The influence of parasitic modes on "weakly" unstable multi-step Finite Difference schemes

Thomas Bellotti

The influence of parasitic modes on "weakly" unstable multi-step Finite Difference schemes

Thomas Bellotti*(thomas.bellotti@math.unistra.fr)

IRMA, Université de Strasbourg, 67000 Strasbourg, France

December 21, 2023

Abstract

Numerical analysis for linear constant-coefficients Finite Difference schemes was developed approximately fifty years ago. It relies on the assumption of scheme stability and in particular—for the L^2 setting—on the absence of multiple roots of the amplification polynomial on the unit circle. This allows to decouple, while discussing the convergence of the method, the study of the consistency of the scheme from the precise knowledge of its parasitic/spurious modes, so that multi-step methods can be studied essentially as they were one-step schemes. In other words, the global truncation error can be inferred from the local truncation error. Furthermore, stability alleviates the need to delve into the complexities of floating-point arithmetic on computers, which can be challenging topics to address. In this paper, we show that in the case of "weakly" unstable schemes with multiple roots on the unit circle, although the schemes may remain stable, the consideration of parasitic modes is essential in studying their consistency and, consequently, their convergence. Otherwise said, the lack of genuine stability prevents bounding the global truncation error using the local truncation error, and one is thus compelled to study the former on its own. This research was prompted by unexpected numerical results on lattice Boltzmann schemes, which can be rewritten in terms of multi-step Finite Difference schemes. Initial expectations suggested that third-order initialization schemes would suffice to maintain the accuracy of a fourth-order multi-step scheme. However, this assumption proved incorrect for "weakly" unstable schemes. This borderline scenario underscores the significance of genuine stability in facilitating the construction of Lax-Richtmyer-like theorems and in mastering the impact of round-off errors. Despite the simplicity and apparent lack of practical usage of the linear transport equation at constant velocity considered throughout the paper, we demonstrate that high-order lattice Boltzmann schemes for this equation can be used to tackle non-linear systems of conservation laws relying on a Jin-Xin approximation and high-order splitting formulæ.

Keywords— Finite Difference; multi-step; lattice Boltzmann; weak instabilities; order of convergence; relaxation system; operator splitting

Introduction

Multi-step constant-coefficients Finite Difference schemes feature several modes [GKO95, Chapter 2]—each one associated with one root of the amplification polynomial of the scheme. For scalar problems being first-order in the time derivative, whom we shall be concerned with in this paper, one physical mode and possibly several parasitic modes are therefore mixed together in the discrete solution. As long as the scheme is stable for the L^2 norm, namely the roots are in the closed unit disk and those on the unit circle are simple, stability rules out any potential influence of the—indeed present—parasitic modes, as far as consistency is concerned. This can be understood in the following way: let $Q+1 \leq n \leq \lfloor T/\delta t \rfloor$, where $Q+1$ is the number of steps of the multi-step scheme, $T>0$ is the final time-horizon of the simulation, and δt is the time step. Denote u^n the solution of the multi-step scheme, obtained by[1] $\hat{\mathsf{u}}^n(\xi) = \hat{\mathsf{g}}^{[n]}(\xi\Delta x)\hat{\mathsf{u}}^0(\xi)$ with $|\xi\Delta x| \leq \pi$, where the amplification factors $\hat{\mathsf{g}}^{[n]}$ are determined by the multi-step scheme itself as well as by the initialization schemes. Let $\hat{\mathsf{g}}_1$ be the physical root of the amplification polynomial: the only one such that $\hat{\mathsf{g}}_1(0)=1$, and construct $\hat{\mathsf{w}}^n(\xi) = \hat{\mathsf{g}}_1(\xi\Delta x)^n\hat{\mathsf{u}}^0(\xi)$, called "pseudo-scheme". The crucial estimate to study the order of the overall method (*i.e.* taking the initialization into account) is

$$\|\mathsf{u}^n - \mathsf{w}^n\|_{\ell^2,\Delta x} \leq C_T \sum_{p=0}^{Q} \|\mathsf{u}^p - \mathsf{w}^p\|_{\ell^2,\Delta x}, \tag{1}$$

*Former affiliation: CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France.

[1] Using the hat to denote a Fourier transform.

thanks to stability. The right-hand side of (1) solely depends on $\hat{\mathsf{g}}_1$—through $\mathsf{w}^Q, \dots, \mathsf{w}^0$—and on the initializations $\mathsf{u}^Q, \dots, \mathsf{u}^0$ *via* $\hat{\mathsf{g}}^{[Q]}, \dots, \hat{\mathsf{g}}^{[0]}$: no influence of the parasitic roots $\hat{\mathsf{g}}_2, \dots, \hat{\mathsf{g}}_{Q+1}$ of the amplification polynomial. An important consequence of this, in the context where $\delta t \propto \Delta x$ as $\Delta x \to 0$ and the target equation is the linear transport equation, is that a scheme of order ω can be initialized with methods of order $\omega - 1$ without lowering the overall order of the method, see [Str04, Theorem 10.6.2]. For concreteness, a leap-frog scheme (*i.e.* $\omega = 2$) can be started using a Lax-Friedrichs scheme, or even an unstable forward-Euler centered scheme.

The situation turns out to be radically different when multiple roots lay on the unit circle. These schemes, which we call "weakly" unstable, can be stable in practice, especially when the frequency at which multiple roots belong to the unit circle is the frequency $\xi = 0$ only. Therefore, provided that the underlying floating point arithmetic is accurate enough, these schemes can be suitable for computations. Genuine stability—*i.e.* (1)—requires control for arbitrary initial data $\mathsf{u}^Q, \dots, \mathsf{u}^0$ and thus cannot be established in these circumstances, due to the potential polynomial growth of the numerical solution in n. Despite this, any "reasonable" initialization scheme—which yields a specific choice of initial data—renders a stable computation as long as it avoids exciting the unstable frequency. Constraining the choice of initialization routines to well-chosen ones in order to achieve desired numerical properties—which we would otherwise unlikely obtain—has been studied in the context of linear multi-step schemes for ordinary differential equations [HRS03]. **The main finding and the subject we try to elucidate in the present paper is the following. In the weakly unstable framework, a scheme of order ω might need to be initialized with methods of the same order ω to preserve the overall order, contrarily to genuinely stable schemes. This is due to the entangled role of the several modes allowed by the multi-step scheme—both physical and parasitic—in the consistency of the method, since the lack of stability forbids from synthesizing consistency merely in terms of the physical root.** To the best of our knowledge, this observation is new in the available literature.

This conclusion has been accidentally drawn while trying to construct a fourth-order lattice Boltzmann scheme [MZ88] for the linear transport equation. Being stable, we reasonably conjectured that it would have needed third-order initializations in order to preserve its overall fourth-order. Surprisingly, despite the fact of having third-order initializations ready to use, numerical results featured orders of convergence stuck at three instead of the expected order four, without any trace of explosion of the numerical solution whatsoever. This fact indicates that the initialization of stable lattice Boltzmann schemes must be handled with particular care. The majority of the content in the present paper is likely to be essentially useless in applications and belatedly tweaks subjects that have been studied a long time ago, at the beginning of the 1970s [BT70]. Still, it emphasizes the key role of stability in making the study of consistency simple, *i.e.* based just on the local truncation error of the bulk multi-step scheme, plus the independent study of the initialization routines (*cf.* (1)). Indeed, these parts of the contribution should be valued for their pedagogical role. However, developing high-order lattice Boltzmann schemes for the linear transport equation can be of interest in applications. Indeed, these explicit methods are praised for their computational efficiency and can be used to approximate the solution of non-linear systems of conservation laws by embedding them into approximations [ADN00] of the Jin-Xin relaxation system [JX95], using high-order symmetric operator splittings [MQ02], in the spirit of [CFH$^+$19].

The paper is structured as follows. Section 1 presents the origin of the study from an empirical observation, where we found unexpected orders of convergence. Theory is developed in Section 2 to understand these results. Section 3 brings the numerical schemes back on the "battlefield" by investigating the role of floating-point arithmetic and thus of round-off errors. Finally, Section 4 presents an application of the fourth-order solver for the linear transport equation to the approximation of non-linear equations.

The model system we consider in the paper is the Cauchy problem associated with the linear transport equation at velocity $V \in \mathbb{R}$, which reads

$$\partial_t u(t,x) + V \partial_x u(t,x) = 0, \qquad t \in (0,T], \qquad x \in \mathbb{R}, \tag{2}$$

$$u(0,x) = u^\circ(x), \qquad x \in \mathbb{R}. \tag{3}$$

The initial datum u° is assumed to be a given smooth function, unless otherwise said. For the problem is linear, we conveniently consider the Fourier transform of (2), obtaining $\partial_t \hat{u}(t,\xi) = -iV\xi\hat{u}(t,\xi)$ for $\xi \in \mathbb{R}$. The explicit solution hence reads $\hat{u}(t,\xi) = e^{-iV\xi t}\hat{u}^\circ(\xi)$.

For the sake of approximating the solution of (2) and (3), we consider a uniform time-space discretization with steps δt and Δx, so that the discrete grid points in time will be $t^n := n\delta t$ with $n \in \mathbb{N}$ and those in space $x_k := k\Delta x$ with $k \in \mathbb{Z}$. For we consider explicit numerical methods for the hyperbolic equation (2), we naturally fix the ratio[2] $\lambda := \Delta x/\delta t$ to some positive real number as the spatial grid shrinks, *i.e.* $\Delta x \to 0$, and so we shall be allowed to use Δx as unique discretization parameter.

[2]We acknowledge that this notation is "standard" in the lattice Boltzmann community, while the Finite Difference community mostly employs the reciprocal $\lambda = \delta t/\Delta x$. We shall stick with the former notation.

