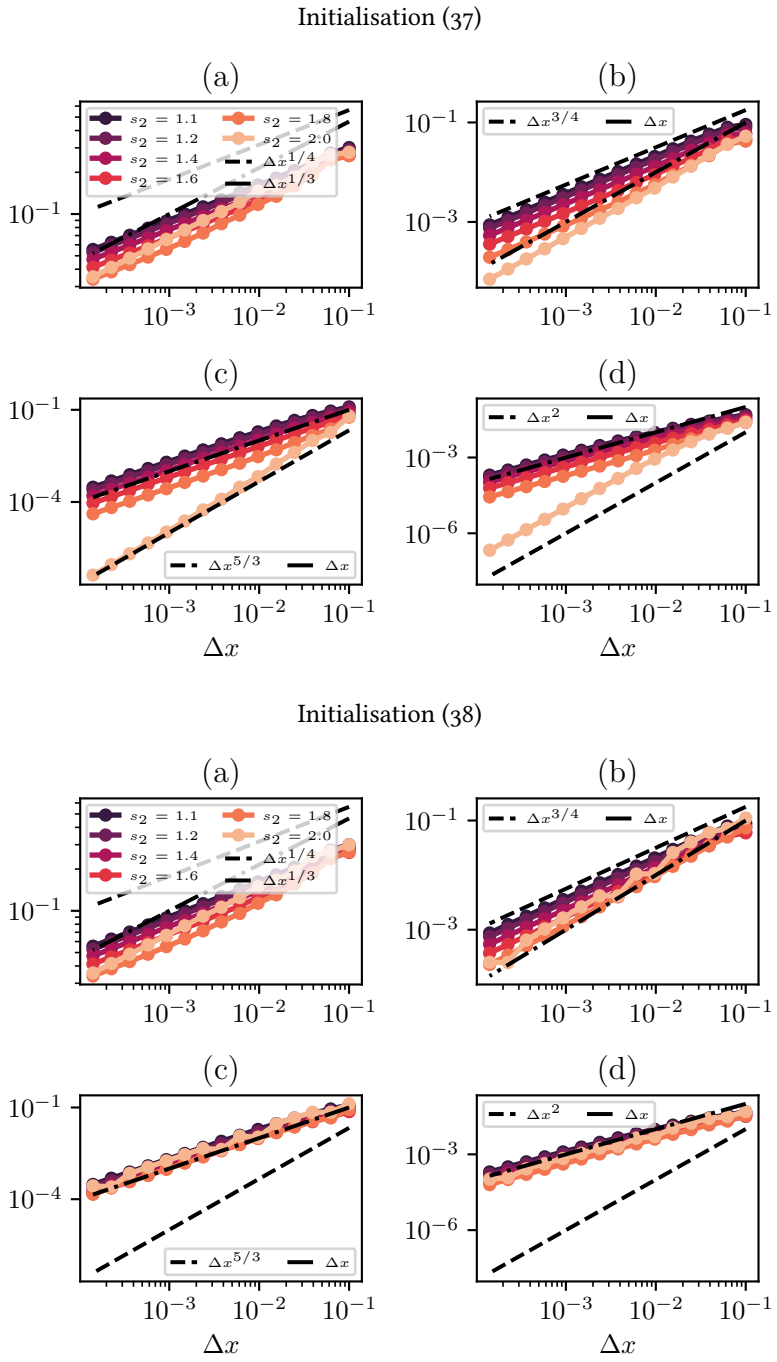


Figure 3: $L^2$ errors at the final time for two forward centered initialisations (37) (top) and (38) (bottom). Since the letter irremediably perturbs the conserved moment feeding the bulk Finite Difference scheme, the orders of convergence above one are lowered.

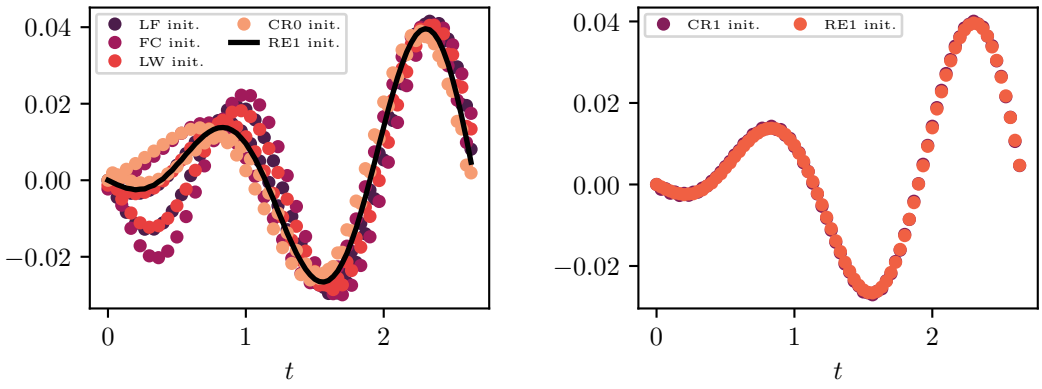


Figure 4: Test for the smoothness in time close to $t = 0$ for $s_2 = 1.99$: difference between exact and numerical solution at the eighth lattice point.

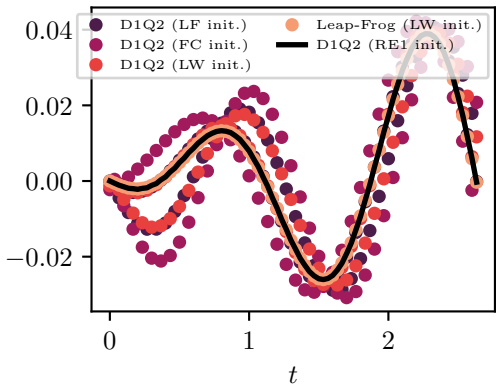


Figure 5: Test for the smoothness in time close to $t = 0$ for $s_2 = 2$: difference between exact and numerical solution at the eighth lattice point. Compared to Figure 4, CR0 and CR1 from [Van Leemput et al., 2009] cannot be used.

- **Lax-Friedrichs** (36).

  **Proposition 5.** *Under acoustic scaling, the modified equations for the starting schemes for the Lax-Friedrichs initialisation given by* (36) *are, for* $n \in \mathbb{N}^*$

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x\Big(\frac{1}{2} + \sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell\Big)(1-\epsilon_2^2)\partial_{xx}\phi(0,x) = O(\Delta x^2), \qquad x \in \mathbb{R}. \tag{45}$$

  *Proof.* This initialisation fulfils the requirements by Corollary 2, which leads to

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) + n\frac{\Delta x}{2\lambda}\partial_{tt}\phi(0,x) - \frac{\lambda\Delta x}{n}\Big((\boldsymbol{\mathcal{E}}^n)^{(2)}_{11} + (\boldsymbol{\mathcal{E}}^n)^{(2)}_{12}\epsilon_2\Big)\phi(0,x) = O(\Delta x^2), \tag{46}$$

  for $n \in \mathbb{N}^*$ and $x \in \mathbb{R}$. Using the previous order to get rid of the second-order time derivative $\partial_{tt}$ [Warming and Hyett, 1974, Carpentier et al., 1997, Dubois, 2008, Dubois, 2021] boils down to

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x\Big(-\frac{n}{2}\epsilon_2^2\partial_{xx} + \frac{1}{n}\Big((\boldsymbol{\mathcal{E}}^n)^{(2)}_{11} + (\boldsymbol{\mathcal{E}}^n)^{(2)}_{12}\epsilon_2\Big)\Big)\phi(0,x) = O(\Delta x^2),$$

  for $n \in \mathbb{N}^*$ and $x \in \mathbb{R}$. We are left to deal with the diffusion term, for $n \in \mathbb{N}^*$

$$(\boldsymbol{\mathcal{E}}^n)^{(2)}_{11} + (\boldsymbol{\mathcal{E}}^n)^{(2)}_{12}\epsilon_2 = \Big(\frac{n}{2} + \sum_{\ell=0}^{n-2}\sum_{p=1}^{n-1-\ell}(1-s_2)^p + \epsilon_2^2\sum_{\ell=0}^{n-2}\sum_{p=0}^{n-2-\ell}\Big(s_2^2 + s_2(1-s_2)\pi_{n-2-\ell-p}(s_2)$$

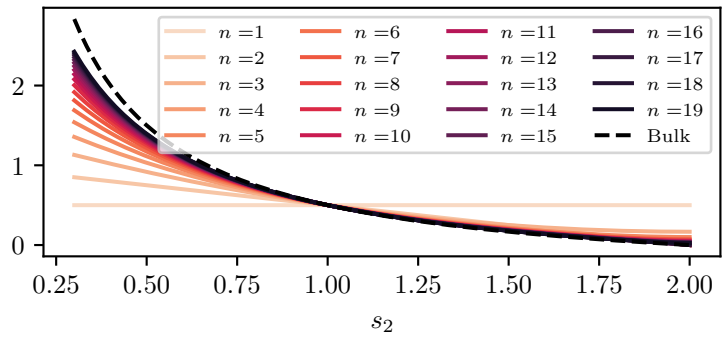


Figure 6: Plot of the polynomial $1/2+\sum_{\ell=1}^{\ell=n-1}(1-\ell/n)(1-s_2)^\ell$ appearing in (45) for different $n$ compared to $1/s_2-1/2$ (bulk).

$$+(1-s_2)\pi_p(s_2)\pi_{n-1-\ell-p}(s_2)+(1-s_2)^{n-1-\ell-p}\pi_{p+1}(s_2)\Big)\Big)\partial_{xx}.$$

Using the expression for $\pi_\ell$ to handle the last term shows that

$$(\boldsymbol{\mathcal{E}}^n)^{(2)}_{11}+(\boldsymbol{\mathcal{E}}^n)^{(2)}_{12}\epsilon_2=\Big(\frac{n}{2}+\sum_{\ell=1}^{n-1}(n-\ell)(1-s_2)^\ell+\epsilon_2^2\Big(\frac{n(n-1)}{2}-\sum_{\ell=1}^{n-1}(n-\ell)(1-s_2)^\ell\Big)\Big)\partial_{xx},$$

for $n\in\mathbb{N}^*$. Plugging into the expansion (46) provides

$$\partial_t\phi(0,x)+\lambda\epsilon_2\partial_x\phi(0,x)-\lambda\Delta x\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell\Big)(1-\epsilon_2^2)\partial_{xx}\phi(0,x)=O(\Delta x^2),\qquad n\in\mathbb{N}^*,$$

for $x\in\mathbb{R}$. □

This proves once more that the origin of the initial boundary layer is the mismatch in the dissipation coefficient of the scheme, see Figure 6. Of course, it must be kept in mind that these expansions are meaningful as long as $n\Delta t\ll 1$, this is, for small times. However, from the simulations and Figure 6, we see that the boundary layer damps in time, since the dissipation coefficient in (45) converges to the bulk one in (34) by taking the formal limit $n\to+\infty$:

$$\lim_{n\to+\infty}\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell\Big)=\lim_{n\to+\infty}\Big(\frac{1}{2}+\frac{(1-s_2)^{n+1}}{ns_2^2}-\frac{(1-s_2)}{ns_2^2}+\frac{(1-s_2)}{s_2}\Big)=\frac{1}{s_2}-\frac{1}{2},$$

unsurprisingly by virtue of Proposition 4. We see that—as previously claimed—this formal limit holds regardless of the fulfilment of the CFL condition. However, it strongly depends on the fact that $s_2\le 2$, otherwise, it would not hold and indeed exponentially diverge. We can also study the behaviour for $s_2\simeq 2$:

$$\lim_{s_2\to 2^-}\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell=\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(-1)^\ell=\frac{1-(-1)^n}{4n}=\begin{cases}0, & \text{for } n \text{ even},\\ \frac{1}{2n}, & \text{for } n \text{ odd},\end{cases}$$

for $n\in\mathbb{N}^*$. This explains why the errors in Figure 4 and Figure 5 are close to the ones of RE1 (40) (up to high order contributions) for even time steps. On the one hand, for $n$ even, the dissipation of the bulk Finite Difference scheme is matched by the starting schemes, producing good agreement. On the other hand, for $n$ odd, the dissipation is strictly positive, though decreasing linearly with $n$, creating the jumping behaviour of the errors. This suggests that the damping of the initial boundary layer should be proportional to $t^{-1}$ and explains the discrepancies with respect to RE1 (40) for the odd time steps. Finally, observe that this decoupling—even as far as the dissipation is concerned—between even and odd time steps for $s_2=2$ is expected since the bulk Finite Difference scheme is a leap-frog.

- **Forward centered scheme** (37). We have, see Appendix C for the proof:

**Proposition 6.** *Under acoustic scaling, the modified equations for the starting schemes for the forward centered initialisation given by* (37) *are, for* $n\in\mathbb{N}^*$

$$\partial_t\phi(0,x)+\lambda\epsilon_2\partial_x\phi(0,x) \tag{47}$$

$$-\lambda\Delta x\Big(\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell\Big)(1-\epsilon_2^2)+\frac{1}{2n}\Big(1-2\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)\partial_{xx}\phi(0,x)=O(\Delta x^2),$$

*with* $x\in\mathbb{R}$.

Again, according to Proposition 4, the bulk viscosity coefficient is asymptotically reached, since

$$\lim_{n\to+\infty}\Big(\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell\Big)(1-\epsilon_2^2)+\frac{1}{2n}\Big(1-2\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)=\Big(\frac{1}{s_2}-\frac{1}{2}\Big)(1-\epsilon_2^2).$$

Concerning the behaviour close to $s_2\simeq 2$, we have

$$\begin{aligned}&\lim_{s_2\to 2^-}\Big(\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell\Big)(1-\epsilon_2^2)+\frac{1}{2n}\Big(1-2\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)\\&=\frac{(1-(-1)^n)}{4n}(1-\epsilon_2^2)+\frac{(-1)^n}{2n}=\begin{cases}\frac{1}{2n}, & \text{for } n \text{ even},\\ -\frac{\epsilon_2^2}{2n}, & \text{for } n \text{ odd},\end{cases}\end{aligned}$$

for $n\in\mathbb{N}^*$. We observe that the even steps of starting schemes have the same diffusivity as the odd steps for the Lax-Friedrichs initialisation (36), whereas the odd ones have negative diffusivity, which remains from having an initialisation scheme with negative dissipation, coupled with the fact that the bulk Finite Difference scheme is a leap-frog scheme. The question which might be risen is on how the overall scheme can remain stable. In terms of Finite Differences, the choice of initial datum only changes the spectrum of the data feeding the bulk Finite Difference scheme, which is stable under (35), for every initial datum. Concerning the previous computation, we have that under the CFL condition $-\epsilon_2^2/(2n)\geq -1/(2n)$, hence steps with negative dissipation are compensated by steps with sufficiently positive dissipation, yielding an overall stable scheme.

- **Forward centered scheme** (38). For this scheme, it is useless to analyze until second order because we know that issues start at $O(\Delta x)$, see Section 5.1.1. We have, see Appendix C:

  **Proposition 7.** *Under acoustic scaling, the modified equations for the starting schemes for the forward centered initialisation given by* (38) *are, for* $n\in\mathbb{N}^*$

  $$\partial_t\phi(0,x)+\lambda\epsilon_2\Big(1+\frac{2}{n}\Big(1-\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)\partial_x\phi(0,x)=O(\Delta x),\qquad x\in\mathbb{R}.$$

  Unsurprisingly, the initialisation scheme is consistent ($n=1$), but the general starting schemes ($n>1$) are not. This does not prevent the overall scheme to converge, since $\omega_1^{(0)}=1$ but only at first-order even when $s_2=2$, see Figure 3. Following Proposition 4

  $$\lim_{n\to+\infty}\Big(1+\frac{2}{n}\Big(1-\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)=1.$$

- **Lax-Wendroff** (39). We have, *cf.* Appendix C:

  **Proposition 8.** *Under acoustic scaling, the modified equations for the starting schemes for the Lax-Wendroff initialisation given by* (39) *are, for* $n\in\mathbb{N}^*$

  $$\begin{aligned}&\partial_t\phi(0,x)+\lambda\epsilon_2\partial_x\phi(0,x)\\&\quad-\lambda\Delta x\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell+\frac{1}{2n}\Big(1-2\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)(1-\epsilon_2^2)\partial_{xx}\phi(0,x)=O(\Delta x^2),\end{aligned}\tag{48}$$

  *for* $x\in\mathbb{R}$.