1 The numerical experiment behind this study

As extensively stressed in the introduction, this study has been stimulated by the following example of multi-step scheme with $Q+1=3$ steps, which origin will be clarified in a moment. Let $n \geq 2$, $k \in \mathbb{Z}$, and consider:

$$\begin{aligned} \mathsf{u}_k^{n+1} =& \tfrac{1}{3}(1-4\mathcal{C}^2+2(\mathcal{C}^2-1)(\mathsf{D}^2+2)-6\mathcal{C}\mathsf{D}_0)\mathsf{u}_k^n \\ & -\tfrac{1}{3}(1-4\mathcal{C}^2+2(\mathcal{C}^2-1)(\mathsf{D}^2+2)+6\mathcal{C}\mathsf{D}_0)\mathsf{u}_k^{n-1}+\mathsf{u}_k^{n-2}, \end{aligned} \tag{4}$$

where the centered first-order finite difference D_0 is defined by $\mathsf{D}_0\mathsf{u}_k := (\mathsf{u}_{k+1}-\mathsf{u}_{k-1})/2$, and the centered second-order finite difference D^2 is given by $\mathsf{D}^2\mathsf{u}_k := \mathsf{u}_{k+1}-2\mathsf{u}_k+\mathsf{u}_{k-1}$. For it will play an important role in what follows, we define the Courant number $\mathcal{C} := V/\lambda$. In (4), we have to interpret $\mathsf{u}_k^n \approx u(t^n, x_k)$.

1.1 Construction of (4) from a lattice Boltzmann scheme

The multi-step Finite Difference scheme (4) is constructed starting from a lattice Boltzmann scheme, as previously described in [BGM22]. Let us emphasize that—though the theory on Finite Difference schemes has been understood for decades now—recent works in the framework of lattice Boltzmann schemes [FS21, BGM22] have demonstrated that these latter are indeed an inextinguishable source of multi-step Finite Difference schemes. Therefore the (still imperfect) rigorous understanding of lattice Boltzmann schemes is likely to pass from that of multi-step Finite Difference schemes [Bel23b].

1.1.1 Lattice Boltzmann algorithm: collide-and-stream

Consider a scheme in one space dimension featuring three discrete velocities, see for example [Dub13]. Without entering into the details on how lattice Boltzmann schemes are devised, we just consider that they stem from a collide-and-stream procedure made up as follows. Here, $n \in \mathbb{N}$.

- The *local collision phase* reads, for $k \in \mathbb{Z}$:

$$\mathsf{u}_k^{n\star} = \mathsf{u}_k^n, \qquad \mathsf{v}_k^{n\star} = (1-s_v)\mathsf{v}_k^n + s_v v^{\mathrm{eq}}(\mathsf{u}_k^n), \qquad \mathsf{w}_k^{n\star} = (1-s_w)\mathsf{w}_k^n + s_w w^{\mathrm{eq}}(\mathsf{u}_k^n). \tag{5}$$

  In these expressions, $s_v, s_w \in (0,2]$ are the relaxation parameters of the non-conserved moments v and w, whereas v^{eq} and w^{eq} are their equilibria: possibly non-linear functions of the conserved moment u. For the problem we aim at solving is linear, we consider linear equilibria and thus take $\epsilon_v, \epsilon_w \in \mathbb{R}$ such that $v^{\mathrm{eq}}(\mathsf{u}) = \epsilon_v \mathsf{u}$ and $w^{\mathrm{eq}}(\mathsf{u}) = \epsilon_w \mathsf{u}$.

- The *non-local stream phase* is written using another basis induced by $\boldsymbol{M}^{-1}$. We take

$$\boldsymbol{M} = \begin{bmatrix} M_{11} & 1 & 1 \\ 0 & 1 & -1 \\ M_{31} & 1 & 1 \end{bmatrix},$$

  where $M_{11}, M_{31} \in \mathbb{R}$ remain free parameters that could be tuned to change some features of the scheme, in particular stability. To keep the matrix $\boldsymbol{M}$ invertible, we have to enforce $M_{11} \neq M_{31}$. The usual choice [Dub13, Fév14, DGR20] is to take $M_{11}=1$, that is, if we interpret the first moment u as a density, all the "particles" have the same mass. It is natural to take $M_{12}=M_{13}$ for symmetry reasons, and we take these two entries equal to one as a normalization. It is also natural to consider $M_{21}=0$, to avoid linear transport terms which do not originate from the equilibria, being intrinsically linear. Usual values for M_{31} are zero [Fév14] (for simplicity) and -2 [DGR20] (to have orthogonal rows in $\boldsymbol{M}$ with respect to the Euclidean scalar product of vectors). Again, $M_{32}=M_{33}=1$ by symmetry and normalization arguments. The post-collisional distribution functions associated with the discrete velocities 0, 1, and -1 are recovered point-by-point by $(\mathsf{f}_{\circ,k}^{n\star}, \mathsf{f}_{+,k}^{n\star}, \mathsf{f}_{-,k}^{n\star})^{\mathsf{T}} = \boldsymbol{M}^{-1}(\mathsf{u}_k^{n\star}, \mathsf{v}_k^{n\star}, \mathsf{w}_k^{n\star})^{\mathsf{T}}$, where $k \in \mathbb{Z}$. The stream reads, for $k \in \mathbb{Z}$

$$\mathsf{f}_{\circ,k}^{n+1} = \mathsf{f}_{\circ,k}^{n\star}, \qquad \mathsf{f}_{+,k}^{n+1} = \mathsf{f}_{+,k-1}^{n\star}, \qquad \mathsf{f}_{-,k}^{n+1} = \mathsf{f}_{-,k+1}^{n\star}. \tag{6}$$

  After this phase, one can recover the moments by setting $(\mathsf{u}_k^{n+1}, \mathsf{v}_k^{n+1}, \mathsf{w}_k^{n+1})^{\mathsf{T}} = \boldsymbol{M}(\mathsf{f}_{\circ,k}^{n+1}, \mathsf{f}_{+,k}^{n+1}, \mathsf{f}_{-,k}^{n+1})^{\mathsf{T}}$.

The lattice Boltzmann scheme can be written on the moments using $\mathbf{E}$, a 3-by-3 matrix with entries on the ring of spatial Finite Difference operators on Cartesian grid, so that $(\mathsf{u}_k^{n+1}, \mathsf{v}_k^{n+1}, \mathsf{w}_k^{n+1})^{\mathsf{T}} = \mathbf{E}(\mathsf{u}_k^n, \mathsf{v}_k^n, \mathsf{w}_k^n)^{\mathsf{T}}$. Therefore $(\mathsf{u}_k^n, \mathsf{v}_k^n, \mathsf{w}_k^n)^{\mathsf{T}} = \mathbf{E}^n(\mathsf{u}_k^0, \mathsf{v}_k^0, \mathsf{w}_k^0)^{\mathsf{T}}$ and by the Parseval's identity, we naturally introduce the following definition of stability.

Definition 1 (Stability of a lattice Boltzmann scheme)**.** *A lattice Boltzmann scheme, such as* (5)/(6)*, is said to be "stable" if and only if* $\hat{\mathbf{E}}(\xi\Delta x)^n$ *is bounded for every* $|\xi\Delta x| \leq \pi$ *and for every* $n \in \mathbb{N}$*.*

Proposition 1 (Stability of a lattice Boltzmann scheme)**.** *A lattice Boltzmann scheme, such as* (5)/(6)*, is stable if and only if, for every* $|\xi\Delta x| \leq \pi$*, the minimal polynomial of* $\hat{\mathbf{E}}(\xi\Delta x)$ *is a simple* von Neumann *polynomial, namely none of its roots is outside the closed unit disk and those on the unit circle are simple.*

Proof. This is a consequence of the Jordan canonical form for complex matrices. If the minimal polynomial of $\hat{\mathbf{E}}(\xi\Delta x)$ is a simple *von Neumann* polynomial, then the maximal size of the Jordan blocks associated to each eigenvalue on the unit circle is one, which prevents polynomial growths of $\hat{\mathbf{E}}(\xi\Delta x)^n$ in n. Exponential growths are not possible since all the roots are in the closed unit disk. □

1.1.2 Tuning of the free parameters

Looking at the algorithm proposed in Section 1.1.1, we see that it features a large number of free parameters to be tuned, namely s_v, s_w, ϵ_v, ϵ_w, M_{11}, and M_{31}. We now select them according to the order of accuracy that we want to achieve with respect to the target equation (2). This could also be obtained by turning the lattice Boltzmann scheme—originally on u, v, and w—into a multi-step Finite Difference scheme solely on u, see [BGM22], and then computing the modified equations [WH74] or expanding the roots of the amplification polynomial in the small wave-number limit. Instead, we compute the modified equations on the original lattice Boltzmann scheme following the procedure proposed in [Dub22], where they are called "equivalent equations". We assume that all the parameters remain fixed as Δx—and *a fortiori* δt—go to zero. The obtained modified equation reads $\partial_t\psi + \Gamma^{(1)}(\psi) + \sum_{h=2}^{h=+\infty} \delta t^{h-1}\Gamma^{(h)}(\psi) = 0$, see [DL09, Equation (38)], where a generic function $\psi = \psi(t,x)$ appears to stress the fact that this is not the solution u of the target problem (2) and (3). The determination of $\Gamma^{(1)}(\psi)$, enforcing that $\Gamma^{(1)}(\psi) = V\partial_x\psi$ secures first-order consistency with (2). Obtaining $\Gamma^{(2)}(\psi) = \cdots = \Gamma^{(\omega)}(\psi) = 0$ ensures accuracy up to order ω. We proceed iteratively order-by-order and progressively incorporate any previous choice on the parameters.

- We obtain $\Gamma^{(1)}(\psi) = \lambda\epsilon_v\partial_x\psi$. To achieve consistency, we have to enforce $\epsilon_v = \mathcal{C}$.
- We have
$$\Gamma^{(2)}(\psi) = \lambda^2\Big(\frac{1}{s_v} - \frac{1}{2}\Big)\Big(-\frac{M_{31}}{M_{31}-M_{11}} + \mathcal{C}^2 + \frac{M_{11}}{M_{31}-M_{11}}\epsilon_w\Big)\partial_{xx}\psi.$$
There are two ways of having $\Gamma^{(2)}(\psi) = 0$ by making each term into parentheses vanish. We adopt $s_v = 2$.
- We obtain
$$\Gamma^{(3)}(\psi) = \frac{\lambda^3\mathcal{C}}{12}\Big(-2\mathcal{C}^2 + \frac{(1-3\epsilon_w)M_{11} + M_{31}}{M_{31}-M_{11}}\Big)\partial_x^3\psi.$$
We achieve $\Gamma^{(3)}(\psi) = 0$ through $\epsilon_w = \frac{1}{3}\Big(1 + \frac{2M_{31}}{M_{11}} - 2\frac{M_{31}-M_{11}}{M_{11}}\mathcal{C}^2\Big)$.
- We have
$$\Gamma^{(4)} = \frac{\lambda^4\mathcal{C}^2(\mathcal{C}^2-1)}{6}\Big(\frac{1}{s_w} - \frac{1}{2}\Big)\partial_x^4\psi,$$
hence achieve fourth-order accuracy by selecting $s_w = 2$.

After this procedure, the coefficients M_{11} and M_{31} are still free. Nevertheless, we are about to see that they do not play any major role in the rest of the paper and we can therefore fix them at our convenience. We finish on the stability of the lattice Boltzmann scheme.

Proposition 2 (Stability of the lattice Boltzmann scheme (5)/(6))**.** *The lattice Boltzmann scheme* (5)/(6) *with the previously selected parameters is stable under the CFL condition* $|\mathcal{C}| < 1/2$*.*

Proof. For $\xi\Delta x \neq 0$, Proposition 3 to come ensures that the characteristic polynomial of $\hat{\mathbf{E}}(\xi\Delta x)$ is a simple *Von Neumann* polynomial, thus this is also true for the minimal polynomial. For $\xi\Delta x = 0$, we have

$$\hat{\mathbf{E}}(0) = \begin{bmatrix} 1 & 0 & 0 \\ \star & -1 & 0 \\ \star & 0 & -1 \end{bmatrix}, \qquad \text{hence} \qquad \hat{\mathbf{E}}(0)^n = \begin{bmatrix} 1 & 0 & 0 \\ \star(1-(-1)^n) & (-1)^n & 0 \\ \star(1-(-1)^n) & 0 & (-1)^n \end{bmatrix},$$

where the $\star$ entries are terms depending on $\mathcal{C}$ and the choice of M_{11} and M_{31}. They are independent of n. Thus $\hat{\mathbf{E}}(0)$ is power bounded and thus the lattice Boltzmann scheme stable. Observe that $\hat{\mathbf{E}}(0)^2 = \boldsymbol{I}$, thus the polynomial $\mathsf{z}^2 - 1$ is the minimal polynomial of $\hat{\mathbf{E}}(0)$. It has two roots on the unit circle which are distinct. □

1.1.3 Corresponding Finite Difference scheme

Having a fourth-order lattice Boltzmann scheme at our disposal, we now describe how to forget about it and obtain (4). In our previous contributions [BGM22, Bel23b, Bel23a], we have shown how to recast any lattice Boltzmann scheme—whether it tackles linear or non-linear equation—under the form of a multi-step Finite Difference scheme on the conserved moments. In the present context, the corresponding Finite Difference scheme will be on u only. The amplification polynomial of the corresponding Finite Difference scheme reads

$$\hat{\Phi}(\xi\Delta x, \mathsf{z}) = \det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(\xi\Delta x)) = \mathsf{z}^3 + \eta(\xi\Delta x)\ \mathsf{z}^2 - \overline{\eta}(\xi\Delta x)\mathsf{z} - 1, \tag{7}$$

where the over-line denotes complex conjugation, $\eta(\xi\Delta x) = -\frac{1}{3}(1 - 4\mathcal{C}^2 + 4(\mathcal{C}^2 - 1)\cos(\xi\Delta x) - 6i\mathcal{C}\sin(\xi\Delta x))$, and $|\xi\Delta x| \leq \pi$. This is readily the amplification polynomial associated with (4). Observe that this scheme does not depend on the specific choice of M_{11} and M_{31}, as previously claimed.

1.1.4 Initialization

We fix $M_{11} = 1$ and $M_{31} = -2$ for simplicity. Since the bulk lattice Boltzmann scheme is fourth-order accurate, it needs to be initialized with at least third-order accurate schemes dictated by the choice of v^0 and w^0. The classical choice of taking them locally at equilibrium, namely selecting $\mathsf{v}_k^0 = \epsilon_v \mathsf{u}_k^0 = \mathcal{C}\mathsf{u}_k^0$ and $\mathsf{w}_k^0 = \epsilon_w \mathsf{u}_k^0 = (2\mathcal{C}^2 - 1)\mathsf{u}_k^0$ is not enough, because it yields first-order accurate solutions, see [Bel23a]. Considering

$$\mathsf{v}_k^0 = \mathcal{C}\mathsf{u}_k^0 + \frac{\mathcal{C}^2 - 1}{6}\mathsf{D}_0\mathsf{u}_k^0, \qquad \mathsf{w}_k^0 = (2\mathcal{C}^2 - 1)\mathsf{u}_k^0 + \mathcal{C}(\mathcal{C}^2 - 1)\mathsf{D}_0\mathsf{u}_k^0 + \delta\mathsf{D}^2\mathsf{u}_k^0 \tag{8}$$

gives a third-order initialization scheme for the first stage and a third-order, for $\delta \neq 0$, or fourth-order, for $\delta = 0$, initialization scheme for the second stage by slightly perturbing the local equilibrium. The amplification factors of the corresponding Finite Difference schemes will be given by $\hat{\mathsf{g}}^{[1]}(\xi\Delta x) = \boldsymbol{e}_1^{\mathsf{T}}\hat{\mathbf{E}}(\xi\Delta x)(1, \mathcal{C} + \frac{i(\mathcal{C}^2-1)}{6}\sin(\xi\Delta x), (2\mathcal{C}^2 - 1) + i\mathcal{C}(\mathcal{C}^2 - 1)\sin(\xi\Delta x) + 2\delta(\cos(\xi\Delta x) - 1))^{\mathsf{T}}$ and $\hat{\mathsf{g}}^{[2]}(\xi\Delta x) = \boldsymbol{e}_1^{\mathsf{T}}\hat{\mathbf{E}}(\xi\Delta x)^2(1, \mathcal{C} + \frac{i(\mathcal{C}^2-1)}{6}\sin(\xi\Delta x), (2\mathcal{C}^2 - 1) + i\mathcal{C}(\mathcal{C}^2 - 1)\sin(\xi\Delta x) + 2\delta(\cos(\xi\Delta x) - 1))^{\mathsf{T}}$ and feature quite involved expressions that we do not provide here. Still, we have the expansions $\hat{\mathsf{g}}^{[1]}(\xi\Delta x) = e^{-i\mathcal{C}\xi\Delta x(1+O(|\xi\Delta x|^3))}$ and $\hat{\mathsf{g}}^{[2]}(\xi\Delta x) = e^{-2i\mathcal{C}\xi\Delta x(1+(\delta-1)O(|\xi\Delta x|^3)+O(|\xi\Delta x|^4))}$ in the limit $|\xi\Delta x| \ll 1$.

1.1.5 A surprising numerical experiment

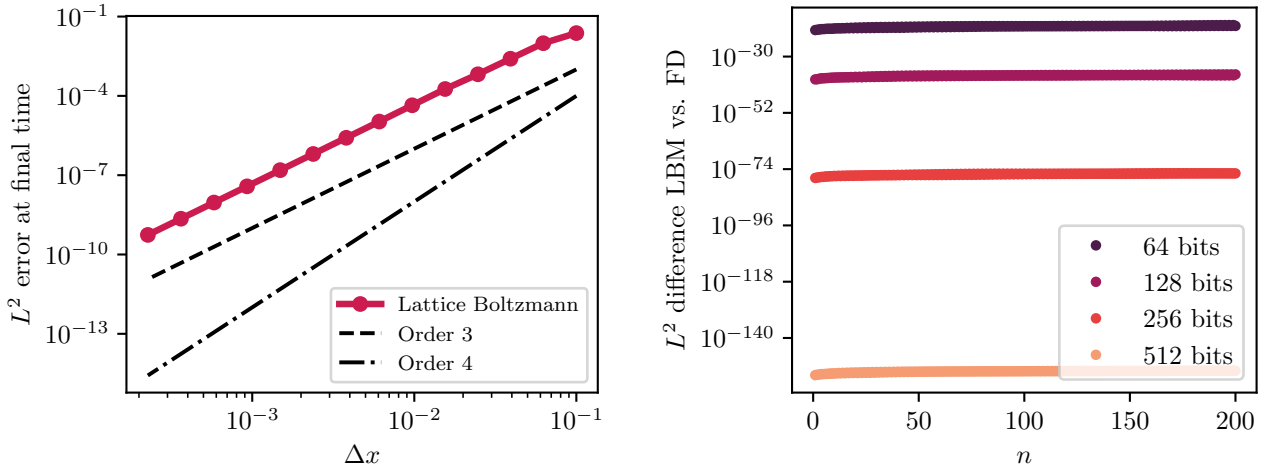

Figure 1: Left: error for the lattice Boltzmann scheme with initialization (8) with $\delta = 1$ at final time $T = 0.2$. Right: difference between lattice Boltzmann and Finite Difference scheme with $N_x = 200$ mesh points as time goes on for different floating point arithmetics.

We now test the order of convergence of the lattice Boltzmann scheme with respect to Δx. We simulate on a bounded domain $[-1, 1]$, enforcing periodic boundary conditions. We employ the original lattice Boltzmann scheme (5)/(6) with (8) using $\delta = 1$, hence having third-order initializations. The initial datum is $u^\circ(x) =$

$\exp(-1/(1-(2x)^2))\mathbb{I}_{(-1,1)}(2x)$, which is a smooth function of class $C_c^\infty([-1,1])$ fulfilling the periodic boundary conditions. We simulate using $\mathcal{C} = 1/4$. Surprisingly, the result on the left of Figure 1 shows third-order convergence instead of the expected fourth-order. We will come back to this fact in a few moments.

1.1.6 Equivalence of the lattice Boltzmann scheme and its corresponding Finite Difference scheme

Upon taking the initialization procedures into account, the unknowns u^n computed using the original lattice Boltzmann method (5)/(6) or its corresponding Finite Difference scheme (4) are mathematically the same. Of course, since the operations implemented on computers can be different, this is actually true up to machine precision. To demonstrate this fact, we follow the illustration by [Del23] and adopt the same setting of Section 1.1.5 with a grid made up of $N_x = 200$ points, using both the original lattice Boltzmann scheme and its corresponding Finite Difference scheme with different machine precisions. The results on the right of Figure 1 confirm our claim: the difference is of the order of the machine epsilon and accumulates in time.