  As expected, the dissipation coefficients tend to the one of the bulk scheme for $n\to+\infty$ and for $s_2\simeq 2$, we find

  $$\lim_{s_2\to 2^-}\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell+\frac{1}{2n}\Big(1-2\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)(1-\epsilon_2^2)=\frac{1+(-1)^n}{4n}=\begin{cases}\frac{1}{2n}, & \text{for } n \text{ even},\\ 0, & \text{for } n \text{ odd},\end{cases}$$

  for $n\in\mathbb{N}^*$. This is the opposite situation compared to the Lax-Friedrichs initialisation (36) and again justifies the jumping behaviour compared to RE1 (40), see Figure 4 and Figure 5. Moreover, we further understand why we still observe the boundary layer: even if the initialisation scheme matches the zero diffusivity of the bulk scheme, the second-order modification $\omega_1^{(2)}\neq 0$ we have imposed on the initial datum to obtain such initialisation scheme reverberates over the following (even) time steps.

- **Smooth initialisation RE1** (40).

  **Proposition 9.** *Under acoustic scaling, the modified equations for the starting schemes for the RE1 initialisation given by* (40) *are, for* $n \in \mathbb{N}^*$

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x\Big(\frac{1}{s_2} - \frac{1}{2}\Big)(1-\epsilon_2^2)\partial_{xx}\phi(0,x) = O(\Delta x^2), \qquad x \in \mathbb{R}.$$

  *Proof.* In Appendix C, we obtain that

$$\partial_t \phi(0,x) + \epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x\Big(\frac{1}{2} - \sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell + \frac{1}{ns_2}\sum_{\ell=1}^{n}(1-s_2)^\ell\Big)(1-\epsilon_2^2)\partial_{xx}\phi(0,x) = O(\Delta x^2). \tag{49}$$

  One can easily show by induction that

$$\frac{1}{2} - \sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell + \frac{1}{ns_2}\sum_{\ell=1}^{n}(1-s_2)^\ell = \frac{1}{s_2} - \frac{1}{2}, \qquad n \in \mathbb{N}^*,$$

  yielding the same modified equation as the bulk Finite Difference scheme. □

  This explains, once more, the smooth behaviour observed in Figure 4 and Figure 5 and also shows an actual application of Proposition 3 for $H = 2$.

## 5.2 Three-velocities $D_1Q_3$ scheme

The previous case of $D_1Q_2$ scheme suggests that particular care must be adopted when prepared initialisations for the conserved moment $\mathsf{m}_1$ are used (*i.e.* $\mathsf{w}_1 \in \mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$). Therefore, in what follows, we treat only local initialisations for any moment. We are now interested in equating the dissipation of the initialisation schemes with the one of the bulk scheme for a richer scheme: the $D_1Q_3$. In particular, we look for a full characterisation of the conditions under which $\mathsf{w}_1, \mathsf{w}_2, \mathsf{w}_3 \in \mathbb{R}$ yield initialisation schemes with the same dissipation as the bulk Finite Difference scheme.

### 5.2.1 Description of the scheme

We consider the $D_1Q_3$ scheme [Dubois et al., 2020, Bellotti et al., 2022], having $d = 1$, $q = 3$, $c_1 = 0$, $c_2 = 1$, $c_3 = -1$ and

$$\boldsymbol{M} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ -2 & 1 & 1 \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} \frac{1}{3}(2\mathsf{S}(\mathsf{x}_1)+1) & \mathsf{A}(\mathsf{x}_1) & \frac{1}{3}(\mathsf{S}(\mathsf{x}_1)-1) \\ \frac{2}{3}\mathsf{A}(\mathsf{x}_1) & \mathsf{S}(\mathsf{x}_1) & \frac{1}{3}\mathsf{A}(\mathsf{x}_1) \\ \frac{2}{3}(\mathsf{S}(\mathsf{x}_1)-1) & \mathsf{A}(\mathsf{x}_1) & \frac{1}{3}(\mathsf{S}(\mathsf{x}_1)+2) \end{bmatrix}, \quad \boldsymbol{K} = \begin{bmatrix} 1 & 0 & 0 \\ s_2\epsilon_2 & 1-s_2 & 0 \\ s_3\epsilon_3 & 0 & 1-s_3 \end{bmatrix}.$$

The modified equation of the bulk Finite Difference scheme from Theorem 1 is

$$\partial_t \phi(t,x) + \lambda\epsilon_2 \partial_x \phi(t,x) - \lambda\Delta x\Big(\frac{1}{s_2} - \frac{1}{2}\Big)\Big(\frac{2}{3} - \epsilon_2^2 + \frac{\epsilon_3}{3}\Big)\partial_{xx}\phi(t,x) = O(\Delta x^2), \qquad (t,x) \in \mathbb{R}_+ \times \mathbb{R}. \tag{50}$$

To have a stable bulk method in the $L^2$ metric, the dissipation coefficient must not be negative, hence $\epsilon_3 < -2 + 3\epsilon_2^2$ is forbidden, because the modulus of the consistency (or "physical") eigenvalue would initially increase above one for small wavenumbers, causing the bulk Finite Difference scheme to be unstable. Sufficient conditions are more involved to determine but can be checked numerically. Observe that the (50) does not depend on the choice of $s_3$. To obtain consistency with (1), we have to enforce $\epsilon_2 = V/\lambda$. Furthermore, two leverages are available to make the bulk Finite Difference scheme second-order consistent with the (1), namely taking $s_2 = 2$ or $s_2 \in ]0,2[$ and $\epsilon_3 = -2 + 3\epsilon_2^2$.

### 5.2.2 Conditions to achieve the time smoothness of the numerical solution

Assuming that $s_2, s_3 \neq 1$, we have that $Q = 2$, thus two initialisation schemes are to consider. Their modified equations, computed with the previous techniques and considering local initialisations following the conditions by Corollary 1—*i.e.* $\mathsf{w}_1 = 1$ and $\mathsf{w}_2 = \epsilon_2$—are as follows.

- **First initialisation scheme**: (9) for $n = 1$

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x\Big(\frac{1}{3} - \frac{\epsilon_2^2}{2} + \frac{s_3\epsilon_3}{6} + \frac{(1-s_3)\mathsf{w}_3}{6}\Big)\partial_{xx}\phi(0,x) = O(\Delta x^2), \qquad x \in \mathbb{R}. \tag{51}$$

  This scheme makes sense as initialisation scheme unless both $s_2 = s_3 = 1$ (*i.e.* $Q = 0$), where we observe that the diffusion coefficient in (51) becomes equal to the one from (50). In this case, the choice of $\mathsf{w}_3$ is unimportant, as expected.

Table 2: Different choices of parameters for the $D_1Q_3$ scheme ensuring match at order $O(\Delta x^2)$ between initialisation schemes and bulk Finite Difference scheme.

| Factors controlling dissipation | | Leverages to obtain compatible dissipation | |
|---|---|---|---|
| $s_2 = 1$ | $\epsilon_3 \geq -2 + 3\epsilon_2^2$ | $s_3 = 1$, any $w_3$ | (a) |
| | | $s_3 \neq 1$, $w_3 = \epsilon_3$ | (b) |
| $s_2 \neq 1$ | $\epsilon_3 > -2 + 3\epsilon_2^2$ | $s_3 = 2 - s_2$, $w_3 = (2(-2 + 3\epsilon_2^2) + (s_2 - 2)\epsilon_3)/s_2$ | (c) |
| | $\epsilon_3 = -2 + 3\epsilon_2^2$ | $s_3 = 1$, any $w_3$ | (d) |
| | | $s_3 \neq 1$, $w_3 = \epsilon_3$ | (e) |

- **Second initialisation scheme**: (9) for $n = 2$

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x\Big(\frac{(2-s_2)}{3} + \frac{(s_2-2)\epsilon_2^2}{2} + \frac{s_3(5-2s_2-s_3)\epsilon_3}{12} + \frac{(1-s_3)(4-2s_2-s_3)w_3}{12}\Big)\partial_{xx}\phi(0,x) = O(\Delta x^2), \tag{52}$$

for $x \in \mathbb{R}$. In the case where both $s_2 = s_3 = 1$ ($Q = 0$), we have the previously described situation. Taking $s_2 \neq 1$ and $s_3 = 1$ ($Q = 1$), we obtain the modified equation of the first starting scheme which is not an initialisation scheme

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x s_2\Big(\frac{1}{s_2} - \frac{1}{2}\Big)\Big(\frac{2}{3} - \epsilon_2^2 + \frac{\epsilon_3}{3}\Big)\partial_{xx}\phi(0,x) = O(\Delta x^2), \qquad x \in \mathbb{R},$$

which equals (50) up to the multiplication of the diffusion coefficient by $s_2$. This discrepancy is the remaining contribution of the initialisation on the evolution of the solution, as we have already observed for the $D_1Q_2$ in Section 5.1 for all initialisation except (40). Taking $s_2 = 1$ and $s_3 \neq 1$ ($Q = 1$), we have

$$\partial_t \phi(0,x) + \lambda\epsilon_2 \partial_x \phi(0,x) - \lambda\Delta x\Big(\frac{1}{3} - \frac{\epsilon_2^2}{2} + \frac{s_3(3-s_3)\epsilon_3}{12} + \frac{(1-s_3)(2-s_3)w_3}{12}\Big)\partial_{xx}\phi(0,x) = O(\Delta x^2),$$

for $x \in \mathbb{R}$, which is utterly different from (50): the choice of initialisation $w_3$ and the relaxation parameter $s_3$ influence the diffusivity, contrarily to (50).

*Remark* 2. The previous discussion again confirms that, for starting schemes which are not initialisation schemes, the choice of initialisations and relaxation parameters can change the modified equations compared to the bulk Finite Difference scheme and thus the dynamics of the method close to the beginning of the simulation. Moreover, even some parameters that do not influence the modified equation of the bulk Finite Difference scheme at a given order (see $s_3$ in this example) impact the modified equations of the starting schemes. This is due to the role of the parasitic eigenvalues in the initialisation process.

According to Proposition 3, it is enough to study the order $O(\Delta x^2)$ for the initialisation schemes to deduce the modified equations for any starting scheme. In order to match the diffusivity in both initialisation schemes, we set the following system

$$\begin{cases} \frac{1}{3} - \frac{\epsilon_2^2}{2} + \frac{s_3\epsilon_3}{6} + \frac{(1-s_3)w_3}{6} & = \Big(\frac{1}{s_2} - \frac{1}{2}\Big)\Big(\frac{2}{3} - \epsilon_2^2 + \frac{\epsilon_3}{3}\Big), \\ \frac{(2-s_2)}{3} + \frac{(s_2-2)\epsilon_2^2}{2} + \frac{s_3(5-2s_2-s_3)\epsilon_3}{12} + \frac{(1-s_3)(4-2s_2-s_3)w_3}{12} & = \Big(\frac{1}{s_2} - \frac{1}{2}\Big)\Big(\frac{2}{3} - \epsilon_2^2 + \frac{\epsilon_3}{3}\Big). \end{cases} \tag{53}$$

We have to interpret $\epsilon_2$ as fixed by the target problem and $\epsilon_3$ as well as $s_2$ by the choice of numerical dissipation of the bulk Finite Difference scheme, *i.e.* the right hand sides in (53). Therefore, the unknowns (or the leverages) are $s_3$ and $w_3$, forming a non-linear system. Eliminating $w_3$ from the second equation in (53) using the first one yields—following some algebra—the equation for $s_3$:

$$(1 - s_2)\Big(\frac{2}{3} - \epsilon_2^2 + \frac{\epsilon_3}{3}\Big)s_3 = (2 - s_2)(1 - s_2)\Big(\frac{2}{3} - \epsilon_2^2 + \frac{\epsilon_3}{3}\Big).$$

We have different cases to discuss which are summarized in Table 2 and which are obtained as detailed in Appendix D. We solely comment on (c): in [Bellotti et al., 2022], we have found that the choice

$$s_3 = 2 - s_2, \tag{54}$$

could yield a bulk Finite Difference scheme with three stages instead of four, in a way reminiscent of [D'Humières and Ginzburg, 2009, Kuzmin et al., 2011]. This phenomenon shall be the focus of Section 6. As far as the stability under this condition is concerned, the analytical conditions in this case are

$$s_2 \in ]0,2], \qquad \text{and} \qquad \begin{cases} |\epsilon_2| \leq 1, & -2 + 3\epsilon_2^2 \leq \epsilon_3 \leq 1, \quad \text{if} \quad s_2 \in ]0,2[, \\ |\epsilon_2| < 1, & \quad \text{if} \quad s_2 = 2, \end{cases}$$

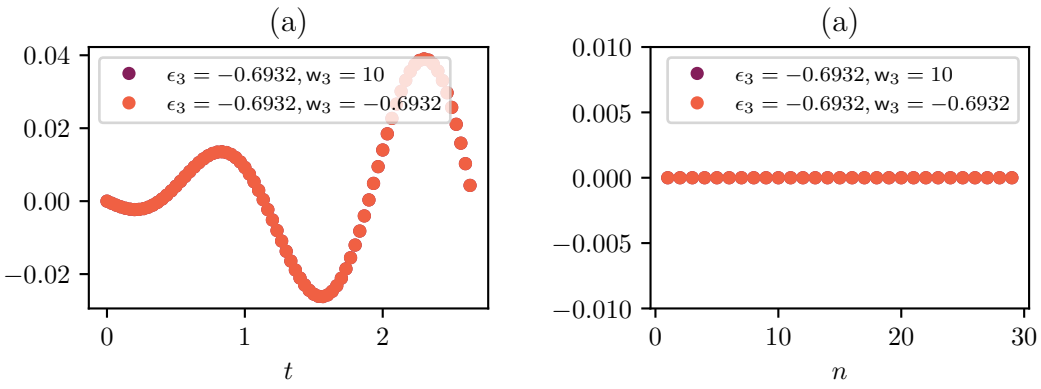


Figure 7: Left: test for smoothness in time close to $t = 0$ for the case (a) in Table 2: difference between exact and numerical solution at the eighth lattice point. As expected, regardless of the choice on $w_3$, the profile is smooth. Right: diffusion coefficient (factor in front of $-\lambda\Delta x\partial_{xx}$) in the modified equations for different $n$.

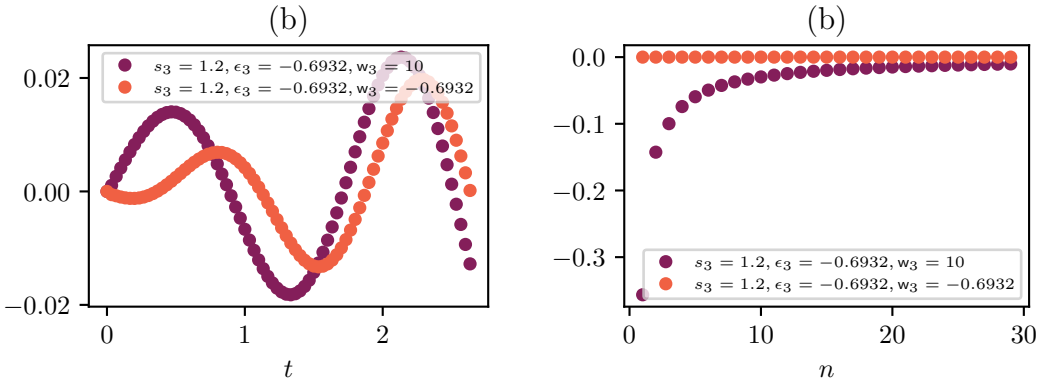


Figure 8: Left: test for smoothness in time close to $t = 0$ for the case (b) in Table 2 ($w_3 = -0.6932$) or violating this condition ($w_3 = 10$): difference between exact and numerical solution at the eighth lattice point. We observe radical differences in the profiles but the smoothness is not affected. Right: diffusion coefficient (factor in front of $-\lambda\Delta x\partial_{xx}$) in the modified equations for different $n$.

see Appendix E, where in $-2 + 3\epsilon_2^2 \leq \epsilon_3 \leq 1$, the left constraint enforces non-negative dissipation (stability for small wavenumbers) whereas the right one concerns large wavenumbers. Notice that the case $s_2 = 2$ corresponds to $s_3 = 0$, meaning that $m_3$ is conserved and thus going against the assumptions of the paper. This particular occurrence will be discussed in what follows.

### 5.2.3 Study of the time smoothness of the numerical solution

We repeat the numerical experiment by [Van Leemput et al., 2009] introduced in Section 5.1.2. Only $L^2$ stable configurations are considered. As long as the dissipation of the bulk Finite Difference scheme is large, time oscillations are damped and thus cannot be observed even if the diffusivities of the bulk Finite Difference scheme and the initialisation schemes are not the same. We therefore look for situations where the numerical diffusion is small or zero.