1.2 An (even more) surprising numerical experiment

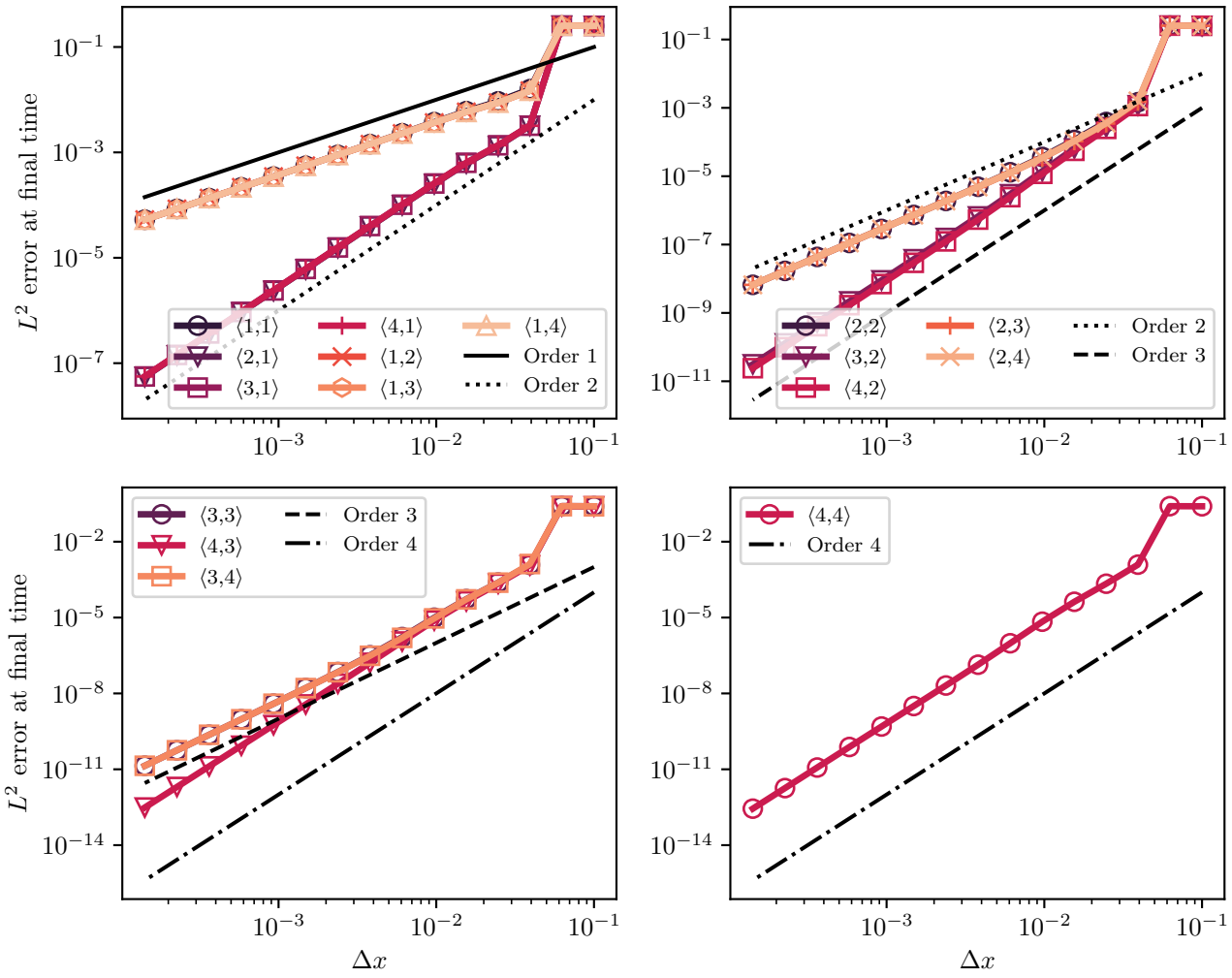

Figure 2: Error for (4) at final time $T = 0.2$ with different initialization schemes.

We now test the order of convergence of the corresponding Finite Difference scheme (4) with respect to Δx. The setting is the same as Section 1.1.5. We vary a large number of initialization schemes. In particular, we shall use the Lax-Friedrichs scheme as prototype of first-order initialization scheme, the Lax-Wendroff scheme for second-order, the OS3 scheme for third-order, and the OS4 scheme for fourth order, see [DT04]. Test cases will be numbered as follows: imagine to deal with a bulk scheme featuring $Q + 1 = 3$ steps, hence needing $Q = 2$

initialization schemes. The test case $\langle \omega_2, \omega_1 \rangle$ corresponds to a ω_2-order scheme (applied twice) for the second initialization and a ω_1-order scheme for the first initialization. In the results given in Figure 2 for the L^2 error at T, we observe an unexpected result:

$$\text{overall order} = \min(\omega, \omega_2, \omega_1 + 1), \quad \textit{in lieu} \text{ of the expected} \quad \text{overall order} = \min(\omega, \omega_2 + 1, \omega_1 + 1)$$

from [Str04, Theorem 10.6.2], where the order of the bulk scheme is ω (in our case $\omega = 4$). This means that we have to feed the second time step $n = 2$ with an initialization of the same order of accuracy as the bulk scheme to preserve the overall order. Observe that the computations shown in Figure 2 remain stable (and converge).

2 Understanding the numerical experiment

We now try to investigate the reasons behind all these unexpected results. To this end, we first show that the Finite Difference scheme is weakly unstable and features—along with the original lattice Boltzmann scheme—two travelling parasitic modes whose speed of propagation sets a specific CFL constraint different from the usual $|\mathcal{C}| \leq 1$. Then, we theoretically study the order of convergence by proving that thanks to the initialization, the scheme remains stable, and that the global truncation error is shaped by the parasitic modes. They also allow to study the qualitative behavior of the error in time and explain supra-convergent results in simulations run under periodic boundary conditions. Finally, we numerically analyze the case where the initial datum u° is not smooth.

2.1 Weak instability for general data

The stability of (4) needs to be studied using the roots of its amplification polynomial (7).

Definition 2 (Stability/weak instability of a Finite Difference scheme)**.** *Consider a Finite Difference scheme explicitly independent of δt and Δx with amplification polynomial $\hat{\Phi}(\xi\Delta x, \mathsf{z})$.*

- *We say that the scheme is "stable" if, for every $|\xi\Delta x| \leq \pi$, $\hat{\Phi}(\xi\Delta x, \mathsf{z})$ is a simple* von Neumann *polynomial.*
- *We say that the scheme is "weakly unstable" if for every $|\xi\Delta x| \leq \pi$, $\hat{\Phi}(\xi\Delta x, \mathsf{z})$ is a* von Neumann *polynomial, namely all the roots are inside the closed unit disk, and for some $|\tilde{\xi}\Delta x| \leq \pi$, $\hat{\Phi}(\tilde{\xi}\Delta x, \mathsf{z})$ has multiple roots on the unit circle.*

If we compute its conjugate reciprocal polynomial (or inversive polynomial) [Vie19] given by $\hat{\Phi}^\dagger(\xi\Delta x, \mathsf{z}) := \mathsf{z}^{Q+1}\hat{\Phi}(-\xi\Delta x, 1/\mathsf{z})$, we obtain

$$\hat{\Phi}^\dagger(\xi\Delta x, \mathsf{z}) = -\mathsf{z}^3 - \eta(\xi\Delta x)\mathsf{z}^2 + \overline{\eta}(\xi\Delta x)\ \mathsf{z} + 1.$$

We observe that since $\hat{\Phi}(\xi\Delta x, \mathsf{z}) = -\hat{\Phi}^\dagger(\xi\Delta x, \mathsf{z})$, the polynomial $\hat{\Phi}(\xi\Delta x, \mathsf{z})$ is said to be "self-inversive" [Mar66, Chapter 10], [MRM94, Chapter 1], and [Vie19], that is, there exists ω on the unit circle such that $\hat{\Phi}(\xi\Delta x, \mathsf{z}) = \omega\hat{\Phi}^\dagger(\xi\Delta x, \mathsf{z})$. In our case, $\omega = -1$ is independent of $\xi\Delta x$. From [Vie19, Theorem 1], since $\hat{\Phi}(\xi\Delta x, \mathsf{z})$ is of odd degree, we deduce that it has at least one root over the unit circle. Moreover, if $\hat{\mathsf{g}}$ is a root of $\hat{\Phi}(\xi\Delta x, \mathsf{z})$, then also $1/\overline{\hat{\mathsf{g}}}$ is a root of $\hat{\Phi}(\xi\Delta x, \mathsf{z})$. The zeros of this kind of polynomial either belong to unit circle or occur in pairs conjugate with respect to the unit circle [KP08]. Moreover $\mathsf{z}^{-1}(\hat{\Phi}^\dagger(\xi\Delta x, 0)\hat{\Phi}(\xi\Delta x, \mathsf{z}) - \hat{\Phi}^\dagger(\xi\Delta x, \mathsf{z})\hat{\Phi}(\xi\Delta x, 0)) \equiv 0$, hence—see [Str04, Chapter 4]—all the roots of the amplification polynomial lie on the unit circle for every wave-number.

Proposition 3 (Weak instability of (4))**.** *Assume that the Courant–Friedrichs–Lewy (CFL) condition $|\mathcal{C}| < 1/2$ holds. Then,* (4) *is weakly unstable. More precisely:*

- *For $|\xi\Delta x| \in (0, \pi]$, the amplification polynomial $\hat{\Phi}(\xi\Delta x, \mathsf{z})$ given by* (7) *has distinct roots on the unit circle.*
- *For $|\xi\Delta x| = 0$, the amplification polynomial $\hat{\Phi}(0, \mathsf{z})$ given by* (7) *has roots* 1 *(single) and* -1 *(double).*

Proposition 3—proved in Appendix A—means that (4) is stable for the L^2-norm except for the frequency zero, which could cause linear growth of the solution in time. Indeed, using the language of linear multi-step schemes for ODEs, the scheme is not zero-stable for $\xi = 0$ and in this case resembles to [SM03, Example 12.5 (d)]. If we have not had the spatial direction x, we could not expect convergence. Still, the presence of the spatial extension helps us in having an overall stable procedure.