- $s_2 = 1$, $\epsilon_3 = -2 + 3\epsilon_2^2$, no dissipation, and $s_3 = 1$. This is the framework of (a) (*cf.* Table 2), where we can consider arbitrary $w_3$. This case is trivial because $Q = 0$. We see in Figure 7 that the profile remains smooth no matter the choice of $w_3$, as predicted by the theory.
- $s_2 = 1$, $\epsilon_3 = -2 + 3\epsilon_2^2$, no dissipation, and $s_3 = 1.2$, close to one for stability reasons. Thus we are in the setting of (b). In Figure 8, we see that the choice of $w_3$ changes the outcome, even if the time smoothness seems to be preserved in both cases. To explain this, on the one hand, we have to take into account that since we are compelled to take $s_3$ close

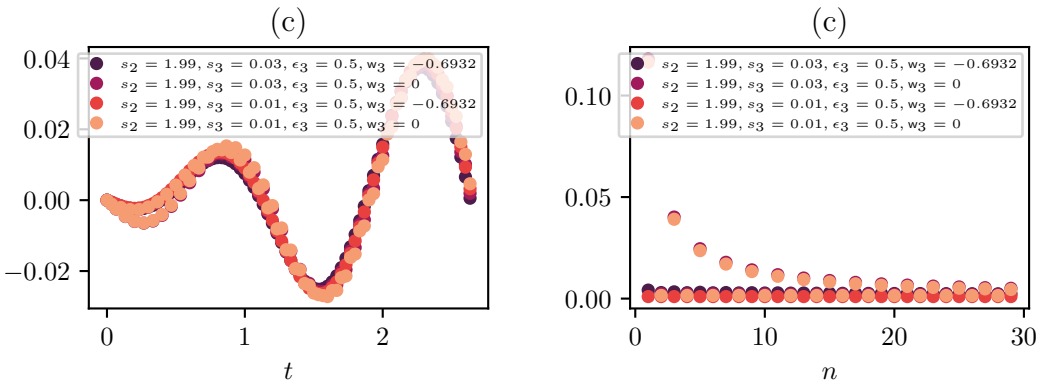


Figure 9: Left: test for smoothness in time close to $t = 0$ for the case (c) in Table 2: difference between exact and numerical solution at the eighth lattice point. The cases where $s_3 = 0.01$ violate the magic relation (54) $s_2 + s_3 = 2$ with minor influences on the spurious oscillation, whereas $w_3 = 0$ violates (73), with more tendency towards an initial boundary layer. Right: diffusion coefficient (factor in front of $-\lambda\Delta x\partial_{xx}$) in the modified equations for different $n$.

to one, we are not far from the previous case. On the other hand, even when the dissipation is not matched, it does not oscillate between time steps, unlike many initialisations for the $D_1Q_2$ scheme in Section 5.1. This is confirmed by the right image in Figure 8: the diffusivity behaves smoothly in $n$ and tends monotonically and quite rapidly to the bulk vanishing one.

- $s_2 = 1.99$, almost zero dissipation. We test (c), since (d) and (e) cannot be considered for stability reasons. In Figure 9, we observe that violating the magic relation (54) still enforcing (73) does not produce large spurious oscillations, likely because this has limited effects on the diffusion coefficient. Quite the opposite, violating (73) both with and without (54) produces an initial oscillating boundary layer. This is corroborated by the right image in Figure 9, where the reason for the observed oscillations is the highly non-smooth behaviour of the diffusion coefficient in $n$, as a result of having taken $s_2 \simeq 2$.

### 5.3 Conclusions

In this Section 5, we have observed in practice that the conditions to obtain consistent starting schemes found in Section 4 preserve second-order convergence when the bulk scheme is second-order consistent. Using an additional order for the modified equations introduced in Section 4, we obtain an extremely precise description of the behaviour of the $D_1Q_2$ close to the initial time, according to the initialisation at hand. The same has been done for a $D_1Q_3$ scheme. Finally, discussing the conditions to have the same dissipation between initialisation and bulk schemes for the $D_1Q_3$ has made the magic relations (54) known in the literature [D'Humières and Ginzburg, 2009, Kuzmin et al., 2011] turn up once more [Bellotti et al., 2022]. The investigation of these relations is central in the following Section 6.

## 6 A more precise evaluation of the number of initialisation schemes

In Section 4, we have observed that describing the behaviour of general lattice Boltzmann schemes close to the initial time above $O(\Delta x)$ order—using the modified equations—seems out of reach, due to the presence of many parasitic modes in the system. The question which we try to answer here—inspired by the findings on the $D_1Q_3$ in Section 5.2—concerns the existence of vast classes of lattice Boltzmann schemes for which a detailed description of the behaviour of the initialisation schemes is indeed possible. The idea is to investigate the possibility of having, from a purely algebraic standpoint, a very small number of initialisation schemes to be considered. For example, this would allow to avoid dealing— when trying to have the same dissipation coefficient between initialisation and bulk—with large non-linear systems such as (53), where the number and the complexity of equations grow with $Q$. The conditions to control the initialisation until a certain order in $\Delta x$ could be simpler thanks to the fact that we have a small number of initialisation steps. In this way, if something similar to Proposition 3 was valid, we could conclude that this control is enough to master the dynamics of the scheme at the considered orders eventually in time.

## 6.1 Lattice Boltzmann schemes as dynamical systems and observability

A preliminary step in this direction is to consider any lattice Boltzmann scheme Algorithm 1 as a linear time-invariant discrete-time system

$$\begin{aligned} \mathsf{z}\mathbf{m}(t,\boldsymbol{x}) &= \mathbf{E}\mathbf{m}(t,\boldsymbol{x}), \qquad && (t,\boldsymbol{x}) \in \Delta t\mathbb{N} \times \Delta x\mathbb{Z}^d, \\ \mathbf{m}(0,\boldsymbol{x}) & && \text{given for } \boldsymbol{x} \in \Delta x\mathbb{Z}^d, \end{aligned}$$

where the output is $\mathbf{y} = \boldsymbol{C}\mathbf{m}$ with matrix $\boldsymbol{C}$ of appropriate dimension. Since, from the very beginning of the paper, we are solely interested in the conserved moment $\mathsf{m}_1$, we select $\boldsymbol{C} = \boldsymbol{e}_1^{\mathsf{t}} \in \mathbb{R}^q$. As we already pointed out, see (9)

$$\mathsf{y}(n\Delta t, \boldsymbol{x}) = \mathsf{m}_1(n\Delta t, \boldsymbol{x}) = (\mathbf{E}^n\mathbf{m})_1(0,\boldsymbol{x}) = \boldsymbol{C}\mathbf{E}^n\mathbf{m}(0,\boldsymbol{x}), \qquad n \in \mathbb{N}, \quad \boldsymbol{x} \in \Delta x\mathbb{Z}^d,$$

thus we introduce the observability matrix of the system

$$\boldsymbol{\Omega} := \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{C}\mathbf{E} \\ \vdots \\ \boldsymbol{C}\mathbf{E}^{q-1} \end{bmatrix} \in \mathcal{M}_q(\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]).$$

If the system were set on a field (*e.g.* $\boldsymbol{\Omega} \in \mathcal{M}_q(\mathbb{R})$ or $\boldsymbol{\Omega} \in \mathcal{M}_q(\mathbb{C})$), it would be customary to call the system "observable" if and only if $\operatorname{rank}(\boldsymbol{\Omega}) = q$. This would mean that we could reconstruct the initial data $\mathbf{m}(0)$ from the observation of $\mathsf{y} = \mathsf{m}_1$ at times $n \in [\![0, q[\![$. Quite the opposite, in our case, since the non-zero entries of $\boldsymbol{\Omega}$ are in general not invertible (for $d = 1$, $\mathsf{S}(\mathsf{x}_1)$ and $\mathsf{A}(\mathsf{x}_1)$ are examples of this), we cannot proceed in the same way, because the observability matrix $\boldsymbol{\Omega}$ can never be a unit.

For systems over commutative rings, different definition of observability are available in the literature: we list a few of them in the following Definition.

**Definition 2** (Observability)**.** The system is said to be

- "observable" according to [Brewer et al., 1986, Theorem 2.6], if the application represented by the left action of $\boldsymbol{\Omega}$ is injective.
- "observable" according to [Fliess and Mounier, 1998], if $\boldsymbol{\Omega}$ has left inverse.
- "hyper-observable" according to [Fliess and Mounier, 1998], if the unobservable sub-space $\mathcal{N} := \ker(\boldsymbol{\Omega})$— where operators act on lattice functions[1]—is trivial: $\mathcal{N} = \{\mathbf{0}\}$.

Furthermore, [Brewer et al., 1986, Theorem 2.6] gives the following criterion to check observability.

**Theorem 2** ([Brewer et al., 1986] Observability criterion)**.** *The system is "observable" according to [Brewer et al., 1986] if and only if the ideal of* $\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$ *generated by* $\det(\boldsymbol{\Omega})$ *is such that its annihilator is zero.*

We also define the "observability index" $o \leq Q + 1$ mimicking the definition for systems over fields as

$$o := \max_{\ell \in \mathbb{N}} \operatorname{rank}(\boldsymbol{\Omega}_\ell), \quad \text{where} \quad \boldsymbol{\Omega}_\ell := \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{C}\mathbf{E} \\ \vdots \\ \boldsymbol{C}\mathbf{E}^{\ell-1} \end{bmatrix} \in \mathcal{M}_{\ell \times q}(\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]),$$

and $\operatorname{rank}(\cdot)$ stands for the row rank of a matrix over a ring according to Definition 10.6 in [Blyth, 2018].

*Example* 4. Considering Example 1 treated in Section 5.1, we have that

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & 0 \\ \mathsf{S}(\mathsf{x}_1) + s_2\epsilon_2\mathsf{A}(\mathsf{x}_1) & (1 - s_2)\mathsf{A}(\mathsf{x}_1) \end{bmatrix},$$

hence $o = Q + 1 = 2$ if $s_2 \neq 1$ and $o = Q + 1 = 1$ if $s_2 = 1$. When $s_2 = 1$, we have $\mathcal{N} = \{(0, \mathsf{m}_2)^{\mathsf{t}} \ : \ \text{for arbitrary } \mathsf{m}_2 = \mathsf{m}_2(x) \text{ lattice function}\}$, which adheres to the intuition that we cannot know the non-conserved moment $\mathsf{m}_2$ by looking at the conserved moment $\mathsf{m}_1$ if the relaxation is made on the equilibrium, regardless of the structure of $\mathsf{m}_2$. When $s_2 \neq 1$,

[1] Observe that the kernel is the left null space: indeed the left action of elements in $\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$ can operate both on lattice function and operators in $\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$, whereas the right action is reserved for operators in $\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \ldots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$.

we have $\mathcal{N} = \{(0, \mathsf{m}_2)^\mathsf{t} \ : \ \text{for any } \mathsf{m}_2 = \mathsf{m}_2(x) \text{ lattice function such that } \mathsf{A}(\mathsf{x}_1)\mathsf{m}_2 = 0\}$. We see that the unobservable subspace is non-trivial even when $o = q = 2$, contrarily to the case of systems with matrix $\mathbf{E}$ and $\mathbf{\Omega}$ with entries in a field. The unobservable states are those in which the first component is zero and the discrete derivative $\mathsf{A}(\mathsf{x}_1)$ of the second component is zero everywhere, for example because the second component is constant or takes one given value on all even point and another one on all odd point. The interesting reader can consult Appendix F to find a numerical experiment showcasing the structure of $\mathcal{N}$ for this scheme. We finally comment on the notions from Definition 2.

- $\det(\mathbf{\Omega}) = (1 - s_2)\mathsf{A}(\mathsf{x}_1)$, thus the ideal to consider (*cf.* Theorem 2) is $\{\mathsf{d}(1 - s_2)\mathsf{A}(\mathsf{x}_1) \ : \ \mathsf{d} \in \mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}]\}$. On the one hand, if $s_2 = 1$, then any operator in $\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}]$ multiplied at the left of any element of the ideal is an annihilator, thus the system is not observable according to [Brewer et al., 1986]. On the other hand, if $s_2 \neq 1$, then the only element annihilating any element of the ideal is zero, thus the system is observable according to [Brewer et al., 1986].
- For any $s_2$, we see that $\mathbf{\Omega}$ does not admit left inverse, therefore it is not observable according to [Fliess and Mounier, 1998].
- For any $s_2$, the system is not hyper-observable according to [Fliess and Mounier, 1998] due to the non-trivial $\mathcal{N}$.

For these reasons, we infer that the observability according to [Brewer et al., 1986] is the one more closely adhering—between those issued from Definition 2—to our definition of observability index $o$.

## 6.2 Reduced number of initialisation schemes for non-observable systems

Following the discussion in [Bellotti et al., 2022], we can introduce $\mathbf{p}_o \in (\mathbb{R}[\mathsf{x}_1, \mathsf{x}_1^{-1}, \dots, \mathsf{x}_d, \mathsf{x}_d^{-1}])^o$ such that

$$\mathbf{p}_o \mathbf{\Omega}_o = -\boldsymbol{C}\mathbf{E}^o. \tag{55}$$

The solution of this problem exists thanks to the definition of the observability index $o$. We then introduce the monic polynomial (keep in mind that the indices in vectors like $\mathbf{p}_o$ start from one)

$$\Psi_o(\mathsf{z}) := \mathsf{z}^o + \sum_{n=1}^{o} \mathsf{p}_{o,n} \mathsf{z}^{n-1}, \tag{56}$$

which by construction (55) annihilates the first row of $\mathbf{E}$, since $\boldsymbol{C} = \boldsymbol{e}_1^\mathsf{t}$. Moreover, we have shown in [Bellotti et al., 2022] that $\Psi_o(\mathsf{z})$ divides $\det(\mathsf{z}\boldsymbol{I} - \mathbf{E})$, whence if $o = Q + 1$, we naturally have $\Psi_o(\mathsf{z}) = \mathsf{z}^{Q+1-q}\det(\mathsf{z}\boldsymbol{I} - \mathbf{E})$. We therefore obtain the following corresponding bulk Finite Difference scheme based on $\Psi_o$ given by Algorithm 3, coinciding with Algorithm 2 when $o = Q + 1$.

**Algorithm 3** Corresponding Finite Difference scheme based on $\Psi_o$.

- Given $\mathbf{m}(0, \boldsymbol{x})$ for every $\boldsymbol{x} \in \Delta x \mathbb{Z}^d$.
- **Initialisation schemes**. For $n \in [\![1, o-1]\!]$

$$\mathsf{m}_1(n\Delta t, \boldsymbol{x}) = \boldsymbol{C}\mathbf{E}^n \mathbf{m}(0, \boldsymbol{x}), \qquad \boldsymbol{x} \in \Delta x \mathbb{Z}^d. \tag{57}$$

- **Corresponding bulk Finite Difference scheme**. For $n \in [\![o-1, +\infty[\![$

$$\mathsf{m}_1((n+1)\Delta t, \boldsymbol{x}) = -\sum_{\ell=q-o}^{q-1} \mathsf{p}_{o,o+\ell+1-q} \mathsf{m}_1((n+\ell+1-q)\Delta t, \boldsymbol{x}), \qquad \boldsymbol{x} \in \Delta x \mathbb{Z}^d. \tag{58}$$

The lack of observability is indeed the reason why, as previously announced in Section 5.2, one can find a bulk Finite Difference scheme with less time steps than what is prescribed by the characteristic polynomial of $\mathbf{E}$. From a different perspective, this is the so-called "pole-zero cancellation" in the transfer function—see for example [Åström and Murray, 2008, Chapter 8.3] or in [Hendricks et al., 2008, Chapter 3.9]—associated with the system and taken from control theory. In our framework, the transfer function is

$$\mathsf{H}(\mathsf{z}) = \boldsymbol{C} \frac{\overbrace{\operatorname{adj}(\mathsf{z}\boldsymbol{I} - \mathbf{A})\mathbf{B}\boldsymbol{\epsilon}}^{\text{control by equilibria}}}{\underbrace{\det(\mathsf{z}\boldsymbol{I} - \mathbf{A})}_{\text{state}}} = \boldsymbol{C}(\mathsf{z}\boldsymbol{I} - \mathbf{A})^{-1}\mathbf{B}\boldsymbol{\epsilon},$$

with $\mathbf{A} := \mathbf{T}(\boldsymbol{I} - \boldsymbol{S})$ and $\mathbf{B} := \mathbf{T}\boldsymbol{S}$ defined as in [Bellotti et al., 2022, Bellotti, 2023], with $\mathbf{E} = \mathbf{A} + \mathbf{B}\boldsymbol{\epsilon} \otimes \boldsymbol{e}_1$.

*Example* 5. We come back to the scheme of Section 5.2 where we have selected the choice of magic parameter equal to 1/4, that is $s_2 + s_3 = 2$. We also assume that $s_2 \neq 1$ to keep things non-trivial. In this case, it can be seen that $o = 2 < 3$, whereas $Q + 1 = 3$. Moreover, we obtain

$$\begin{aligned}
&\det(\mathsf{z}\boldsymbol{I} - \mathbf{E}) = (\mathsf{z} + (1 - s_2))\Psi_2(\mathsf{z}),\\
&\qquad \text{with} \quad \Psi_2(\mathsf{z}) = \mathsf{z}^2 + (-s_2\epsilon_2\mathsf{A}(\mathsf{x}_1) + \tfrac{1}{3}(s_2 - 2)(2\mathsf{S}(\mathsf{x}_1) + 1) + \tfrac{1}{3}\epsilon_3(s_2 - 2)(\mathsf{S}(\mathsf{x}_1) - 1))\mathsf{z} + (1 - s_2),\\
&\qquad \text{or equivalently} \quad \mathsf{H}(\mathsf{z}) = \frac{(\mathsf{z} + (1 - s_2))(s_2\epsilon_2\mathsf{A}(\mathsf{x}_1) + \frac{1}{3}\epsilon_3(2 - s_2)(\mathsf{S}(\mathsf{x}_1) - 1))\mathsf{z}}{(\mathsf{z} + (1 - s_2))(\mathsf{z}^2 + \frac{1}{3}(s_2 - 2)(2\mathsf{S}(\mathsf{x}_1) + 1)\mathsf{z} + (1 - s_2))}.
\end{aligned}$$

The Finite Difference scheme coming from $\Psi_2(\mathsf{z})$ becomes a leap-frog scheme for $s_2 = 2$. Otherwise, it is a centered discretisation with a certain amount of numerical dissipation. A first question which might arise concerns the modified equation for the bulk Finite Difference scheme obtained using $\det(\mathsf{z}\boldsymbol{I} - \mathbf{E})$, see Algorithm 2, *versus* those obtained by $\Psi_2(\mathsf{z})$, see Algorithm 3. The answer is that they are same at any order because the eigenvalue $(s_2 - 1)$ does not contribute to the consistency (being constant through wavenumbers and thus being a mere numerical eigenvalue) and it can be easily checked that $\Psi_2(\mathsf{z})$ yields the same modified equation, since it contains the consistency eigenvalue [Strikwerda, 2004]. As far as stability is concerned, the stability constraints for the two bulk Finite Difference schemes are the same because $|s_2 - 1| < 1$ for $s_2 \in ]0, 2[$. The case $s_2 = 2$ might produce instabilities because of the presence of multiple roots in $\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}})$ on the unit circle. However, in this case, there is an additional conserved moment $\mathsf{m}_3$ and we know that the *von Neumann* condition for systems is that no root is outside the unit circle (no precise indication is provided for those on the unit circle), but this is only necessary for stability [Gustafsson et al., 1995, Theorem 5.2.2]. Therefore, the presence of multiple eigenvalues on the unit circle (and in particular those concerning consistency which are now more than one) cannot allow to deduce that the scheme is unstable. The stability conditions are analytically computed in Appendix E.

Concerning the notion of observability by [Brewer et al., 1986], we have that $\det(\boldsymbol{\Omega}) = 0$, hence the system is not observable, according to Theorem 2. If we want to characterize the unobservable sub-space, we have that, since $s_2 \neq 1$, it is given by $\mathcal{N} = \{(0, \mathsf{m}_2, \mathsf{m}_3)^\mathsf{t} \ : \ \text{for any } \mathsf{m}_2 = \mathsf{m}_2(x),\ \mathsf{m}_3 = \mathsf{m}_3(x) \text{ lattice functions such that } \mathsf{A}(\mathsf{x}_1)\mathsf{m}_2 = 1/3(\mathsf{S}(\mathsf{x}_1) - 1)\mathsf{m}_3\}$. Recall that $\mathsf{A}(\mathsf{x}_1) = (\mathsf{x}_1 - \mathsf{x}_1{}^{-1})/2$ and $\mathsf{S}(\mathsf{x}_1) - 1 = (\mathsf{x}_1 - 2 + \mathsf{x}_1{}^{-1})/2$, which means that the initial states belonging to $\mathcal{N}$ are those with zero first moment everywhere and such that the centered approximation of the first derivative of the second moment is proportional—with ratio 1/6—to the centered approximation of the second derivative of the third moment, at any point of the lattice. The numerical verification of the expression found for $\mathcal{N}$ is provided in Appendix G for the interested reader.

As already remarked in [Saad, 1989, Bellotti et al., 2022], the cases were $Q + 1 \neq o$ are extremely peculiar. Indeed, the situation described in Example 5 and in the forthcoming Section 6.3 are the only examples we were able to find. Loosely speaking, both $o$ and $Q$ measure the speed of saturation of the image of the scheme $\mathbf{E}$ concerning the conserved moment. Once the generated sub-spaces saturate, the evolution of the conserved moment at the new time-step can be recast as function of itself at the previous steps. The fact that the Cayley-Hamilton theorem holds (concerning $Q$) and that the polynomial $\Psi_o$ (concerning $o$) annihilates the first row of $\mathbf{E}$ introduce—as previously shown—a set of linear constraints on $\mathsf{m}_1$, solution of the lattice Boltzmann scheme.

It should be emphasized that Proposition 3 is still valid turning $Q$ into $o - 1$, (9) into (57) and (10) into (58). This is fundamental, because Proposition 3 ensures to control the whole dynamics of the scheme by mastering it in the initialisation layer. The aim of studying observability is to characterise in which case the initialisation layer (57), thus what we need to control, is simple but still determines the dynamics eventually in time. This property comes from the fact that the root of $\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(\boldsymbol{\xi}\Delta x))$ setting the consistency of the scheme—*i.e.* being one in the low-frequency limit—is also a root of $\hat{\Psi}_o(\mathsf{z})$. This is a consequence of the fact that $\hat{\Psi}_o(\mathsf{z})$ annihilates the first row of $\hat{\mathbf{E}}(\boldsymbol{\xi}\Delta x)$.

**Proposition 10.** *Let $\hat{\mathsf{g}}_1 \equiv \hat{\mathsf{g}}_1(\boldsymbol{\xi}\Delta x)$ be the unique root of $det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(\boldsymbol{\xi}\Delta x))$ such that*

$$\hat{\mathsf{g}}_1(\boldsymbol{\xi}\Delta x) = 1 + O(|\boldsymbol{\xi}\Delta x|) \tag{59}$$

*in the limit $|\boldsymbol{\xi}\Delta x| \ll 1$, which is the one determining the consistency and the modified equation of the numerical scheme. Then, $\hat{\mathsf{g}}_1$ is also a root of $\hat{\Psi}_o(\mathsf{z})$.*

*Proof.* Let us show this, using the Fourier representation and considering $d = 1$ for the sake of keeping notations simple. Recall that

$$\hat{\Psi}_o(\hat{\mathbf{E}}(\xi\Delta x)) = \hat{\mathbf{E}}(\xi\Delta x)^o + \sum_{n=1}^{o} \hat{\mathsf{p}}_{o,n}(\xi\Delta x)\hat{\mathbf{E}}(\xi\Delta x)^{n-1} = \begin{bmatrix} 0 & \cdots & 0 \\ \star & \cdots & \star \\ \vdots & & \vdots \\ \star & \cdots & \star \end{bmatrix}, \tag{60}$$

where the starred $\star$ entries are not necessarily zero. Notice that, whatever the scaling between space and time, we have

that for every $\ell \in \mathbb{N}$

$$\hat{\mathbf{E}}(\xi\Delta x)^\ell = \begin{bmatrix} 1 & \cdots & 0 \\ \star & \cdots & \star \\ \vdots & & \vdots \\ \star & \cdots & \star \end{bmatrix} + O(|\xi\Delta x|) \tag{61}$$

in the limit $|\xi\Delta x| \ll 1$. In particular, for the acoustic scaling, we have $\hat{\mathbf{E}}(\xi\Delta x)^\ell = \boldsymbol{K}^\ell + O(|\xi\Delta x|)$, where $\boldsymbol{K}^\ell$ has the property stated by (61) and $\boldsymbol{K}$ is the collision matrix. Again taking $|\xi\Delta x| \ll 1$ and considering that $\hat{\mathsf{p}}_{o,n}(\xi\Delta x) = \hat{p}^{(0)}_{o,n} + O(|\xi\Delta x|)$, selecting the very first entry in (60) yields, using (61)

$$1 + \sum_{n=1}^{o} \hat{p}^{(0)}_{o,n} = O(|\xi\Delta x|). \tag{62}$$

Since $\mathbb{C}$ is an algebraically closed field, we can write $\hat{\Psi}_o(\mathsf{z}) = \prod_{\ell=1}^{\ell=o}(\mathsf{z} - \hat{\mathsf{r}}_\ell(\xi\Delta x))$, where $\hat{\mathsf{r}}_\ell$ for $\ell \in [\![1, o]\!]$ are the roots of $\hat{\Psi}_o(\mathsf{z})$. These are also part of the roots of $\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(\xi\Delta x))$ since $\hat{\Psi}_o(\mathsf{z})$ divides $\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(\xi\Delta x))$. The question is whether the roots of $\hat{\Psi}_o(\mathsf{z})$ include the one of $\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(\xi\Delta x))$, indicated by $\hat{\mathsf{g}}_1(\xi\Delta x)$, being the only one such that $\hat{\mathsf{g}}_1(\xi\Delta x) = 1 + O(|\xi\Delta x|)$ in the limit $|\xi\Delta x| \ll 1$, see (26), and which totally dictates consistency (and the modified equations). Considering $\mathsf{z} = 1$ in (56) gives

$$\hat{\Psi}_o(1) = 1 + \sum_{n=1}^{o} \hat{\mathsf{p}}_{o,n}(\xi\Delta x).$$

Taking the limit $|\xi\Delta x| \ll 1$, we are left with

$$\prod_{\ell=1}^{o}(1 - \hat{r}^{(0)}_\ell) + O(|\xi\Delta x|) = 1 + \sum_{n=1}^{o} \hat{p}^{(0)}_{o,n} + O(|\xi\Delta x|) = O(|\xi\Delta x|),$$

thanks to (62), where $\hat{\mathsf{r}}_\ell = \hat{r}^{(0)}_\ell + O(|\xi\Delta x|)$. This gives $\prod_{\ell=1}^{\ell=o}(1 - \hat{r}^{(0)}_\ell) = 0$, hence at least one $\hat{r}^{(0)}_\ell = 1$. Since the roots $\hat{\mathsf{r}}_\ell$ for $\ell \in [\![1, o]\!]$ are a subset of those of $\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(\xi\Delta x))$, where only one has the desired property (59), then the latter is also a root of $\hat{\Psi}_o(\mathsf{z})$, let us say $\hat{\mathsf{r}}_1 \equiv \hat{\mathsf{g}}_1$. □

## 6.3 An important case: link $\mathrm{D}_d\mathrm{Q}_{1+2W}$ two-relaxation-times schemes with magic parameters equal to $1/4$

We are now ready to consider a quite wide class of schemes [D'Humières and Ginzburg, 2009] for which very little initialisation schemes are to consider, namely $o$ is particularly small. The "observable" features of these schemes are to some extent independent from $d$ and the choice of the $q = 1 + 2W$ discrete velocities. This boils down to a quite general application of the ideas of Section 6.2.

### 6.3.1 Description of the schemes

Consider any spatial dimension $d$ and $q = 1 + 2W$ velocities with $W \in \mathbb{N}^*$, which is the number of so-called "links". The velocities should be opposite along each link such that

$$\boldsymbol{c}_1 = \boldsymbol{0}, \qquad \boldsymbol{c}_{2j} = -\boldsymbol{c}_{2j+1} \in \mathbb{Z}^d, \qquad j \in [\![1, W]\!], \tag{63}$$

and the moment matrix

$$\boldsymbol{M} = \left[\begin{array}{c|cc|c|cc} 1 & 1 & 1 & \cdots & 1 & 1 \\ \hline 0 & 1 & -1 & & & \\ 0 & 1 & 1 & & & \\ \hline \vdots & & & \ddots & & \\ \hline 0 & & & & 1 & -1 \\ 0 & & & & 1 & 1 \end{array}\right] \in \mathcal{M}_{1+2W}(\mathbb{R}). \tag{64}$$

Here, empty blocks shall indicate null blocks of suitable size. The relaxation parameters should be such that

$$s_{2\ell} = s \in ]0, 2], \qquad s_{2\ell+1} = 2 - s, \qquad \ell \in [\![1, W]\!]. \tag{65}$$

This is equivalent to having the so-called "magic parameter" of every block $\ell \in [\![1, W]\!]$, given by $(1/s_{2\ell} - 1/2)(1/s_{2\ell+1} - 1/2)$, equal to $1/4$.

### 6.3.2 Observability and number of initialisation steps

The study of the observability of the previously described schemes is carried in the following result.

**Proposition 11.** *The characteristic polynomial of the scheme matrix* $\mathbf{E}$ *for the schemes given by* (63), (64) *and* (65) *is given by*

$$det(\mathsf{z}\boldsymbol{I}-\mathbf{E}) = (\mathsf{z}+(1-s))(\mathsf{z}^2-(1-s)^2)^{W-1}\Psi_2(\mathsf{z}),$$

*where*

$$\Psi_2(\mathsf{z}) = \mathsf{z}^2+(s-2)\mathsf{z}+(1-s)-\mathsf{z}s\sum_{\ell=1}^{W}\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})\epsilon_{2\ell}+\mathsf{z}(s-2)\sum_{\ell=1}^{W}(\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})-1)\epsilon_{2\ell+1} \tag{66}$$

*annihilates the first row of the matrix* $\mathbf{E}$. *Therefore* $o=2$ *if* $s\neq 1$ *and* $o=1$ *if* $s=1$. *Equivalently, the transfer function of the system is given by*

$$\mathsf{H}(\mathsf{z}) = \frac{(\mathsf{z}+(1-s))(\mathsf{z}^2-(1-s)^2)^{W-1}\Big(s\sum_{\ell=1}^{\ell=W}\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})\epsilon_{2\ell}+(2-s)\sum_{\ell=1}^{\ell=W}(\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})-1)\epsilon_{2\ell+1}\Big)\mathsf{z}}{(\mathsf{z}+(1-s))(\mathsf{z}^2-(1-s)^2)^{W-1}(\mathsf{z}^2+(s-2)\mathsf{z}+(1-s))}.$$

By Proposition 11, $\Psi_2(\mathsf{z})$ yields a bulk Finite Difference scheme according to Algorithm 3. As for Example 5, the modified equations of the method obtained by $\Psi_2(\mathsf{z})$ and by $\det(\mathsf{z}\boldsymbol{I}-\mathbf{E})$ are the same because the remaining roots do not concern consistency with respect to (1). The only consistency eigenvalue is one of the two roots of $\Psi_2(\mathsf{z})$, thus present in both schemes. The case $s=2$ apparently questions the previous claim since by looking at the proof of Proposition 3, a scheme consistent with (1) has only one eigenvalue equal to one for small wavenumbers. This is not a contradiction because in this case $s_{2\ell+1}=0$ for $\ell\in[\![1,W]\!]$, thus the corresponding moments are conserved [Ginzburg et al., 2008], whereas Theorem 1 has been demonstrated under the assumption that $s_i\neq 0$ for $i\in[\![2,q]\!]$ and the whole paper relies on the assumption that we deal only with one conserved moment. The moments $\mathsf{m}_{2\ell+1}$ for $\ell\in[\![1,W]\!]$ are conserved not because their equilibrium value equals the respective moment itself, but rather since their corresponding relaxation parameter is zero. A valid proof of Theorem 1 for several conserved moments has to follow the indications of [Bellotti et al., 2022, Bellotti, 2023] and would still lead to (11). Using the results of [Bellotti, 2023] with $s=2$, we would get, for the first moment

$$\partial_t\phi(t,\boldsymbol{x})+\lambda\sum_{\ell=1}^{W}\epsilon_{2\ell}\sum_{|\mathbf{n}|=1}\boldsymbol{c}_{2\ell}^{\mathbf{n}}\partial_{\boldsymbol{x}}^{\mathbf{n}}\phi(t,\boldsymbol{x})=O(\Delta x). \tag{67}$$

For the conserved moments $\mathsf{m}_{2\ell+1}$ for $\ell\in[\![1,W]\!]$, we obtain

$$\partial_t m_{2\ell+1}(t,\boldsymbol{x})+\lambda\epsilon_{2\ell}\sum_{|\mathbf{n}|=1}\boldsymbol{c}_{2\ell}^{\mathbf{n}}\partial_{\boldsymbol{x}}^{\mathbf{n}}\phi(t,\boldsymbol{x})=O(\Delta x). \tag{68}$$

Observe that the equation (67) for the moment of interest is indeed independent of the other conserved moments, as desired (no possible coupling *via* the equilibria, for they depend only on the first conserved moment). Quite the opposite, the equations (68) for the "inadvertently" conserved moments couple them with the first one. The first conserved moment is going to evolve alone, as usual, and the dynamics of the other conserved moments is going to be coupled with the one of $\mathsf{m}_1$ according to (68). Still, we are not interested in the latter moments. Coming back to our case, where we operate as only one moment was conserved even when $s=2$, the multiplicative factor $(\mathsf{z}+(1-s))(\mathsf{z}^2-(1-s)^2)^{W-1}\asymp(\exp(\frac{\Delta x}{\lambda}\partial_t)-1)(\exp(2\frac{\Delta x}{\lambda}\partial_t)-1)^{W-1}=O(\Delta x^W)$ in front of the amplification polynomial $\Psi_2$ of a leap-frog scheme is a series of time differential operators starting with a term of kind $\Delta x^W\partial_t^W$. Thus, if we compute the modified equation of the corresponding Finite Difference scheme obtained as we were dealing only with one conserved moment (*i.e.* (7)) whereas several conserved moments are present, we would obtain a sort of wave equation with time derivative of order $1+W$. This is unsurprising since this kind of equation feature $1+W$ "consistency" eigenvalues (one of which, the actual one, is inside $\Psi_2(\mathsf{z})$) which values equal one for small wavenumbers.