One legitimate question concerns the meaning of the CFL condition $|\mathcal{C}| < 1/2$. This can be seen by explicitly computing the roots of the amplification polynomial and perform Taylor expansions in the low-frequency limit $|\xi\Delta x| \ll 1$. This is

$$\hat{\mathsf{g}}_1(\xi\Delta x) = 1 - i\mathcal{C}\xi\Delta x - \tfrac{1}{2}\mathcal{C}^2\xi^2\Delta x^2 + \tfrac{i}{6}\mathcal{C}^3\xi^3\Delta x^3 + \tfrac{1}{24}\mathcal{C}^4\xi^4\Delta x^4 + \tfrac{i\mathcal{C}}{360}\big(5\mathcal{C}^4 - 10\mathcal{C}^2 + 2\big)\xi^5\Delta x^5 + O(|\xi\Delta x|^6)$$
$$= e^{-i\mathcal{C}\xi\Delta x(1+O(|\xi\Delta x|^4))}, \quad (9)$$

which proves that—as already emphasized—the method is fourth-order accurate. Even more precisely, *cf.* [Str04, CL20]: it exists a constant $C > 0$ such that $\frac{1}{\delta t}|e^{-i\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}_1(\xi\Delta x)| \leq C\Delta x^4|\xi|^5$ for $|\xi\Delta x| \leq \pi$. For the parasitic eigenvalues:

$$\hat{\mathsf{g}}_2(\xi\Delta x) = -e^{\frac{i\sqrt{3}}{6}(\sqrt{3}\mathcal{C}+\sqrt{8-5\mathcal{C}^2})\xi\Delta x(1+O(|\xi\Delta x|^2))}, \quad (10)$$

$$\hat{\mathsf{g}}_3(\xi\Delta x) = -e^{\frac{i\sqrt{3}}{6}(\sqrt{3}\mathcal{C}-\sqrt{8-5\mathcal{C}^2})\xi\Delta x(1+O(|\xi\Delta x|^2))}. \quad (11)$$

The exponential form in (10) and (11) is inspired by [MQ02, Theorem 19] and emphasizes that the parasitic eigenvalues essentially behave like pseudo-schemes of order Δx^2 for a different flow compared to the target equation. Remark that $8 - 5\mathcal{C}^2 > 0$ by the constraint $|\mathcal{C}| \leq 1$ that must naturally hold for an explicit method with spatial stencil of one, see [Str62]. The first parasitic mode—brought by $\hat{\mathsf{g}}_2$—propagates backward, whatever the sign of $\mathcal{C}$, whereas the second mode—carried by $\hat{\mathsf{g}}_3$—always propagates forward. Both produce rapid checkerboard-like oscillating solutions since $\hat{\mathsf{g}}_2(0) = \hat{\mathsf{g}}_3(0) = -1$. We would like the parasitic waves to propagate slower than the speed of information of the scheme, which is equal to $\lambda = \Delta x/\delta t$. This can be stated as

$$\begin{cases} \frac{\sqrt{3}}{6}(\sqrt{3}\mathcal{C} + \sqrt{8-5\mathcal{C}^2}) < 1, & \to \quad \mathcal{C} \in (-1, 1/2), \\ \frac{\sqrt{3}}{6}(\sqrt{3}\mathcal{C} - \sqrt{8-5\mathcal{C}^2}) > -1, & \to \quad \mathcal{C} \in (-1/2, 1), \end{cases} \quad \to \quad |\mathcal{C}| < 1/2,$$

which is indeed the CFL condition by Proposition 3. This demonstrates that in this case, the CFL constraint is a condition on the speed of propagation of information by the parasitic modes. By studying $\frac{\sqrt{3}}{6}(\sqrt{3}\mathcal{C}+\sqrt{8-5\mathcal{C}^2}) > \mathcal{C}$, we see that the velocity of the parasitic waves is always larger than the one of the physical wave, as expected, because otherwise the CFL constraint would have stemmed from the speed of propagation of information by the physical mode.

We now understand why the numerical simulations in Section 1.2 go against [Str04, Theorem 10.6.2]. This results needs the numerical scheme to be genuinely stable... and (4) is not—according to Proposition 3. The aim of the sections to come is to understand why the numerical scheme still behaves in a stable fashion and converges, though with unexpected rates.

Remark 1 (A simpler numerical scheme with the similar features)**.** *We can construct a simpler multi-step scheme with features analogous to* (4)*, building an* ad hoc *amplification polynomial. We would like it to have a stable method for all wave-numbers except for the frequency zero, where a double root* -1 *is present. We do not request all the roots to be on the unit circle for all* $|\xi\Delta x| \leq \pi$*. We thus consider*

$$\hat{\Phi}(\xi\Delta x, \mathsf{z}) = (\mathsf{z} - \hat{\mathsf{g}}_{\mathrm{OS4}}(\xi\Delta x))(\mathsf{z} + \cos(\xi\Delta x))(\mathsf{z} + 1), \quad (12)$$

where $\hat{\mathsf{g}}_{\mathrm{OS4}}(\xi\Delta x)$ *is the amplification factor of the OS4 scheme. It could indeed be replaced by the one of any dissipative one-step scheme (this constraint would exclude the Lax-Friedrichs scheme since we would have a double root* -1 *at* $\xi\Delta x = \pi$*). This type of scheme gives the same surprising results as in Section 1.2, which are not included in the paper. This is rather paradoxical but instructive: if we had taken the fourth-order one-step scheme associated with* $\hat{\mathsf{g}}_{\mathrm{OS4}}$ *and made a (finite) number of iterations at the very beginning with a third-order scheme, we would have preserved an overall fourth-order. Similarly, if we had considered parasitic roots in of the amplification polynomial* $\hat{\mathsf{g}}_2$ *and* $\hat{\mathsf{g}}_3$ *fulfilling the stability condition also at* $\xi\Delta x = 0$*, for example* $\tilde{\Phi}(\xi\Delta x, \mathsf{z}) = (\mathsf{z} - \hat{\mathsf{g}}_{\mathrm{OS4}}(\xi\Delta x))(\mathsf{z} + \frac{1}{2})(\mathsf{z} + \frac{1}{3})$*, third-order initializations would have been enough to preserve fourth order. These two observations confirm that in this weakly unstable framework, the parasitic roots that we have artificially put along* $\hat{\mathsf{g}}_{\mathrm{OS4}}$ *start playing a role as far as consistency (and thus the order of accuracy) is concerned.*

Remark 2 (Trying to re-establish stability)**.** *Since, if we do not set* $s_v = s_w = 2$ *as we did in Section 1.1.2 to enforce fourth-order consistency, we have that* $\mathrm{sp}(\hat{\mathbf{E}}(0)) = \{1, 1 - s_v, 1 - s_w\}$*, or equivalently* $\hat{\Phi}(0, \mathsf{z}) = \det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(0)) = (\mathsf{z} - 1)(\mathsf{z} + s_v - 1)(\mathsf{z} + s_w - 1)$*, we could hope to solve the "collision" between the second and the third root lying on the unit disk and coinciding by considering* $s_v = 2$ *and* $s_w = 2 - \Delta x^\alpha$ *with* $\alpha \in \mathbb{R}$*. Here,* α *would be chosen large enough not to perturb the fourth order of the scheme. This setting does not provide the expected*

result and gives the same result as Section 1.2. This comes from the fact that now [Str04, Theorem 4.2.2] applies and stability requires that $\hat{\mathsf{g}}_2(0)$ and $\hat{\mathsf{g}}_3(0)$ are apart by a positive quantity independent of Δx when the space step is small, i.e. $|\hat{\mathsf{g}}_2(0) - \hat{\mathsf{g}}_3(0)| \geq c_1$. *However, in our case $|\hat{\mathsf{g}}_2(0) - \hat{\mathsf{g}}_3(0)| = \Delta x^\alpha$, hence the scheme has not been stabilized.*

2.2 Understanding convergence

As observed at the very beginning and through Remark 1, the consistency analysis of the whole scheme does no longer boil down to consider the behavior of the scheme as determined only by $\hat{\mathsf{g}}_1$, but we have to take all the roots into account and clarify how these different modes are excited and interact by the choice of initialization schemes.

2.2.1 Several decompositions of the discrete scheme

This is achieved using several kinds of decompositions of the discrete solution. Given the amplification polynomial $\hat{\Phi}(\xi\Delta x, \mathsf{z}) = \mathsf{z}^{Q+1} + \sum_{n=0}^{n=Q} \hat{\mathsf{c}}_n(\xi\Delta x)\mathsf{z}^n$ of an explicit scheme, we introduce its companion matrix (or amplification matrix [Cou21])

$$\hat{\mathbf{C}}(\xi\Delta x) = \begin{bmatrix} -\hat{\mathsf{c}}_Q(\xi\Delta x) & \cdots & -\hat{\mathsf{c}}_1(\xi\Delta x) & -\hat{\mathsf{c}}_0(\xi\Delta x) \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 1 & 0 \end{bmatrix}.$$

We form the amplification factors, given, for $n \geq Q+1$ by

$$\hat{\mathsf{g}}^{[n]}(\xi\Delta x) = \boldsymbol{e}_1^\mathsf{T} \hat{\mathbf{C}}(\xi\Delta x)^{n-Q} (\hat{\mathsf{g}}^{[Q]}(\xi\Delta x), \dots, \hat{\mathsf{g}}^{[1]}(\xi\Delta x), 1)^\mathsf{T}, \tag{13}$$

so that we have $\hat{\mathsf{u}}^n(\xi) = \hat{\mathsf{g}}^{[n]}(\xi\Delta x)\hat{\mathsf{u}}^0(\xi)$. Here $\hat{\mathsf{g}}^{[Q]}(\xi\Delta x), \dots, \hat{\mathsf{g}}^{[1]}(\xi\Delta x)$ are the amplification factors of the initialization schemes. Let us point out a feature concerning the companion matrix of a weakly unstable scheme.

Proposition 4. *Let $\hat{\mathbf{C}}(\xi\Delta x)$ for $|\xi\Delta x| \leq \pi$ be the companion matrix of a weakly unstable Finite Difference scheme. Then $\hat{\mathbf{C}}(\tilde{\xi}\Delta x)^n$, for some $|\tilde{\xi}\Delta x| \leq \pi$, grows (polynomially) with $n \in \mathbb{N}$.*

Proof. Let $|\tilde{\xi}\Delta x| \leq \pi$ be one of the frequencies where multiple roots of the amplification polynomial happen to be on the unit circle. The companion matrix has coinciding characteristic (*i.e.* the amplification polynomial) and minimal polynomial. Therefore, for this frequency, its Jordan canonical form inferred from the minimal polynomial features a block for the multiple eigenvalue on the unit circle having size larger than one, giving the claimed growth. □

Green decomposition A tool to understand the different role of the initialization $\hat{\mathsf{g}}^{[2]}$ compared to $\hat{\mathsf{g}}^{[1]}$ in (4) can be found using Green functions [CL99, CF22, Cou22].We remark that the phenomenon observed in Section 1.2 corresponds to the fact that the minimal requirement to preserve the overall order four is that the vector $(\hat{\mathsf{g}}^{[2]}, \hat{\mathsf{g}}^{[1]}, 1)$ must to be equal to the eigenvector $(\hat{\mathsf{g}}_1^2, \hat{\mathsf{g}}_1, 1)$ of the companion matrix $\hat{\mathbf{C}}$ associated with $\hat{\mathsf{g}}_1$ at different orders for $|\xi\Delta x| \ll 1$ according to the component (four for the first and third component, three for the second one). Otherwise said, one wants at least $(\hat{\mathsf{g}}^{[2]}, \hat{\mathsf{g}}^{[1]}, 1) = (\hat{\mathsf{g}}_1^2 + O(|\xi\Delta x|^5), \hat{\mathsf{g}}_1 + O(|\xi\Delta x|^4), 1 + O(|\xi\Delta x|^5))$. We therefore isolate the role of each initialization scheme by considering the ℓ-th Green functions $\hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)$—for $\ell \in [0, Q]$—defined by

$$\begin{cases} \hat{\mathsf{G}}_\ell^{[n+1]} = -\sum_{p=0}^{Q} \hat{\mathsf{c}}_p \hat{\mathsf{G}}_\ell^{[n+p-Q]} = \boldsymbol{e}_1^\mathsf{T} \hat{\mathbf{C}} (\hat{\mathsf{G}}_\ell^{[n]}, \dots, \hat{\mathsf{G}}_\ell^{[n-Q]})^\mathsf{T}, & \text{for} \quad n \geq Q, \\ \hat{\mathsf{G}}_\ell^{[n]} = \delta_{n,\ell}, & \text{for} \quad n \in [0, Q]. \end{cases}$$

Then, for $n \geq Q+1$, we have that $\hat{\mathsf{G}}_\ell^{[n]} = \boldsymbol{e}_1^\mathsf{T} \hat{\mathbf{C}}^{n-Q} \boldsymbol{e}_{Q-\ell+1}$ and, by adding and subtracting well-selected quantities:

$$\begin{aligned} \hat{\mathsf{g}}^{[n]}(\xi\Delta x) = \sum_{\ell=0}^{Q} \hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)\hat{\mathsf{g}}^{[\ell]}(\xi\Delta x) &= \sum_{\ell=0}^{Q} \hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)\hat{\mathsf{g}}_1(\xi\Delta x)^\ell + \sum_{\ell=1}^{Q} \hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)(\hat{\mathsf{g}}^{[\ell]}(\xi\Delta x) - \hat{\mathsf{g}}_1(\xi\Delta x)^\ell) \\ &= \boldsymbol{e}_1^\mathsf{T} \hat{\mathbf{C}}(\xi\Delta x)^{n-Q} (\hat{\mathsf{g}}_1(\xi\Delta x)^Q, \dots, \hat{\mathsf{g}}_1(\xi\Delta x), 1)^\mathsf{T} + \sum_{\ell=1}^{Q} \hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)(\hat{\mathsf{g}}^{[\ell]}(\xi\Delta x) - \hat{\mathsf{g}}_1(\xi\Delta x)^\ell) \\ &= \hat{\mathsf{g}}_1(\xi\Delta x)^n + \sum_{\ell=1}^{Q} \hat{\mathsf{G}}_\ell^{[n]}(\xi\Delta x)(\hat{\mathsf{g}}^{[\ell]}(\xi\Delta x) - \hat{\mathsf{g}}_1(\xi\Delta x)^\ell), \end{aligned} \tag{14}$$

where the last equality is obtained using the fact that $(\hat{\mathsf{g}}_1(\xi\Delta x)^Q, \dots, \hat{\mathsf{g}}_1(\xi\Delta x), 1)^\mathsf{T}$ is the eigenvector of $\hat{\mathbf{C}}$ relative to the eigenvalue $\hat{\mathsf{g}}_1$.

Modal decomposition Another decomposition of $\hat{\mathsf{g}}^{[n]}$ can be found as follows. If we assume that for a given wave-number ξ such that $|\xi\Delta x| \leq \pi$, all the roots $\hat{\mathsf{g}}_1(\xi\Delta x), \dots, \hat{\mathsf{g}}_{Q+1}(\xi\Delta x)$ are distinct, we have the so-called "modal" decomposition, directly inspired from the theory of linear recurrences, which reads

$$\hat{\mathsf{g}}^{[n]}(\xi\Delta x) = \sum_{\ell=1}^{Q+1} \hat{\sigma}_\ell(\xi\Delta x)\hat{\mathsf{g}}_\ell(\xi\Delta x)^n, \tag{15}$$

and where the coefficients $\hat{\sigma}_\ell$ are determined by the initialization schemes $\hat{\mathsf{g}}^{[Q]}, \dots, \hat{\mathsf{g}}^{[1]}$. In particular, introducing the Vandermonde matrix

$$\hat{\mathbf{V}}(\xi\Delta x) = \begin{bmatrix} \hat{\mathsf{g}}_1(\xi\Delta x)^Q & \cdots & \hat{\mathsf{g}}_{Q+1}(\xi\Delta x)^Q \\ \vdots & & \vdots \\ \hat{\mathsf{g}}_1(\xi\Delta x) & \cdots & \hat{\mathsf{g}}_{Q+1}(\xi\Delta x) \\ 1 & & 1 \end{bmatrix},$$

gives that $\hat{\sigma}_1, \dots \hat{\sigma}_{Q+1}$ satisfy the linear system $\hat{\mathbf{V}}(\xi\Delta x)(\hat{\sigma}_1(\xi\Delta x), \dots, \hat{\sigma}_{Q+1}(\xi\Delta x))^\mathsf{T} = (\hat{\mathsf{g}}^{[Q]}(\xi\Delta x), \hat{\mathsf{g}}^{[Q-1]}(\xi\Delta x), \dots, 1)^\mathsf{T}$. The fact that the Vandermonde matrix can be inverted comes from the assumption of dealing with distinct roots.

Comparison between decompositions Let us comment on the different properties of the Green decomposition (14) *vs.* the modal decomposition (15).

- Green decomposition (14). It focuses on the consistency eigenvalue $\hat{\mathsf{g}}_1$: one sees the overall scheme after n time-steps as the application of the pseudo-scheme associated with $\hat{\mathsf{g}}_1$ n-times, plus some perturbation induced by the initialization schemes $\hat{\mathsf{g}}^{[Q]}, \dots, \hat{\mathsf{g}}^{[1]}$ and their deviation from $\hat{\mathsf{g}}_1^Q, \dots, \hat{\mathsf{g}}_1$. These deviations are weighted by the Green functions. This decomposition is therefore used to understand consistency and the fact that the overall scheme reacts differently according to the step fed by a given initialization scheme. Its main drawback is that it mixes the parasitic modes $\hat{\mathsf{g}}_2, \dots, \hat{\mathsf{g}}_{Q+1}$ and thus hides their role of rapidly oscillating transport modes at different velocities (see (10) and (11)) compared to the physical mode.
- Modal decomposition (15). It can informally described by the sentence "All modes are created equal", for it does not emphasize any root between $\hat{\mathsf{g}}_1, \hat{\mathsf{g}}_2, \dots, \hat{\mathsf{g}}_{Q+1}$. It aims at alleviating the drawback of the Green decomposition, allowing to see the discrete solution as the superposition of $Q+1$ modes, each one having its own velocity. This helps the study of qualitative properties of the discrete solution. However, this approach is less suitable (at least without explicit computation of the coefficients at hand) to proceed to a rigorous study of the consistency of the scheme.

2.2.2 Stability

In the cases where the amplification polynomials are given by (7) and (12), $\hat{\mathbf{C}}(\xi\Delta x)^{n-Q}$ is not uniformly power bounded [LT84], because the schemes are weakly unstable, *cf.* Proposition 4. In both cases, we have

$$\boldsymbol{e}_1^\mathsf{T}\hat{\mathbf{C}}(0)^{n-2} = (\tfrac{1}{4}(-1)^n(2n-1) + \tfrac{1}{4}, \tfrac{1}{2}(-1)^{n+1} + \tfrac{1}{2}, \tfrac{1}{4}(-1)^{n+1}(2n-3) + \tfrac{1}{4}), \tag{16}$$

where the first and last entries diverge linearly with n. However, the reason why the simulations remain stable is that this instability is not excited. Indeed, any reasonable—meaning at least zero-order accurate—initialization scheme that one can consider is such that $\hat{\mathsf{g}}^{[2]}(0) = \hat{\mathsf{g}}^{[1]}(0) = 1$, hence we obtain from (13) that $\hat{\mathsf{g}}^{[n]}(0) = 1$ for any $n \in \mathbb{N}$, which is bounded as n grows, and thus keeps the simulation stable, *cf.* Section 1.2. The other frequencies $\xi \neq 0$ are stable (*i.e.* $\boldsymbol{e}_1^\mathsf{T}\hat{\mathbf{C}}(\xi\Delta x)^{n-2}$ is bounded with n) thanks to Proposition 3. This indeed shows that there exists $C_\mathrm{s} > 0$ independent of $n \in \mathbb{N}$ and $|\xi\Delta x| \leq \pi$ such that

$$|\hat{\mathsf{g}}^{[n]}(\xi\Delta x)| \leq C_\mathrm{s}. \tag{17}$$

Observe that the constant C_s can be—for the considered tests—slightly larger than one, but this is unimportant. This constant does not depend on the final time T. Notice that (17) exactly coincides—for the considered class of initializations—with the notion of stability stated in [GKO95, Theorem 2.1.1].