Coming back to generic $s$, the stability conditions of the corresponding Finite Difference obtained by using $\Psi_2(\mathsf{z})$ instead of $\det(\mathsf{z}\boldsymbol{I}-\mathbf{E})$ are the same because the remaining roots are constant in wavenumber and do not exceed modulus one when $s\in]0,2[$. The case $s=2$ might produce instabilities because of the presence of multiple roots of $\det(\mathsf{z}\boldsymbol{I}-\hat{\mathbf{E}})$ on the unit circle. However, in this case, there are additional conserved moments and the *von Neumann* condition for systems is that no root is outside the unit circle, with no precision concerning multiple ones on the unit circle. Still this condition is only necessary for stability. Therefore, the presence of multiple eigenvalues on the unit circle (and in particular those concerning consistency which are now $1+W$) cannot allow to deduce that the scheme is unstable. This should be precisely tested in the case where $W\geq 2$, for example, taking a $\mathrm{D_1Q_5}$ scheme.

Since $\det(\boldsymbol{\Omega})=0$, the system is not observable according to [Brewer et al., 1986]. However, it is not easy to generally characterize the unobservable sub-space $\mathcal{N}$, because this sub-space inflates with $d$ and $W$ due to the rank-nullity theorem. To explain this difficulty, consider that $\Psi_2$ from (66) is essentially scheme independent and concerns the "observable" part of

the system relative to $\mathrm{span}(\boldsymbol{\Omega})$, whereas $\mathcal{N} = \ker(\boldsymbol{\Omega})$ must be highly scheme dependent because it pertains to the remaining "unobservable" part of the system, which is encoded in the quotient $\det(z\boldsymbol{I} - \mathbf{E})/\Psi_2(z)$ between polynomials. Let us now proceed to the proof of Proposition 11.

*Proof of Proposition 11.* The transport matrix is given by

$$\mathbf{T} = \left[\begin{array}{c|cc|c|cc} 1 & \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_2}) & \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_2}) - 1 & \cdots & \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2W}}) & \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2W}}) - 1 \\ \hline 0 & \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_2}) & \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_2}) & & & \\ 0 & \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_2}) & \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_2}) & & & \\ \hline \vdots & & & \ddots & & \\ \hline 0 & & & & \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2W}}) & \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2W}}) \\ 0 & & & & \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2W}}) & \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2W}}) \end{array}\right] \in \mathcal{M}_{1+2W}(\mathbb{R}[x_1, x_1^{-1}, \ldots, x_d, x_d^{-1}]).$$

We have that $\det(z\boldsymbol{I} - \mathbf{E}) = \det(z\boldsymbol{I} - \mathbf{A} - \mathbf{B}\boldsymbol{\epsilon} \otimes \boldsymbol{e}_1) = \det(z\boldsymbol{I} - \mathbf{A}) - \boldsymbol{e}_1^{\mathsf{t}} \mathrm{adj}(z\boldsymbol{I} - \mathbf{A})\mathbf{B}\boldsymbol{\epsilon}$, using the matrix determinant lemma [Horn and Johnson, 2012]. Also

$$\mathrm{adj}(z\boldsymbol{I} - \mathbf{A})_{11} = \prod_{\ell=1}^{W} \det(z\boldsymbol{I} - \tilde{\mathbf{T}}_\ell \mathrm{diag}(1-s, s-1)) = \prod_{\ell=1}^{W} (z^2 - (1-s)^2) = (z^2 - (1-s)^2)^W,$$

with

$$\tilde{\mathbf{T}}_\ell = \begin{bmatrix} \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) & \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) \\ \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) & \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) \end{bmatrix}.$$

We only treat $\mathrm{adj}(z\boldsymbol{I} - \mathbf{A})_{12}$ and $\mathrm{adj}(z\boldsymbol{I} - \mathbf{A})_{13}$, since the following entries read the same except for the indices of the involved shift operators.

$$\mathrm{adj}(z\boldsymbol{I} - \mathbf{A})_{12}$$
$$= -\det \left[\begin{array}{cc|c|c|c} (s-1)\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_2}) & (1-s)(\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_2}) - 1) & \star & \cdots & \star \\ (s-1)\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_2}) & z + (1-s)\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_2}) & & & \\ \hline & & z\boldsymbol{I} - \tilde{\mathbf{T}}_2 \mathrm{diag}(1-s, s-1) & \star & \star \\ \hline & & & \ddots & \star \\ \hline & & & & z\boldsymbol{I} - \tilde{\mathbf{T}}_W \mathrm{diag}(1-s, s-1) \end{array}\right],$$

where the $\star$ blocks are not necessarily zero but do not need to be further characterized, since this is a determinant of a block upper triangular matrix, whence

$$\begin{aligned} \mathrm{adj}(z\boldsymbol{I} - \mathbf{A})_{1,2\ell} &= -(z^2 - (1-s)^2)^{W-1} \det \begin{bmatrix} (s-1)\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) & (1-s)(\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) - 1) \\ (s-1)\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) & z + (1-s)\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) \end{bmatrix}, \\ &= (1-s)(z + (1-s))(z^2 - (1-s)^2)^{W-1} \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}), \qquad \ell \in [\![1, W]\!]. \end{aligned}$$

The analogous computation for the odd moments yields

$$\begin{aligned} \mathrm{adj}(z\boldsymbol{I} - \mathbf{A})_{1,2\ell+1} &= (z^2 - (1-s)^2)^{W-1} \det \begin{bmatrix} (s-1)\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) & (1-s)(\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) - 1) \\ z(s-1)\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) & (1-s)\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) \end{bmatrix} \\ &= (s-1)(z + (1-s))(z^2 - (1-s)^2)^{W-1} (\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) - 1), \qquad \ell \in [\![1, W]\!]. \end{aligned}$$

Some algebra provides, for $\ell \in [\![1, W]\!]$

$$(\mathbf{B}\boldsymbol{\epsilon})_p = \begin{cases} s\sum_{r=1}^{W} \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2r}})\epsilon_{2r} + (2-s)\sum_{r=1}^{r=W} (\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2r}}) - 1)\epsilon_{2r+1}, & p = 1, \\ s\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})\epsilon_{2\ell} + (2-s)\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})\epsilon_{2\ell+1}, & p = 2\ell, \\ s\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})\epsilon_{2\ell} + (2-s)\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})\epsilon_{2\ell+1}, & p = 2\ell + 1, \end{cases}$$

and thus, after tedious computations

$$\boldsymbol{e}_1^{\mathsf{t}} \mathrm{adj}(z\boldsymbol{I} - \mathbf{A})\mathbf{B}\boldsymbol{\epsilon} = z(z + (1-s))(z^2 - (1-s)^2)^{W-1} \Big( s \sum_{\ell=1}^{W} \mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}})\epsilon_{2\ell} + (2-s) \sum_{\ell=1}^{W} (\mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) - 1)\epsilon_{2\ell+1} \Big).$$

To finish up, since the matrix $z\boldsymbol{I} - \mathbf{A}$ is upper block triangular, we have

$$\det(z\boldsymbol{I} - \mathbf{A}) = (z-1) \prod_{\ell=1}^{W} \det(z\boldsymbol{I} - \tilde{\mathbf{T}}_\ell \mathrm{diag}(1-s, s-1)) = (z-1)(z^2 - (1-s)^2)^W,$$

giving the characteristic polynomial of the scheme. The property of $\Psi_2(z)$ annihilating for the first row of $\mathbf{E}$ can be checked analogously to [Bellotti et al., 2022]. Observe that $\Psi_2(z)$ could also be found solving (55) by hand. □

### 6.3.3 Modified equations under acoustic scaling

The discussion of Section 6.3.2 is fully discrete. Now we come back to the asymptotic analysis using modified equations of Section 4 and considering local initialisations, *i.e.* $\mathbf{w} \in \mathbb{R}^q$.

**Proposition 12** (Modified equations)**.** *Under acoustic scaling, that is, when $\lambda > 0$ is fixed as $\Delta x \to 0$, the modified equation for the bulk Finite Difference scheme* (58) *where the lattice Boltzmann scheme is determined by* (63), (64) *and* (65) *is*

$$\partial_t \phi(t,\boldsymbol{x}) + \lambda \sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}} \phi(t,\boldsymbol{x}) \\ - \lambda \Delta x \Big(\frac{1}{s} - \frac{1}{2}\Big)\Big(2 \sum_{\ell=1}^{W} \epsilon_{2\ell+1} \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}} - \Big(\sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}}\Big)^2\Big)\phi(t,\boldsymbol{x}) = O(\Delta x^2),$$

*for $(t,\boldsymbol{x}) \in \mathbb{R}_+ \times \mathbb{R}^d$. Under the assumption of local initialisation $\mathbf{w} \in \mathbb{R}^q$ fulfilling Corollary 1, thus having $\mathsf{w}_1 = 1$ and $\mathsf{w}_{2\ell} = \epsilon_{2\ell}$ for $\ell \in [\![1, W]\!]$, the modified equation for the unique initialisation scheme ((57) with $n = 1$) is, for $\boldsymbol{x} \in \mathbb{R}^d$*

$$\partial_t \phi(0,\boldsymbol{x}) + \lambda \sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}} \phi(0,\boldsymbol{x}) \\ - \frac{\lambda \Delta x}{2}\Big(2 \sum_{\ell=1}^{W} \big((2-s)\epsilon_{2\ell+1} + (s-1)\mathsf{w}_{2\ell+1}\big) \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}} - \Big(\sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}}\Big)^2\Big)\phi(0,\boldsymbol{x}) = O(\Delta x^2).$$

*Proof.* We have

$$\mathsf{A}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) \asymp -\Delta x \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}} + O(\Delta x^3), \qquad \mathsf{S}(\mathbf{x}^{\boldsymbol{c}_{2\ell}}) \asymp 1 + \Delta x^2 \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}} + O(\Delta x^4).$$

It can be easily checked that the modified equation of the bulk Finite Difference scheme reads as in the claim. For the initialisation scheme, only the computation of $\boldsymbol{\mathcal{E}}^{(0)}$, $\boldsymbol{\mathcal{E}}^{(1)}$ and $\boldsymbol{\mathcal{E}}^{(2)}$ is needed:

$$\boldsymbol{\mathcal{E}}_{1j}^{(0)} = \delta_{1j}, \quad \boldsymbol{\mathcal{E}}_{1,\cdot}^{(1)} = \Big[-s \sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}}, (s-1) \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}}, 0, \dots, (s-1) \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2W}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}}, 0\Big], \\ \boldsymbol{\mathcal{E}}_{1,\cdot}^{(2)} = \Big[(2-s) \sum_{\ell=1}^{W} \epsilon_{2\ell+1} \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}}, 0, (s-1) \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}}, \dots, 0, (s-1) \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2W}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}}\Big].$$

Using the assumptions on the choice of initialisation, the modified equation for the initialisation scheme, which reads for $\boldsymbol{x} \in \mathbb{R}^d$

$$\partial_t \phi(0,\boldsymbol{x}) + \lambda \sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}} \phi(0,\boldsymbol{x}) \\ - \frac{\lambda \Delta x}{2}\Big(2 \sum_{\ell=1}^{W} \big((2-s)\epsilon_{2\ell+1} + (s-1)\mathsf{w}_{2\ell+1}\big) \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}} - \Big(\sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}}\Big)^2\Big)\phi(0,\boldsymbol{x}) = O(\Delta x^2).$$

depends on the choice of initialisation $\mathsf{w}_{2\ell+1}$ of the odd moments, which still need to be fixed. □

Enforcing the equality between the dissipation coefficients of the initialisation scheme and the bulk Finite Difference scheme according to Proposition 12 provides the differential constraint

$$\sum_{\ell=1}^{W} \mathsf{w}_{2\ell+1} \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}} = \frac{1}{s}\Big(\Big(\sum_{\ell=1}^{W} \epsilon_{2\ell} \sum_{|\mathbf{n}|=1} \boldsymbol{c}_{2\ell}^{\mathbf{n}} \partial_{\boldsymbol{x}}^{\mathbf{n}}\Big)^2 + (s-2) \sum_{\ell=1}^{W} \epsilon_{2\ell+1} \sum_{|\mathbf{n}|=2} \frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!} \partial_{\boldsymbol{x}}^{\mathbf{n}}\Big).$$

We now provide some examples where this differential constraint can or cannot be fulfilled.

*Example* 6.

- $\mathrm{D_1Q_3}$, having $d = 1$, $W = 1$ and $c_2 = 1$. After simplifying the second-order derivative operator, the condition reads $\mathsf{w}_3 = (2\epsilon_2^2 + (s-2)\epsilon_3)/s$, which has to be compared with (73).
- $\mathrm{D_2Q_5}$, having $d = 2$, $W = 2$, $\boldsymbol{c}_2 = [1,0]^{\mathsf{t}}$ and $\boldsymbol{c}_4 = [0,1]^{\mathsf{t}}$, we obtain

  $$\mathsf{w}_3 \partial_{x_1 x_1} + \mathsf{w}_5 \partial_{x_2 x_2} = \frac{1}{s}\big((2\epsilon_2^2 + (s-2)\epsilon_3)\partial_{x_1 x_1} + 4\epsilon_2\epsilon_4 \partial_{x_1 x_2} + (2\epsilon_4^2 + (s-2)\epsilon_5)\partial_{x_2 x_2}\big),$$

  which cannot be fulfilled—except when either $\epsilon_2$ or $\epsilon_4$ are zero rendering an essentially 1d problem—due to the presence of the mixed term in $\partial_{x_1 x_2}$ on the right hand side, arising from the hyperbolic part. In order to deal with this term, one is compelled to consider a richer scheme with diagonal discrete velocities, such as the $\mathrm{D_2Q_9}$ scheme.

- $\mathrm{D_2Q_9}$, having $d=2$, $W=4$, $\boldsymbol{c}_2=[1,0]^\mathsf{t}$, $\boldsymbol{c}_4=[0,1]^\mathsf{t}$, $\boldsymbol{c}_6=[1,1]^\mathsf{t}$ and $\boldsymbol{c}_8=[-1,1]^\mathsf{t}$, we obtain

$$\begin{aligned}
&\frac{1}{2}(\mathsf{w}_3+\mathsf{w}_7+\mathsf{w}_9)\partial_{x_1x_1}+(\mathsf{w}_7-\mathsf{w}_9)\partial_{x_1x_2}+\frac{1}{2}(\mathsf{w}_5+\mathsf{w}_7+\mathsf{w}_9)\partial_{x_2x_2}\\
&=\frac{1}{s}\Big(\underbrace{\Big(\epsilon_2^2+\epsilon_6^2+\epsilon_8^2+\epsilon_2\epsilon_6-\epsilon_2\epsilon_8-\epsilon_6\epsilon_8+\frac{1}{2}(s-2)(\epsilon_3+\epsilon_7+\epsilon_9)\Big)}_{R_{x_1x_1}}\partial_{x_1x_1}\\
&\quad+\underbrace{\Big(2\epsilon_6^2-2\epsilon_8^2+2\epsilon_2\epsilon_4+\epsilon_2\epsilon_6+\epsilon_2\epsilon_8+\epsilon_4\epsilon_6-\epsilon_4\epsilon_8+(s-2)(\epsilon_7-\epsilon_9)\Big)}_{R_{x_1x_2}}\partial_{x_1x_2}\\
&\quad+\underbrace{\Big(\epsilon_4^2+\epsilon_6^2+\epsilon_8^2+\epsilon_4\epsilon_6+\epsilon_4\epsilon_8+\epsilon_6\epsilon_8+\frac{1}{2}(s-2)(\epsilon_5+\epsilon_7+\epsilon_9)\Big)}_{R_{x_2x_2}}\partial_{x_2x_2}\Big).
\end{aligned}$$

  This system is under-determined, thus it has several solutions. For example, picking $\mathsf{w}_9=0$, we necessarily enforce $\mathsf{w}_7=R_{x_1x_2}/s$ and then we have that $\mathsf{w}_3=(2R_{x_1x_1}-R_{x_1x_2})/s$ and $\mathsf{w}_5=(2R_{x_2x_2}-R_{x_1x_2})/s$.