2.2.3 Consistency

Since initializations must be seriously taken into account due to instabilities, we cannot use the notion of local truncation error to derive an estimate on the global truncation error—the quantity one is eventually interested in. We therefore have to propose estimates directly on the global truncation error, as in [GKO95, Section 2.2], using the decomposition on the Green functions (14).

Remark 3 (Smoothness)**.** *We remark that* $\hat{\mathsf{g}}^{[n]}(\xi\Delta x)$ *and the Green functions* $\hat{\mathsf{G}}_2^{[n]}(\xi\Delta x), \hat{\mathsf{G}}_1^{[n]}(\xi\Delta x)$*, and* $\hat{\mathsf{G}}_0^{[n]}(\xi\Delta x)$ *are smooth functions of their argument since they are products of smooth functions of the form* $e^{\pm i\xi\Delta x}$*. This is distinct from the difficulties that can arise in the explicit determination of formulæ or of their Taylor expansions as* $|\xi\Delta x| \ll 1$*, due to the fact that the multiplicity of the roots of the amplification polynomial changes with* $\xi\Delta x$*,* cf. *Proposition 3. Since these quantities are smooth functions, their Taylor expansions for* $|\xi\Delta x| \ll 1$ *will be determined using their explicit expressions for the case* $|\xi\Delta x| \neq 0$*, which are found by the standard theory of linear constant coefficient recurrence relations, and then carefully passing to the limit.*

For every $|\xi\Delta x| \leq \pi$ (arguments are sometimes omitted for the sake of compactness), (14) provides

$$\hat{\mathsf{g}}^{[n]}(\xi\Delta x) = \hat{\mathsf{g}}_1(\xi\Delta x)^n + \hat{\mathsf{G}}_2^{[n]}(\xi\Delta x)(\hat{\mathsf{g}}^{[2]}(\xi\Delta x) - \hat{\mathsf{g}}_1(\xi\Delta x)^2) + \hat{\mathsf{G}}_1^{[n]}(\xi\Delta x)(\hat{\mathsf{g}}^{[1]}(\xi\Delta x) - \hat{\mathsf{g}}_1(\xi\Delta x)). \tag{18}$$

The explicit formulæ in terms of the roots are

$$\hat{\mathsf{G}}_2^{[n]}(\xi\Delta x) = \begin{cases} \frac{\hat{\mathsf{g}}_1^n(\hat{\mathsf{g}}_2-\hat{\mathsf{g}}_3)-\hat{\mathsf{g}}_2^n(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_3)+\hat{\mathsf{g}}_3^n(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_2)}{\hat{\mathsf{g}}_1^2(\hat{\mathsf{g}}_2-\hat{\mathsf{g}}_3)-\hat{\mathsf{g}}_2^2(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_3)+\hat{\mathsf{g}}_3^2(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_2)}, & |\xi\Delta x| \neq 0, \\ \frac{1}{4}(2(-1)^n n + (-1)^{n+1} + 1), & |\xi\Delta x| = 0. \end{cases} \tag{19}$$

We observe that this Green function is the potentially explosive one: still it is sufficient to have a zero-order scheme for $\hat{\mathsf{g}}^{[2]}$ in order to disengage this term. Since $\hat{\mathsf{G}}_2^{[n]}(\xi\Delta x)$ is smooth, we can obtain higher order terms for the limit $|\xi\Delta x| \ll 1$ using (19) for $|\xi\Delta x| \neq 0$. This yields

$$\hat{\mathsf{G}}_2^{[n]}(\xi\Delta x) = (-1)^n \left\lfloor \tfrac{n}{2} \right\rfloor + (-1)^n \left\lfloor \tfrac{n-1}{2} \right\rfloor \left(\left\lfloor \tfrac{n-1}{2} \right\rfloor + 1 \right) i\mathcal{C}\xi\Delta x + O(n^3)(\xi\Delta x)^2 + O(|\xi\Delta x|^3).$$

The growth of this Green function with n shows why we need—besides stability—one order more for the initialization $\hat{\mathsf{g}}^{[2]}$, the error coming from this choice is amplified (we might say that it resonates) and accumulates in time analogously to the consistency mode $\hat{\mathsf{g}}_1^n$. Thus, it needs to be of the same order as $\hat{\mathsf{g}}_1$ not to lower the overall order. The explicit form of (19), in particular the denominator, comes from the fact that for $|\xi\Delta x| \neq 0$, the companion matrix is diagonalisable:

$$\begin{aligned} \hat{\mathbf{C}}(\xi\Delta x)^{n-2} &= \hat{\mathbf{V}}(\xi\Delta x)^{-1} \mathrm{diag}(\hat{\mathsf{g}}_1(\xi\Delta x)^{n-2}, \hat{\mathsf{g}}_2(\xi\Delta x)^{n-2}, \hat{\mathsf{g}}_3(\xi\Delta x)^{n-2}) \hat{\mathbf{V}}(\xi\Delta x) \\ &= \frac{\mathbf{adj}(\hat{\mathbf{V}}(\xi\Delta x))}{\det(\hat{\mathbf{V}}(\xi\Delta x))} \mathrm{diag}(\hat{\mathsf{g}}_1(\xi\Delta x)^{n-2}, \hat{\mathsf{g}}_2(\xi\Delta x)^{n-2}, \hat{\mathsf{g}}_3(\xi\Delta x)^{n-2}) \hat{\mathbf{V}}(\xi\Delta x), \end{aligned}$$

with $\det(\hat{\mathbf{V}}(\xi\Delta x)) = \hat{\mathsf{g}}_1^2(\hat{\mathsf{g}}_2-\hat{\mathsf{g}}_3) - \hat{\mathsf{g}}_2^2(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_3) + \hat{\mathsf{g}}_3^2(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_2)$. Unsurprisingly, $\det(\hat{\mathbf{V}}(0)) = 0$. Still, also the numerator $\hat{\mathsf{g}}_1^n(\hat{\mathsf{g}}_2 - \hat{\mathsf{g}}_3) - \hat{\mathsf{g}}_2^n(\hat{\mathsf{g}}_1 - \hat{\mathsf{g}}_3) + \hat{\mathsf{g}}_3^n(\hat{\mathsf{g}}_1 - \hat{\mathsf{g}}_2) = 0$ for $\xi\Delta x = 0$. For the other Green function:

$$\hat{\mathsf{G}}_1^{[n]}(\xi\Delta x) = \begin{cases} -\frac{\hat{\mathsf{g}}_1^n(\hat{\mathsf{g}}_2^2-\hat{\mathsf{g}}_3^2)-\hat{\mathsf{g}}_2^n(\hat{\mathsf{g}}_1^2-\hat{\mathsf{g}}_3^2)+\hat{\mathsf{g}}_3^n(\hat{\mathsf{g}}_1^2-\hat{\mathsf{g}}_2^2)}{\hat{\mathsf{g}}_1^2(\hat{\mathsf{g}}_2-\hat{\mathsf{g}}_3)-\hat{\mathsf{g}}_2^2(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_3)+\hat{\mathsf{g}}_3^2(\hat{\mathsf{g}}_1-\hat{\mathsf{g}}_2)}, & |\xi\Delta x| \neq 0, \\ -\frac{1}{2}((-1)^n - 1), & |\xi\Delta x| = 0. \end{cases}$$

As before

$$\hat{\mathsf{G}}_1^{[n]}(\xi\Delta x) = \tfrac{1-(-1)^n}{2} + ((-1)^n - 1) \left\lfloor \tfrac{n}{2} \right\rfloor i\mathcal{C}\xi\Delta x + O(n^2)(\xi\Delta x)^2 + O(|\xi\Delta x|^3).$$

Remark that the first two terms in this expansions are zero whenever n is even. This means that the solution at even time steps experiences—in the low frequency limit—a very reduced influence of the first initialization scheme.

To finish this part, the previous arguments show that there exist constants $C_1, C_2 > 0$ and $C_3 > 0$ such that, for every $|\xi\Delta x| \leq \pi$ and $n \geq Q$

$$|\hat{\mathsf{G}}_2^{[n]}(\xi\Delta x)| \leq C_1 n + C_2, \qquad |\hat{\mathsf{G}}_1^{[n]}(\xi\Delta x)| \leq C_3. \tag{20}$$

2.2.4 Convergence

Theorem 1 (Convergence of (4))**.** *Let the CFL condition $|\mathcal{C}| < 1/2$ be satisfied. Let $u^\circ \in H^{\omega+1}(\mathbb{R})$, where $\omega = 4$ is the order of the bulk scheme and $\omega_2, \omega_1 \geq 0$ are the orders of accuracy of the second and first initialization schemes. Start the discrete scheme using $\mathsf{u}^0 = \mathcal{E}u^\circ$, where $\mathcal{E} : L^2(\mathbb{R}) \to \ell^2(\Delta x\mathbb{Z})$ such that $(\mathcal{E}u)_k = u(x_k)$ for $k \in \mathbb{Z}$. Then, there exists $C(t^n, u^\circ, \omega, \omega_2, \omega_1) > 0$ such that, for Δx small enough, the leading-order error estimate for (4) reads:*

$$\|\mathcal{E}u(t^n) - \mathsf{u}^n\|_{\ell^2,\Delta x} \leq C(t^n, u^\circ, \omega, \omega_2, \omega_1)\Delta x^{\min(\omega,\omega_2,\omega_1+1)},$$

where

$$\overline{\sigma} = \begin{cases} \omega_2 + 1, & \text{if} \quad \min(\omega,\omega_2) < \min(\omega,\omega_1) + 1, \quad \omega_2 < \omega, & \text{(I)}, \\ \omega + 1, & \text{if} \quad \min(\omega,\omega_2) < \min(\omega,\omega_1) + 1, \quad \omega_2 \geq \omega, & \text{(II)}, \\ \omega_2 + 1, & \text{if} \quad \min(\omega,\omega_2) = \min(\omega,\omega_1) + 1, \quad \omega_2 < \omega, & \text{(III)}, \\ \omega + 1, & \text{if} \quad \min(\omega,\omega_2) = \min(\omega,\omega_1) + 1, \quad \omega_2 \geq \omega, & \text{(IV)}, \\ \omega_1 + 1, & \text{if} \quad \min(\omega,\omega_2) > \min(\omega,\omega_1) + 1, \quad \omega_1 < \omega - 1, & \text{(V)}, \end{cases}$$

and

$$C(t^n, u^\circ, \omega, \omega_2, \omega_1) = \begin{cases} (Ct^n + C)|u^\circ|_{H^{\overline{\sigma}}}, & \text{(I)}, \\ (Ct^n + C)|u^\circ|_{H^{\overline{\sigma}}}, & \text{(II)}, \\ (C(t^n)^2|u^\circ|^2_{H^{\overline{\sigma}}} + Ct^n|u^\circ|^2_{H^{\overline{\sigma}-1/2}} + C|u^\circ|^2_{H^{\overline{\sigma}-1}})^{1/2}, & \text{(III)}, \\ (C(t^n)^2|u^\circ|^2_{H^{\overline{\sigma}}} + Ct^n|u^\circ|^2_{H^{\overline{\sigma}-1/2}} + C|u^\circ|^2_{H^{\overline{\sigma}-1}})^{1/2}, & \text{(IV)}, \\ C|u^\circ|_{H^{\overline{\sigma}}}, & \text{(V)}, \end{cases}$$

and $C > 0$ indicates unknown constants that can change at each occurrence.