### 6.3.4 Modified equations under diffusive scaling

As mentioned in the Introduction, the literature also features lattice Boltzmann schemes used under diffusive scaling [Zhao and Yong, 2017] between time and space discretisations. We therefore finally consider this scaling where $\Delta t\propto\Delta x^2$ as $\Delta x\to 0$, allowing to approximate the solution of

$$\begin{cases}\partial_t u(t,\boldsymbol{x})+\boldsymbol{V}\cdot\nabla_{\boldsymbol{x}}u(t,\boldsymbol{x})-\nabla_{\boldsymbol{x}}\cdot(\boldsymbol{D}\nabla_{\boldsymbol{x}}u)(t,\boldsymbol{x})=0, & (t,\boldsymbol{x})\in\mathbb{R}_+\times\mathbb{R}^d, \quad (69)\\ u(0,\boldsymbol{x})=u^\circ(\boldsymbol{x}), & \boldsymbol{x}\in\mathbb{R}^d, \quad (70)\end{cases}$$

where the diffusion matrix is $\boldsymbol{D}\in\mathcal{M}_d(\mathbb{R})$. This scaling is difficult to treat in full generality because it requires a consistency study up to order $O(\Delta t)=O(\Delta x^2)$ included. Still, as previously highlighted, the unobservable framework of the current Section 6 allows us to circumvent these difficulties. The assumptions are slightly different than the rest of the paper.

**Theorem 3** ([Bellotti, 2023] Modified equation of the bulk scheme)**.** *Under diffusive scaling, that is, when $\lambda=\mu/\Delta x$ with $\mu>0$ fixed as $\Delta x\to 0$, assuming that $\epsilon_{2\ell}=\Delta x\tilde{\epsilon}_{2\ell}$ where $\tilde{\epsilon}_{2\ell}$ and $\epsilon_{2\ell+1}$ are fixed as $\Delta x\to 0$ for $\ell\in[\![1,W]\!]$, the modified equation for the bulk Finite Difference scheme* (58) *where the lattice Boltzmann scheme is determined by* (63), (64) *and* (65) *is*

$$\partial_t\phi(t,\boldsymbol{x})+\mu\sum_{\ell=1}^{W}\tilde{\epsilon}_{2\ell}\sum_{|\mathbf{n}|=1}\boldsymbol{c}_{2\ell}^{\mathbf{n}}\partial_{\boldsymbol{x}}^{\mathbf{n}}\phi(t,\boldsymbol{x})-2\mu\Big(\frac{1}{s}-\frac{1}{2}\Big)\sum_{\ell=1}^{W}\epsilon_{2\ell+1}\sum_{|\mathbf{n}|=2}\frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!}\partial_{\boldsymbol{x}}^{\mathbf{n}}\phi(t,\boldsymbol{x})=O(\Delta x^2),$$

*for $(t,\boldsymbol{x})\in\mathbb{R}_+\times\mathbb{R}^d$.*

However, we observe that since $\Delta t\propto\Delta x^2$, the second-order consistency of the bulk scheme is preserved even when the initialisation schemes are not consistent, provided that $\mathsf{w}_1=1$. This is radically different from the acoustic scaling $\Delta t\propto\Delta x$ and comes from the fact that the errors coming from the initialisation routine are now of order $O(\Delta t)=O(\Delta x^2)$. Hence, under diffusive scaling, enforcing that the initialisation schemes are consistent is merely a question of obtaining time smoothness of the numerical solution.

**Proposition 13.** *Under diffusive scaling, that is, when $\lambda=\mu/\Delta x$ with $\mu>0$ fixed as $\Delta x\to 0$, assuming that $\epsilon_{2\ell}=\Delta x\tilde{\epsilon}_{2\ell}$ where $\tilde{\epsilon}_{2\ell}$ and $\epsilon_{2\ell+1}$ are fixed as $\Delta x\to 0$ for $\ell\in[\![1,W]\!]$, the modified equation of the unique initialisation scheme for the lattice Boltzmann scheme determined by* (63), (64) *and* (65)*— considering a local initialisation $\mathbf{w}\in\mathbb{R}^q$ with $\mathsf{w}_1=1$—is*

$$\begin{aligned}
&\partial_t\phi(0,\boldsymbol{x})+\mu\sum_{\ell=1}^{W}(s\tilde{\epsilon}_{2\ell}+(1-s)\tilde{\mathsf{w}}_{2\ell})\sum_{|\mathbf{n}|=1}\boldsymbol{c}_{2\ell}^{\mathbf{n}}\partial_{\boldsymbol{x}}^{\mathbf{n}}\phi(0,\boldsymbol{x})\\
&\qquad-\mu\sum_{\ell=1}^{W}((2-s)\epsilon_{2\ell+1}+(s-1)\mathsf{w}_{2\ell+1})\sum_{|\mathbf{n}|=2}\frac{\boldsymbol{c}_{2\ell}^{\mathbf{n}}}{\mathbf{n}!}\partial_{\boldsymbol{x}}^{\mathbf{n}}\phi(0,\boldsymbol{x})=O(\Delta x),
\end{aligned}$$

*for $\boldsymbol{x}\in\mathbb{R}^d$, where $\mathsf{w}_{2\ell}=\Delta x\tilde{\mathsf{w}}_{2\ell}$ with fixed $\tilde{\mathsf{w}}_{2\ell}$ as $\Delta x\to 0$ for $\ell\in[\![1,W]\!]$.*

Therefore, the initialisation scheme is consistent with the bulk scheme under the conditions

$$\tilde{\mathsf{w}}_{2\ell}=\tilde{\epsilon}_{2\ell},\qquad \mathsf{w}_{2\ell+1}=\frac{s-2}{s}\epsilon_{2\ell+1},\qquad \ell\in[\![1,W]\!],$$

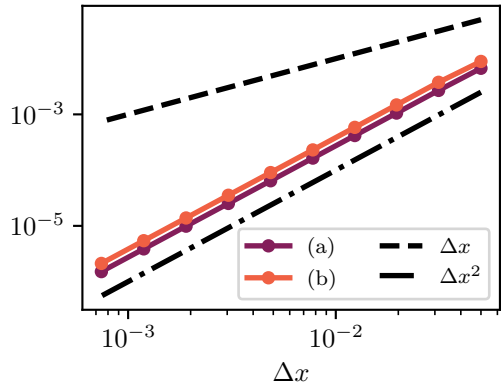


Figure 10: $L^2$ errors at the final time for the initialisations (a) and (b). We observe second-order convergence irrespective of the consistency of the initialisation scheme.

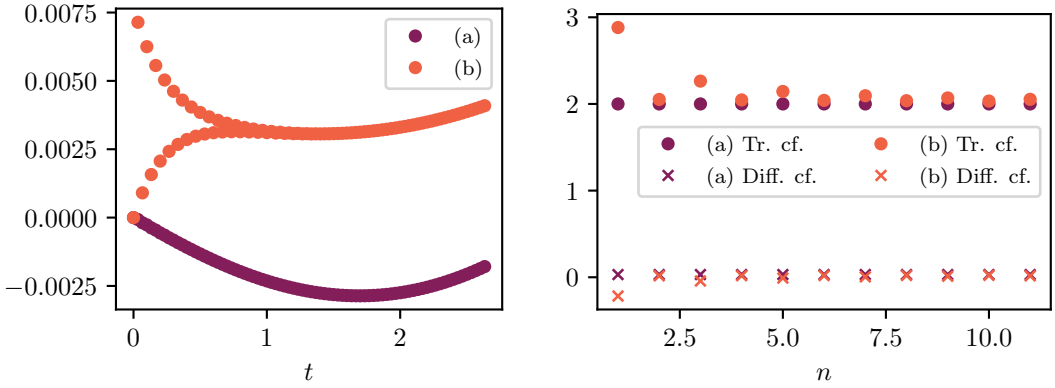


Figure 11: Left: test for smoothness in time close to $t = 0$ for the initialisations (a) and (b): difference between exact and numerical solution at the eighth lattice point. The first one gives a smooth profile whereas the second one oscillates with damping. Right: transport and diffusion coefficient in the modified equations for different $n$.

which are only set to ensure—as previously stated—time smoothness.

To numerically verify the previous claims, we consider the $D_1Q_3$ introduced in Example 6. Considering the bounded domain $[0,1]$ with periodic boundary conditions, using $u^\circ(x) = \cos(2\pi x)$ renders the exact solution $u(t,x) = e^{-4\pi^2 Dt}\cos(2\pi(x - Vt))$ to (69)/(70). We utilize $\mu = 1$, $V = 2$ and $D = 1/32$. These physical constant are set taking $\tilde{\epsilon}_2 = V$, $\epsilon_3 = 1$ and $s = 1/(D + 1/2)$. We consider two kinds of initialisations, which are

$$\text{(a)} \qquad w_1 = 1, \quad w_2 = \Delta x \tilde{\epsilon}_2, \quad w_3 = \frac{s-2}{s}\epsilon_3,$$

$$\text{(b)} \qquad w_1 = 1, \quad w_2 = \frac{\Delta x \tilde{\epsilon}_2}{2}, \quad w_3 = 10\frac{s-2}{s}\epsilon_3,$$

with the first condition (a) yielding a consistent initialisation scheme and the second condition (b) giving an inconsistent one.

**Study of the convergence order** We simulate until the final time 0.05 and measure the $L^2$ errors progressively decreasing the space step $\Delta x$. The results are given in Figure 10, confirming that, regardless of the consistency of the initialisation scheme, the overall method is second-order convergent, since $\Delta t \propto \Delta x^2$. As expected, the error constant is slightly better when the initialisation scheme is consistent.

**Study of the time smoothness of the numerical solution** We consider a framework analogous to the one of Section 5.1.2 with $\Delta x = 1/30$ and the previously introduced parameters, measuring the discrepancy between numerical and

exact solution at the eighth point of the lattice. The results in Figure 11 confirm the previous theoretical discussion: considering consistent initialisation schemes allows to avoid initial oscillating boundary layers in the simulation. Furthermore, we see that for the even steps, the transport and diffusion coefficients from the modified equations of the starting schemes are always closer to the one in the bulk than the ones for the odd steps, explaining why the discrepancies in terms of error with respect to the exact solution are smaller for even steps than for odd steps.

### 6.4 Conclusions

In Section 6, we have defined a notion of observability for lattice Boltzmann schemes, allowing to identify the schemes with very little initialisation schemes as those being strongly unobservable. In control theory, it is well known that unobservable systems can be represented by other systems which order has been reduced removing unobservable modes. It is therefore easy to analyze the initialisation phase of these schemes with the technique of the modified equation. In particular, we have found that a well-known and vast class of lattice Boltzmann schemes, namely the so-called link two-relaxation-times schemes with magic parameters equal to one-fourth, fits this framework. We have exploited this fact in order to provide the constraints on the initial data for having a smooth initialisation, both under acoustic and diffusive scaling.

## 7 General conclusions and perspectives

Due to the fact that lattice Boltzmann schemes feature more unknowns than variables of interest, their initialisation—especially for the non-conserved moments—can have an important impact on the outcomes of the simulations. This is a side effect due to the fact that the discrete scheme supports parasitic—or spurious—modes. The aim of the present contribution was indeed to study the role of the initialisation on the numerical behaviour of general lattice Boltzmann schemes. To this end, we have introduced a modified equation analysis which has ensured to propose initialisations yielding consistent starting schemes, which is crucial to preserve the second-order accuracy of many methods. The modified equation has also allowed to precisely describe the behaviour of the lattice Boltzmann schemes close to the beginning of the numerical simulation—where the different dynamics were essentially driven by numerical dissipation—and identify initialisations yielding smooth discrete solution without oscillatory initial layers. Finally, we have introduced a notion of observability for lattice Boltzmann schemes which has allowed to characterize schemes with a small number of initialisation schemes. This feature makes the study of the initialisation for these schemes way more accessible than for general ones. Consistent and smooth initialisations have been a hot topic in the lattice Boltzmann community for quite a long time [Van Leemput et al., 2009, Caiazzo, 2005, Junk and Yang, 2015, Huang et al., 2015]. However, to the best of our knowledge, no general approach to the analysis of these features were available. They are important in order to ensure that the order of the schemes is preserved. From another perspective, although the novel notion of observability for lattice Boltzmann schemes has been exploited solely to study the number of needed initialisation schemes, we do believe that it can be useful to investigate other features of these schemes. For example, one interesting topic would be the one linked to "realisation" [Brewer et al., 1986, Chapter 4] and "minimal realisations" [De Schutter, 2000]: given a target Finite Difference scheme (*i.e.* a transfer function), how can we construct the smallest lattice Boltzmann scheme of which it is the corresponding Finite Difference scheme. This will be the object of future investigations.

## Acknowledgments

The author thanks his PhD advisors B. Graille and M. Massot for the useful advice and L. François for having hinted the importance of dealing with simple eigenvalues of modulus one in the study of initial conditions.

## Funding

The author is supported by a PhD funding (year 2019) from Ecole polytechnique.

## References

[Ascher and Petzold, 1998] Ascher, U. M. and Petzold, L. R. (1998). *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM.

[Åström and Murray, 2008] Åström, K. J. and Murray, R. M. (2008). *Feedback systems: an introduction for scientists and engineers*. Princeton University Press.

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

[Bellotti et al., 2022] Bellotti, T., Graille, B., and Massot, M. (2022). Finite difference formulation of any lattice Boltzmann scheme. *Numerische Mathematik*, 152:1–40.

[Bender et al., 1999] Bender, C. M., Orszag, S., and Orszag, S. A. (1999). *Advanced mathematical methods for scientists and engineers I: Asymptotic methods and perturbation theory*, volume 1. Springer Science & Business Media.

[Blyth, 2018] Blyth, T. S. (2018). *Module theory: an approach to linear algebra*. University of St Andrews.

[Brewer et al., 1986] Brewer, J. W., Bunce, J. W., and Van Vleck, F. S. (1986). *Linear systems over commutative rings*. CRC Press.

[Caetano et al., 2019] Caetano, F., Dubois, F., and Graille, B. (2019). A result of convergence for a mono-dimensional two-velocities lattice Boltzmann scheme. *arXiv preprint arXiv:1905.12393*.

[Caiazzo, 2005] Caiazzo, A. (2005). Analysis of lattice Boltzmann initialization routines. *Journal of Statistical Physics*, 121(1):37–48.

[Carpentier et al., 1997] Carpentier, R., de La Bourdonnaye, A., and Larrouturou, B. (1997). On the derivation of the modified equation for the analysis of linear numerical methods. *ESAIM: Mathematical Modelling and Numerical Analysis*, 31(4):459–470.

[Chai and Shi, 2020] Chai, Z. and Shi, B. (2020). Multiple-relaxation-time lattice Boltzmann method for the Navier-Stokes and nonlinear convection-diffusion equations: Modeling, analysis, and elements. *Physical Review E*, 102(2):023306.

[Cheng and Lu, 1999] Cheng, S. S. and Lu, Y.-F. (1999). General solutions of a three-level partial difference equation. *Computers & Mathematics with Applications*, 38(7-8):65–79.

[De Schutter, 2000] De Schutter, B. (2000). Minimal state-space realization in linear system theory: an overview. *Journal of Computational and Applied Mathematics*, 121(1-2):331–354.

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

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

[D'Humières and Ginzburg, 2009] D'Humières, D. and Ginzburg, I. (2009). Viscosity independent numerical errors for lattice boltzmann models: From recurrence equations to "magic" collision numbers. *Computers & Mathematics with Applications*, 58(5):823–840.

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

[Dubois, 2021] Dubois, F. (2021). Nonlinear fourth order Taylor expansion of lattice Boltzmann schemes. *Asymptotic Analysis*, (1 Jan. 2021):1–41.

[Dubois et al., 2020] Dubois, F., Graille, B., and Rao, S. R. (2020). A notion of non-negativity preserving relaxation for a mono-dimensional three velocities scheme with relative velocity. *Journal of Computational Science*, 47:101181.

[Fliess and Mounier, 1998] Fliess, M. and Mounier, H. (1998). Controllability and observability of linear delay systems: an algebraic approach. *ESAIM: Control, Optimisation and Calculus of Variations*, 3:301–314.

[Fučík and Straka, 2021] Fučík, R. and Straka, R. (2021). Equivalent finite difference and partial differential equations for the lattice Boltzmann method. *Computers & Mathematics with Applications*, 90:96–103.

[Ginzburg et al., 2008] Ginzburg, I., Verhaeghe, F., and d'Humières, D. (2008). Two-relaxation-time lattice Boltzmann scheme: About parametrization, velocity, pressure and mixed boundary conditions. *Communications in Computational Physics*, 3(2):427–478.

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

[Gustafsson et al., 1995] Gustafsson, B., Kreiss, H.-O., and Oliger, J. (1995). *Time dependent problems and difference methods*, volume 24. John Wiley & Sons.

[Hairer et al., 2008] Hairer, E., Nørsett, S. P., and Wanner, G. (2008). *Solving Ordinary Differential Equations I*, volume 8. Springer. Second Revised Edition.

[Hendricks et al., 2008] Hendricks, E., Jannerup, O., and Sørensen, P. H. (2008). *Linear systems control: deterministic and stochastic methods*. Springer.

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

[Huang et al., 2015] Huang, J., Wu, H., and Yong, W.-A. (2015). On initial conditions for the lattice Boltzmann method. *Communications in Computational Physics*, 18(2):450–468.

[Hundsdorfer and Ruuth, 2006] Hundsdorfer, W. and Ruuth, S. (2006). On monotonicity and boundedness properties of linear multistep methods. *Mathematics of Computation*, 75(254):655–672.

[Hundsdorfer et al., 2003] Hundsdorfer, W., Ruuth, S. J., and Spiteri, R. J. (2003). Monotonicity-preserving linear multistep methods. *SIAM Journal on Numerical Analysis*, 41(2):605–623.

[Jin and Xin, 1995] Jin, S. and Xin, Z. (1995). The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Communications on Pure and Applied Mathematics*, 48(3):235–276.

[Junk et al., 2005] Junk, M., Klar, A., and Luo, L.-S. (2005). Asymptotic analysis of the lattice Boltzmann equation. *Journal of Computational Physics*, 210(2):676–704.

[Junk and Yang, 2015] Junk, M. and Yang, Z. (2015). L2 convergence of the lattice Boltzmann method for one dimensional convection-diffusion-reaction equations. *Communications in Computational Physics*, 17(5):1225–1245.

[Kuzmin et al., 2011] Kuzmin, A., Ginzburg, I., and Mohamad, A. (2011). The role of the kinetic parameter in the stability of two-relaxation-time advection–diffusion lattice Boltzmann schemes. *Computers & Mathematics with Applications*, 61(12):3417–3442.

[Kuznik et al., 2013] Kuznik, F., Luo, L.-S., and Krafczyk, M. (2013). Mesoscopic Methods in Engineering and Science. *Computers & Mathematics with Applications*, 65(6):813–814.

[Lallemand and Luo, 2000] Lallemand, P. and Luo, L.-S. (2000). Theory of the lattice Boltzmann method: Dispersion, dissipation, isotropy, Galilean invariance, and stability. *Physical Review E*, 61(6):6546.

[Miller, 1971] Miller, J. J. (1971). On the location of zeros of certain classes of polynomials with applications to numerical analysis. *IMA Journal of Applied Mathematics*, 8(3):397–406.

[O'Malley, 1991] O'Malley, R. E. (1991). *Singular perturbation methods for ordinary differential equations*, volume 89. Springer.

[Saad, 1989] Saad, Y. (1989). Overview of Krylov subspace methods with applications to control problems. Technical report.

[Simonis et al., 2020] Simonis, S., Frank, M., and Krause, M. J. (2020). On relaxation systems and their relation to discrete velocity Boltzmann models for scalar advection–diffusion equations. *Philosophical Transactions of the Royal Society A*, 378(2175):20190400.

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

[Suga, 2010] Suga, S. (2010). An accurate multi-level finite difference scheme for 1D diffusion equations derived from the lattice Boltzmann method. *Journal of Statistical Physics*, 140(3):494–503.

[Trefethen, 1996] Trefethen, L. N. (1996). *Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations*. unpublished text.

[Van Leemput et al., 2009] Van Leemput, P., Rheinländer, M., and Junk, M. (2009). Smooth initialization of lattice Boltzmann schemes. *Computers & Mathematics with Applications*, 58(5):867–882.

[Warming and Hyett, 1974] Warming, R. F. and Hyett, B. (1974). The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of Computational Physics*, 14(2):159–179.

[Yong et al., 2016] Yong, W.-A., Zhao, W., Luo, L.-S., et al. (2016). Theory of the lattice Boltzmann method: Derivation of macroscopic equations via the Maxwell iteration. *Physical Review E*, 93(3):033310.

[Zhang et al., 2019] Zhang, M., Zhao, W., and Lin, P. (2019). Lattice Boltzmann method for general convection-diffusion equations: MRT model and boundary schemes. *Journal of Computational Physics*, 389:147–163.

[Zhao and Yong, 2017] Zhao, W. and Yong, W.-A. (2017). Maxwell iteration for the lattice Boltzmann method with diffusive scaling. *Physical Review E*, 95(3):033311.

# A Stability of the $D_1Q_2$ scheme of Section 5.1

There are several ways of checking the roots of amplification polynomial of the corresponding bulk scheme for Section 5.1. In this case, we can proceed directly by solving the characteristic equation or by using the procedure by [Graille, 2014]. Thanks to its generality, we here present the computations using the technique by Miller [Miller, 1971, Strikwerda, 2004]. The amplification polynomial reads $\hat{\Phi}_2(z,\xi\Delta x) = z^2+((s_2-2)\cos(\xi\Delta x)+is_2\epsilon_2\sin(\xi\Delta x))z+(1-s_2)$, where $\xi\in[-\pi/\Delta x,\pi/\Delta x]$. We have that

$$\hat{\Phi}_2^\star(z,\xi\Delta x) := z^2\hat{\Phi}_2(z^{-1},-\xi\Delta x) = (1-s_2)z^2+((s_2-2)\cos(\xi\Delta x)-is_2\epsilon_2\sin(\xi\Delta x))z+1.$$

- Let $s_2\in]0,2[$. A first condition to bound the roots of $\hat{\Phi}_2(z,\xi\Delta x)$ in modulus by one regardless of the frequency is that $|\hat{\Phi}_2(0,\xi\Delta x)|<|\hat{\Phi}_2^\star(0,\xi\Delta x)|$, which yields the condition $0<s_2<2$. Then we compute $\hat{\Phi}_1(z,\xi\Delta x)$ as

  $$\hat{\Phi}_1(z,\xi\Delta x) := z^{-1}(\hat{\Phi}_2^\star(0,\xi\Delta x)\hat{\Phi}_2(z,\xi\Delta x)-\hat{\Phi}_2(0,\xi\Delta x)\hat{\Phi}_2^\star(z,\xi\Delta x)) = s_2(2-s_2)(z-\cos(\xi\Delta x)+i\epsilon_2\sin(\xi\Delta x)).$$

  The final condition to check is that the root of $\hat{\Phi}_1(z,\xi\Delta x)$ is bounded by one in modulus for any frequency. This is $\cos^2(\xi\Delta x)+\epsilon_2^2\sin^2(\xi\Delta x) = 1+(\epsilon_2^2-1)\sin^2(\xi\Delta x)\le 1$ taking place for any $\xi\in[-\pi/\Delta x,\pi/\Delta x]$ if and only if $\epsilon_2^2\le 1$.

- Let $s_2=2$. In this case $\hat{\Phi}_1\equiv 0$. We then have to use the second condition from [Strikwerda, 2004, Theorem 4.3.2], hence we check

  $$\frac{d\hat{\Phi}_2(z,\xi\Delta x)}{dz} = 2z+2i\epsilon_2\sin(\xi\Delta x),$$

  which unique root should be strictly in the unit circle for any frequency $\xi\in[-\pi/\Delta x,\pi/\Delta x]$. This is achieved by $|\epsilon_2|<1$.

This achieves the proof of the stability conditions (35).

# B Derivation of the forward centered initialisation schemes for the $D_1Q_2$ of Section 5.1

We can first unsuccessfully attempt to obtain a forward centered scheme as initialisation scheme, using a local initialisation of the conserved moment, that is $w_1=1$ and prepared initialisation of the non-conserved one, thus $w_2\in\mathbb{R}[x_1,x_1^{-1},\dots,x_d,x_d^{-1}]$. Using the notation (21), this corresponds to find a compactly supported solution of the following infinite system

$$\dots,\quad w_{2,1}-w_{2,3}=0,\quad w_{2,0}-w_{2,2}=-\frac{1-\epsilon_2+s_2\epsilon_2}{1-s_2},\quad w_{2,-1}-w_{2,1}=\frac{2}{1-s_2},$$
$$w_{2,-2}-w_{2,0}=-\frac{1+\epsilon_2-s_2\epsilon_2}{1-s_2},\quad w_{2,-3}-w_{2,-1}=0,\quad \dots$$

This problem cannot be solved by a compactly supported sequence, in particular, because of the median term. This would go back to perform a deconvolution in the ring of Finite Difference operators, which is not solvable because the operator $A(x_1)$ is not invertible in such ring. If we consider to work on a bounded domain with $N_x\in\mathbb{N}^*$ points and endow the shift operators with periodic boundary conditions [Van Leemput et al., 2009], some of these deconvolution problems become solvable at the price of dealing with non-compactly supported solutions, *i.e.* stemming from a full inverse of a sparse matrix. The previous problem can be seen as the one of inverting a circulant matrix, which eigenvalues are $\sigma_\ell = \exp(2\pi i(N_x-1)\ell/N_x)-\exp(2\pi i\ell/N_x)$ for $\ell\in[\![0,N_x[\![$. Since $\sigma_0=0$, the circulant matrix is not invertible. Therefore, even in the periodic setting, this procedure does not work. This can be interpreted—if we see the equilibria as a control on the system—as due to the lack of "reachability" of the system at hand, *cf.* [Brewer et al., 1986, Chapter 2]. Since the term $A(x_1)$ is not a unit, which causes the lack of reachability, it cannot be compensated by its inverse contained in the equilibrium to generate the desired initialisation scheme. This is why we are compelled to consider $w_1\in\mathbb{R}[x_1,x_1^{-1},\dots,x_d,x_d^{-1}]$ to obtain the requested forward centered scheme.

Considering a prepared initialisation for both moments, thus $w_1,w_2\in\mathbb{R}[x_1,x_1^{-1},\dots,x_d,x_d^{-1}]$, several choices are possible to recover this scheme. The infinite system to solve reads

…

$$
\begin{aligned}
\frac{1+s_2\epsilon_2}{2}w_{1,1}+\frac{1-s_2}{2}w_{2,1}+\frac{1-s_2\epsilon_2}{2}w_{1,3}-\frac{1-s_2}{2}w_{2,3}&=0,\\
\frac{1+s_2\epsilon_2}{2}w_{1,0}+\frac{1-s_2}{2}w_{2,0}+\frac{1-s_2\epsilon_2}{2}w_{1,2}-\frac{1-s_2}{2}w_{2,2}&=\frac{\epsilon_2}{2},\\
\frac{1+s_2\epsilon_2}{2}w_{1,-1}+\frac{1-s_2}{2}w_{2,-1}+\frac{1-s_2\epsilon_2}{2}w_{1,1}-\frac{1-s_2}{2}w_{2,1}&=1,\\
\frac{1+s_2\epsilon_2}{2}w_{1,-2}+\frac{1-s_2}{2}w_{2,-2}+\frac{1-s_2\epsilon_2}{2}w_{1,0}-\frac{1-s_2}{2}w_{2,0}&=-\frac{\epsilon_2}{2},\\
\frac{1+s_2\epsilon_2}{2}w_{1,-3}+\frac{1-s_2}{2}w_{2,-3}+\frac{1-s_2\epsilon_2}{2}w_{1,-1}-\frac{1-s_2}{2}w_{2,-1}&=0,\\
&\dots
\end{aligned}
$$

In order to construct a (non-unique) solution, we first enforce the compactness: $w_{1,\ell}=w_{2,\ell}=0$ for $|\ell|\geq 2$. From this, we obtain the finite system

$$
\begin{aligned}
(1+s_2\epsilon_2)w_{1,1}+(1-s_2)w_{2,1}&=0,\\
(1+s_2\epsilon_2)w_{1,0}+(1-s_2)w_{2,0}&=\epsilon_2,\\
(1+s_2\epsilon_2)w_{1,-1}+(1-s_2)w_{2,-1}+(1-s_2\epsilon_2)w_{1,1}-(1-s_2)w_{2,1}&=2,\\
(1-s_2\epsilon_2)w_{1,0}-(1-s_2)w_{2,0}&=-\epsilon_2,\\
(1-s_2\epsilon_2)w_{1,-1}-(1-s_2)w_{2,-1}&=0.
\end{aligned}
$$

We then split the central equation using a parameter $\theta\in\mathbb{R}$, having $(1+s_2\epsilon_2)w_{1,-1}+(1-s_2)w_{2,-1}=\theta$ and $(1-s_2\epsilon_2)w_{1,1}-(1-s_2)w_{2,1}=2-\theta$. Introducing the matrix

$$
\boldsymbol{A}=\begin{bmatrix}1+s_2\epsilon_2 & 1-s_2\\ 1-s_2\epsilon_2 & s_2-1\end{bmatrix},
$$

we solve the systems $\boldsymbol{A}[w_{1,1},w_{2,1}]^{\mathsf{t}}=[0,2-\theta]^{\mathsf{t}}$, $\boldsymbol{A}[w_{1,0},w_{2,0}]^{\mathsf{t}}=[\epsilon_2,-\epsilon_2]^{\mathsf{t}}$ and $\boldsymbol{A}[w_{1,-1},w_{2,-1}]^{\mathsf{t}}=[\theta,0]^{\mathsf{t}}$, yielding

$$
\begin{aligned}
w_{1,1}&=\frac{2-\theta}{2}, & w_{2,1}&=-\frac{(1+s_2\epsilon_2)(2-\theta)}{2(1-s_2)}, & w_{1,0}&=0, & w_{2,0}&=\frac{\epsilon_2}{1-s_2},\\
w_{1,-1}&=\frac{\theta}{2}, & w_{2,-1}&=\frac{(1-s_2\epsilon_2)\theta}{2(1-s_2)}.
\end{aligned}
$$

Unsurprisingly, these coefficients are defined for $s_2\neq 1$, since otherwise there is no initialisation scheme to devise. The only way to fulfill (22), (23) and (24) is to take $\theta=1$, giving

$$
w_{1,\pm1}=\frac{1}{2}, \qquad w_{2,\pm1}=\mp\frac{1\pm s_2\epsilon_2}{2(1-s_2)}, \qquad w_{2,0}=\frac{\epsilon_2}{1-s_2}.
$$

Allowing more non-vanishing coefficients and through a similar procedure, another possible choice to obtain the desired scheme would be

$$
w_{1,\pm2}=\pm\frac{\epsilon_2}{2}, \qquad w_{1,\pm1}=\frac{1}{2}, \qquad w_{2,\pm2}=-\frac{\epsilon_2(1\pm s_2\epsilon_2)}{2(1-s_2)}, \qquad w_{2,\pm1}=\mp\frac{1\pm s_2\epsilon_2}{2(1-s_2)}.
$$

# C Derivation of the remaining modified equations of the starting schemes for the $\mathrm{D}_1\mathrm{Q}_2$ of Section 5.1.3

In Section 5.1.3, we have left the derivation of several modified equations of the starting schemes for the $\mathrm{D}_1\mathrm{Q}_2$. We now explain how to reach them.

- **Forward centered scheme** (37). This scheme fulfills the conditions of Corollary 2, hence for $n\in\mathbb{N}^*$

$$
\partial_t\phi(0,x)+\lambda\epsilon_2\partial_x\phi(0,x)-\lambda\Delta x\Big(-\frac{n}{2}\epsilon_2^2\partial_{xx}+\frac{1}{n}\Big((\boldsymbol{\mathcal{E}}^n)_{11}^{(2)}+(\boldsymbol{\mathcal{E}}^n)_{12}^{(2)}\epsilon_2+(\boldsymbol{\mathcal{E}}^n)_{12}^{(1)}\omega_2^{(1)}+\omega_1^{(2)}\Big)\Big)\phi(0,x)=O(\Delta x^2),
$$

for $x\in\mathbb{R}$, where only the terms $\omega_2^{(1)}=1/(1-s_2)\partial_x$ and $\omega_1^{(2)}=\frac{1}{2}\partial_{xx}$ introduce discrepancies from the Lax-Friedrichs initialisation (36). Using (71) we obtain for $n\in\mathbb{N}^*$ and $x\in\mathbb{R}$

$$
\begin{aligned}
&\partial_t\phi(0,x)+\lambda\epsilon_2\partial_x\phi(0,x)\\
&\quad-\lambda\Delta x\Big(\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell\Big)(1-\epsilon_2^2)+\frac{1}{2n}\Big(1-2\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)\partial_{xx}\phi(0,x)=O(\Delta x^2).
\end{aligned}
$$

- **Forward centered scheme** (38). We have

$$\partial_t\phi(0,x)-\frac{\lambda}{n}\Big((\boldsymbol{\mathcal{E}}^n)^{(1)}_{11}+(\boldsymbol{\mathcal{E}}^n)^{(1)}_{12}\omega_2^{(0)}+\omega_1^{(1)}\Big)\phi(0,x)=O(\Delta x),\qquad n\in\mathbb{N}^*,\quad x\in\mathbb{R}.$$

where in this case $\omega_2^{(0)}=-(1+s_2)/(1-s_2)\epsilon_2$ and $\omega_1^{(1)}=-2\epsilon_2\partial_x$. Recalling that

$$(\boldsymbol{\mathcal{E}}^n)^{(1)}_{11}=-\epsilon_2\sum_{\ell=0}^{n-1}\pi_{n-\ell}(s_2)\partial_x,\qquad (\boldsymbol{\mathcal{E}}^n)^{(1)}_{12}=-\sum_{\ell=0}^{n-1}(1-s_2)^{n-\ell}\partial_x,\qquad n\in\mathbb{N}^*,\tag{71}$$

yields

$$(\boldsymbol{\mathcal{E}}^n)^{(1)}_{11}+(\boldsymbol{\mathcal{E}}^n)^{(1)}_{12}\omega_2^{(0)}+\omega_1^{(1)}=-\epsilon_2\Big(n+2\Big(1-\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)\partial_x,\qquad \in\mathbb{N}^*,$$

thus

$$\partial_t\phi(0,x)+\lambda\epsilon_2\Big(1+\frac{2}{n}\Big(1-\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)\partial_x\phi(0,x)=O(\Delta x),\qquad \in\mathbb{N}^*,\quad x\in\mathbb{R}.$$

- **Lax-Wendroff** (39). The computation is similar to the previous ones, taking into account that the only terms to change are $\omega_2^{(1)}=(1-\epsilon_2^2)/(1-s_2)\partial_x$ and $\omega_1^{(2)}=(1-\epsilon_2^2)/2\partial_{xx}$. This provides the modified equations, for $n\in\mathbb{N}^*$ and $x\in\mathbb{R}$

$$\begin{aligned}&\partial_t\phi(0,x)+\lambda\epsilon_2\partial_x\phi(0,x)\\&\quad-\lambda\Delta x\Big(\frac{1}{2}+\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell+\frac{1}{2n}\Big(1-2\sum_{\ell=0}^{n-1}(1-s_2)^\ell\Big)\Big)(1-\epsilon_2^2)\partial_{xx}\phi(0,x)=O(\Delta x^2).\end{aligned}$$

- **Smooth initialisation RE1** (40). This scheme fulfills Corollary 2 and we have for $n\in\mathbb{N}^*$ and $x\in\mathbb{R}$

$$\partial_t\phi(0,x)+\lambda\epsilon_2\partial_x\phi(0,x)-\lambda\Delta x\Big(-\frac{n}{2}\epsilon_2^2\partial_{xx}+\frac{1}{n}\Big((\boldsymbol{\mathcal{E}}^n)^{(2)}_{11}+(\boldsymbol{\mathcal{E}}^n)^{(2)}_{12}\epsilon_2+(\boldsymbol{\mathcal{E}}^n)^{(1)}_{12}\omega_2^{(1)}\Big)\Big)\phi(0,x)=O(\Delta x^2),$$

where only $\omega_2^{(1)}=-(1-\epsilon_2^2)/s_2\partial_x$ introduces differences compared to the Lax-Friedrichs initialisation (36). We therefore obtain for $n\in\mathbb{N}^*$ and $x\in\mathbb{R}$

$$\partial_t\phi(0,x)+\epsilon_2\partial_x\phi(0,x)-\lambda\Delta x\Big(\frac{1}{2}-\sum_{\ell=1}^{n-1}\Big(1-\frac{\ell}{n}\Big)(1-s_2)^\ell+\frac{1}{ns_2}\sum_{\ell=1}^{n}(1-s_2)^\ell\Big)(1-\epsilon_2^2)\partial_{xx}\phi(0,x)=O(\Delta x^2).$$

# D Conditions to obtain dissipation matching for the $D_1Q_3$ scheme of Section 5.2

Here, we present the detailed discussion of the conditions to obtain dissipation matching for the $D_1Q_3$ scheme of Section 5.2

- $s_2=1$. Then the equation is trivially satisfied for any choice of $s_3$. Enforcing the choice of $s_2=1$ in the first equation of (53) yields $(1-s_3)(w_3-\epsilon_3)=0$. This equation is trivially satisfied for $s_3=1$. If $s_3\neq 1$, then we must initialize at equilibrium, that is, consider $w_3=\epsilon_3$.
- $s_2\neq 1$. Then the equation for $s_3$ reads $(2/3-\epsilon_2^2+\epsilon_3/3)s_3=(2-s_2)(2/3-\epsilon_2^2+\epsilon_3/3)$. We distinguish two cases
  - $\epsilon_3>-2+3\epsilon_2^2$. In this case, we have to enforce

$$s_3=2-s_2.\tag{72}$$

    This is very interesting because it corresponds to the choice of "magic parameter" [D'Humières and Ginzburg, 2009, Kuzmin et al., 2011] equal to 1/4. Using this choice of $s_3$ into the first equation from (53), we obtain that $w_3$ has to be taken as

$$w_3=\frac{1}{s_2}(2(-2+3\epsilon_2^2)+(s_2-2)\epsilon_3).\tag{73}$$

    Remark that in this case, the only way of making the bulk scheme to be of second-order is to take $s_2=2$. This results in $s_3=0$, which means that one more moment is conserved by the scheme. Still, the equilibria do not depend on it. Moreover, the initialisation has to be $w_3=-2+3\epsilon_2^2\neq\epsilon_3$.
  - $\epsilon_3=-2+3\epsilon_2^2$. The equation is trivially true. Considering the first equation in (53) once more, we obtain $(1-s_3)(w_3+2-3\epsilon_2^2)=0$. If $s_3=1$, this equation is satisfied regardless of the choice of $w_3$. If $s_3\neq 1$, then the initialisation should be $w_3=-2+3\epsilon_2^2=\epsilon_3$.

# E Stability of the $D_1Q_3$ scheme of Section 5.2 when $s_2 + s_3 = 2$

We again employ the technique by Miller [Miller, 1971, Strikwerda, 2004]. We have to control the roots of $\hat{\Psi}_2(z, \xi\Delta x) = z^2 + ((s_2-2)(2\cos(\xi\Delta x)+1)/3 + (s_2-2)\epsilon_3(\cos(\xi\Delta x)-1)/3 + i s_2\epsilon_2 \sin(\xi\Delta x))z + (1-s_2)$, by computing

$$\hat{\Psi}_2^\star(z, \xi\Delta x) = (1-s_2)z^2 + ((s_2-2)(2\cos(\xi\Delta x)+1)/3 + (s_2-2)\epsilon_3(\cos(\xi\Delta x)-1)/3 - i s_2\epsilon_2 \sin(\xi\Delta x))z + 1.$$

- Let $s_2 \in ]0,2[$. Checking the first condition $|\hat{\Psi}_2(0,\xi\Delta x)| < |\hat{\Psi}_2^\star(0,\xi\Delta x)|$ trivially gives $0 < s_2 < 2$. Then we have

  $$\hat{\Psi}_1(z, \xi\Delta x) = s_2(2-s_2)(z - (2\cos(\xi\Delta x)+1)/3 - \epsilon_3(\cos(\xi\Delta x)-1)/3 + i\epsilon_2 \sin(\xi\Delta x)).$$

  Checking that the unique root of this polynomial has modulus bounded by one comes back at considering $((2\cos(\xi\Delta x)+1) + \epsilon_3(\cos(\xi\Delta x)-1))^2/9 + \epsilon_2^2\sin^2(\xi\Delta x) \le 1$. Using the trigonometric identities $\cos(\xi\Delta x) = 1 - 2\sin^2(\xi\Delta x/2)$ and $\sin^2(\xi\Delta x) = 4\sin^2(\xi\Delta x/2)(1-\sin^2(\xi\Delta x/2))$ and calling $\mu = \sin^2(\xi\Delta x/2) \in [0,1]$, we obtain

  $$\mu((\epsilon_3+2)^2/9 - \epsilon_2^2) + (-(\epsilon_3+2)/3 + \epsilon_2^2) \le 0, \qquad \forall \mu \in [0,1].$$

  This is an affine expression on $\mu$, thus the maximum is reached on the boundary of $[0,1]$. Assume, without loss of generality that $\epsilon_2 > 0$ and the standard CFL condition $\epsilon_2 \le 1$.

  - $(\epsilon_3+2)^2/9 - \epsilon_2^2 \ge 0$, corresponding to

    $$\epsilon_3 \le -2 - 3\epsilon_2, \qquad \text{or} \qquad \epsilon_3 \ge -2 + 3\epsilon_2.$$

    In this case the maximum is reached at $\mu = 1$, thus we want $(\epsilon_3+2)(\epsilon_3-1) \le 0$, hence $-2 \le \epsilon_3 \le 1$. Under the CFL condition $\epsilon_2 \le 1$ (otherwise all the computations can be adapted accordingly but no stability can be deduced), we easily find the first overall condition $-2 + 3\epsilon_2 \le \epsilon_3 \le 1$.
  - $(\epsilon_3+2)^2/9 - \epsilon_2^2 < 0$, corresponding to

    $$-2 - 3\epsilon_2 < \epsilon_3 < -2 + 3\epsilon_2.$$

    In this case the maximum is reached on $\mu = 0$, providing $-(\epsilon_3+2)/3 + \epsilon_2^2 \le 0$ thus comparing with the other conditions taking the CFL condition into account, we have $-2 + 3\epsilon_2^2 \le \epsilon_3 \le -2 + 3\epsilon_2$.

  In this case, the necessary and sufficient stability condition reads $|\epsilon_2| \le 1$ and $-2 + 3\epsilon_2^2 \le \epsilon_3 \le 1$.

- Let $s_2 = 2$. In this case $\hat{\Psi}_1 \equiv 0$, hence we compute

  $$\frac{d\hat{\Psi}_2(z, \xi\Delta x)}{dz} = 2z + 2i\epsilon_2 \sin(\xi\Delta x),$$

  hence to have its roots strictly in the unit circle for any frequency, we have the strict CFL condition $|\epsilon_2| < 1$.

# F Numerical experiments on the unobservable sub-space for Example 4

To validate the finding concerning the unobservable sub-space $\mathcal{N}$ given in Example 4, we consider two sets of initial data

$$\text{(a)} \qquad m_1(0,\cdot) = 0, \qquad m_2(0, j\Delta x) = \frac{1 + 3(-1)^j}{8},$$

$$\text{(b)} \qquad m_1(0,\cdot) = 0, \qquad m_2(0, j\Delta x) = \frac{1}{10}\exp\Big(-\frac{1}{1 - (4(j\Delta x - 0.5))^2}\Big).$$

The first datum (a) lies in $\mathcal{N}$ whereas the second one (b) does not. Observe that both data do not adhere to the guidelines to choose initial data according to the analysis in Section 4: they are uniquely selected for the current test. We shall take $j \in [\![0, N_x[\![$ in the simulations and $\Delta x = 1/N_x$. Periodic boundary conditions are enforced. The results of the simulation given in Figure 12 confirm the theory. The unobservable initial datum (a) yields zero conserved (observed) moment for any time step, whereas the observable one (b) does not, even if the conserved moment is initialized as zero everywhere. For the observable datum (b), we see that the solution converges linearly to the exact solution of the Cauchy problem, meaning the zero solution.

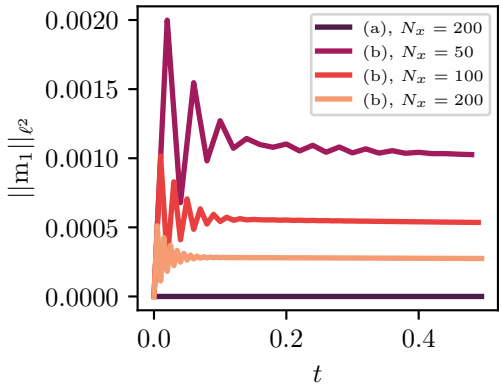


Figure 12: $L^2$ norm of the conserved moment as function of the time for the $\mathrm{D_1Q_2}$ scheme choosing $\lambda = 1$, $\epsilon_2 = 1/2$ and $s_2 = 1.8$. The test is performed for different initial data (both observable and unobservable) with different $\Delta x$.

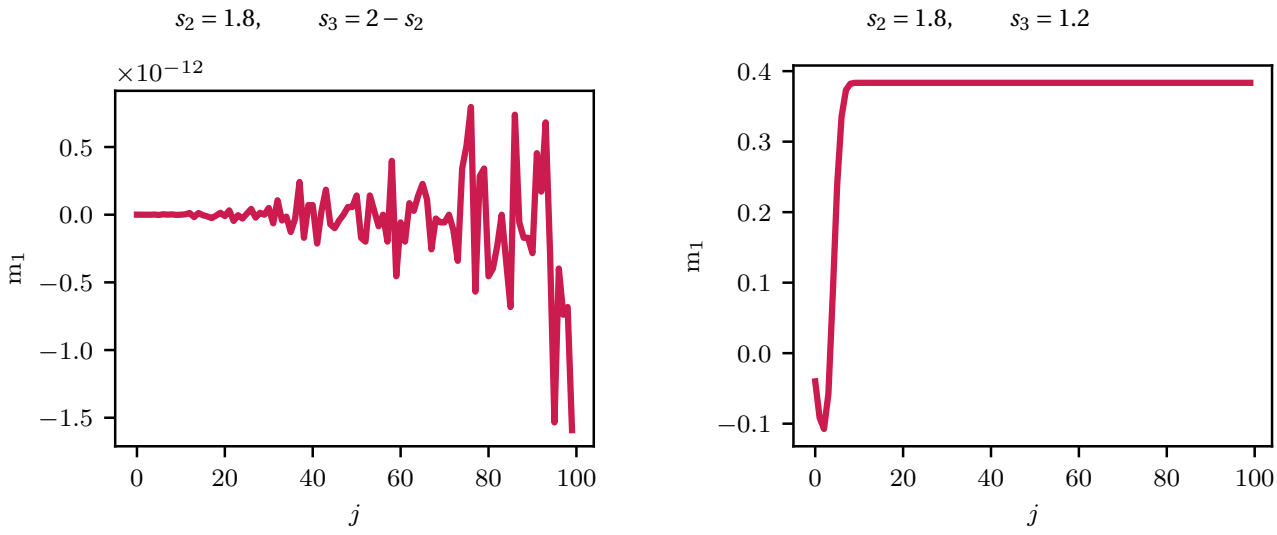


Figure 13: Conserved moment after 10 iterations for the $\mathrm{D_1Q_3}$ scheme choosing $\lambda = 1$, $\epsilon_2 = 1/2$, $\epsilon_3 = 1/10$, $N_x = 100$ and different relaxation parameters.

# G Numerical experiments on the unobservable sub-space for Example 5

To check the findings concerning $\mathcal{N}$ for the scheme in Example 5, we select

$$\mathsf{m}_1(0,\cdot) = 0, \qquad \mathsf{m}_2(0, j\Delta x) = j, \qquad \mathsf{m}_3(0, j\Delta x) = -3j^2,$$

which thus belongs to $\mathcal{N}$. We discretize with $j \in [\![0, 100[\![$ using anti-bounce-back boundary conditions $\mathsf{f}_2((n+1)\Delta t, 99\Delta x) = -\mathsf{f}_3^\star(n\Delta t, 99\Delta x)$ on the inflow and a second-order extrapolation $\mathsf{f}_3((n+1)\Delta t, 99\Delta x) = 3\mathsf{f}_3^\star(n\Delta t, 99\Delta x) - 3\mathsf{f}_3^\star(n\Delta t, 98\Delta x) + 3\mathsf{f}_3^\star(n\Delta t, 97\Delta x)$. These boundary conditions are compatible with the initial data. The result of the simulation is proposed in Figure 13. For the choice where $s_3 = 2 - s_2$, thus for which the initial datum belongs to $\mathcal{N}$, we see that the conserved moment remains zero (up to machine precision). When $s_3 \neq 2 - s_2$, thus the initial datum is observable, we remark that inside the domain, the conserved moment is non-zero (around 0.383).