Remark 4 (On Theorem 1)**.** *We remark the following facts:*

- *The assumption $u^\circ \in H^5(\mathbb{R})$ may be sub-optimal. Especially when the overall order is low due to bad initialization schemes, this order can be observed for initial data which are less than H^5,* cf. *Section 2.4.*
- *The error constants generally trend like $C(t^n, \dots) \sim t^n$ except for (V). When they depend on t^n, this means that the error accumulates at most linearly in time. For (V), the leading-order contribution comes into play at the beginning of the process due to the initialization schemes and cannot be reduced.*

Proof of Theorem 1. The proof is extremely similar to the ones given in [Str04, Chapter 10]: we provide it for the interested reader. Knowing that $u^\circ \in H^5(\mathbb{R})$, the initial function and the exact solution are continuous and their point-wise values are well defined at any time. Using the triangle inequality

$$\|\mathcal{E}u(t^n) - \mathsf{u}^n\|_{\ell^2,\Delta x} \leq \|\mathcal{E}u(t^n) - \mathcal{T}u(t^n)\|_{\ell^2,\Delta x} + \|\mathcal{T}u(t^n) - \mathsf{w}^n\|_{\ell^2,\Delta x} + \|\mathsf{w}^n - \mathsf{u}^n\|_{\ell^2,\Delta x},$$

where the solution w^n is obtained by applying the multi-step scheme and its initializations on the initial datum $\mathsf{w}^0_k = (\mathcal{T}u^\circ)(x_k)$. Here, the truncation operator $\mathcal{T} : L^2(\mathbb{R}) \to \ell^2(\Delta x\mathbb{Z})$ is given by $\widehat{\mathcal{T}u}(\xi) = \hat{u}(\xi)\mathbb{I}_{[0,\pi/\Delta x]}(|\xi|)$. The interpolation operator $\mathcal{S} : \ell^2(\Delta x\mathbb{Z}) \to L^2(\mathbb{R})$ is given by $\widehat{\mathcal{S}\mathsf{u}}(\xi) = \hat{\mathsf{u}}(\xi)\mathbb{I}_{[0,\pi/\Delta x]}(|\xi|)$. For the first term, by [Str04, Theorem 10.1.3]

$$\|\mathcal{E}u(t^n) - \mathcal{T}u(t^n)\|_{\ell^2,\Delta x} \leq C\Delta x^5|u^\circ|_{H^5}. \tag{21}$$

For the last term, we have

$$\|\mathsf{w}^n - \mathsf{u}^n\|^2_{\ell^2,\Delta x} = \int_{|\xi\Delta x|\leq\pi} |\hat{\mathsf{g}}^{[n]}(\xi\Delta x)(\widehat{\mathcal{T}u^\circ}(\xi) - \widehat{\mathcal{E}u^\circ}(\xi))|^2\mathrm{d}\xi \leq C_{\mathrm{s}}^2 \int_{|\xi\Delta x|\leq\pi} |\widehat{\mathcal{T}u^\circ}(\xi) - \widehat{\mathcal{E}u^\circ}(\xi)|^2\mathrm{d}\xi$$
$$\leq C_{\mathrm{s}}^2 C\Delta x^5|u^\circ|^2_{H^5}, \tag{22}$$

where the stability (17) and [Str04, Theorem 10.1.3] have been used. For the central term:

$$\|\mathcal{T}u(t^n) - \mathsf{w}^n\|^2_{\ell^2,\Delta x} = \int_{|\xi\Delta x|\leq\pi} |\hat{u}(t^n,\xi) - \hat{\mathsf{w}}^n(\xi)|^2\mathrm{d}\xi = \int_{\mathbb{R}} |\hat{u}(t^n,\xi) - \widehat{\mathcal{S}\mathsf{w}^n}(\xi)|^2\mathrm{d}\xi - \int_{|\xi\Delta x|>\pi} |\hat{u}(t^n,\xi)|^2\mathrm{d}\xi$$
$$\leq \|u(t^n,\cdot) - \mathcal{S}\mathsf{w}^n\|^2_{L^2(\mathbb{R})} = \int_{|\xi\Delta x|\leq\pi} |e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}(\xi\Delta x)|^2|\hat{u}^\circ(\xi)|^2\mathrm{d}\xi + \int_{|\xi\Delta x|>\pi} |e^{-in\mathcal{C}\xi\Delta x}|^2|\hat{u}^\circ(\xi)|^2\mathrm{d}\xi,$$

thanks to the Parseval's identity. We have

$$
\begin{aligned}
|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}| &\leq |e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}_1^n| + |\hat{\mathsf{G}}_2^{[n]}||\hat{\mathsf{g}}^{[2]} - \hat{\mathsf{g}}_1^2| + |\hat{\mathsf{G}}_1^{[n]}||\hat{\mathsf{g}}^{[1]} - \hat{\mathsf{g}}_1| \\
&\leq n|e^{-i\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}_1| + (C_1 n + C_2)|\hat{\mathsf{g}}^{[2]} - \hat{\mathsf{g}}_1^2| + C_3|\hat{\mathsf{g}}^{[1]} - \hat{\mathsf{g}}_1| \\
&\leq C_4 n|\xi\Delta x|^{\omega+1} + C_5(C_1 n + C_2)|\xi\Delta x|^{\min(\omega,\omega_2)+1} + C_3C_6|\xi\Delta x|^{\min(\omega,\omega_1)+1} \\
&= \underbrace{\tfrac{C_5}{\lambda} t^n \Delta x^{\omega}|\xi|^{\omega+1}}_{\text{bulk scheme}} + \underbrace{C_6(C_2\tfrac{t^n}{\lambda}\Delta x^{\min(\omega,\omega_2)} + C_3\Delta x^{\min(\omega,\omega_2)+1})|\xi|^{\min(\omega,\omega_2)+1}}_{\text{2nd initialization scheme}} + \underbrace{C_4C_7\Delta x^{\min(\omega,\omega_1)+1}|\xi|^{\min(\omega,\omega_1)+1}}_{\text{1st initialization scheme}}.
\end{aligned}
\tag{23}
$$

The first inequality comes from the triangle inequality applied using (18). The second one uses [Str04, Equation (10.1.7)] and the found dependence of the Green functions in n, see (20). The third inequality comes from the order of the schemes, *cf.* [CL20]. When taking the square of the previous inequality back into the integral, all the stemming Sobolev semi-norms exist thanks to the smoothness assumption on the initial datum. It is therefore time to let $\Delta x \to 0$ and identify which term is the leading order term. This can be done directly on the global truncation error term $|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}|$. We distinguish all the cases

- $\min(\omega,\omega_2) < \min(\omega,\omega_1) + 1$: the second initialization scheme is the limiting one.

  (I) $\min(\omega,\omega_2) < \omega$, equivalently $\omega_2 < \omega$. In this case, which covers $\langle 1,1\rangle$, the leading order term in terms of Δx will be

  $$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}| \leq C_2C_6\tfrac{t^n}{\lambda}\Delta x^{\min(\omega,\omega_2)}|\xi|^{\min(\omega,\omega_2)+1} = C_2C_6\tfrac{t^n}{\lambda}\Delta x^{\omega_2}|\xi|^{\omega_2+1}.$$

  (II) $\min(\omega,\omega_2) = \omega$, equivalently $\omega_2 \geq \omega$, covering the case $\langle 4,4\rangle$:

  $$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}| \leq \tfrac{C_5+C_2C_6}{\lambda} t^n\Delta x^{\omega}|\xi|^{\omega+1}.$$

- $\min(\omega,\omega_2) = \min(\omega,\omega_1) + 1$: both initialization scheme contribute equally.

  (III) $\min(\omega,\omega_2) = \min(\omega,\omega_1)+1 < \omega$, *i.e.* $\omega_2 < \omega$ and $\omega_1 < \omega - 1$ (this latter condition is redundant), which covers $\langle 1,2\rangle$. We have

  $$\begin{aligned}|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}| &\leq (C_2C_6\tfrac{t^n}{\lambda}|\xi|^{\min(\omega,\omega_2)+1} + C_4C_7|\xi|^{\min(\omega,\omega_2)})\Delta x^{\min(\omega,\omega_2)} \\ &= (C_2C_6\tfrac{t^n}{\lambda}|\xi|^{\omega_2+1} + C_4C_7|\xi|^{\omega_2})\Delta x^{\omega_2}.\end{aligned}$$

  (IV) $\min(\omega,\omega_2) = \min(\omega,\omega_1)+1 = \omega$, *i.e.* $\omega_2 \geq \omega$ and $\omega_1 \geq \omega - 1$ (this latter condition is redundant), which covers $\langle 3,4\rangle$. We have

  $$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}| \leq (\tfrac{C_5+C_2C_6}{\lambda} t^n|\xi|^{\omega+1} + C_4C_7|\xi|^{\omega})\Delta x^{\omega}.$$

- $\min(\omega,\omega_2) > \min(\omega,\omega_1) + 1$: the first initialization scheme is the limiting one. The only possible case is $\min(\omega,\omega_1) + 1 < \omega$, equivalently $\omega_1 < \omega - 1$, indicated by (V). In this case, which covers $\langle 1,3\rangle$:

  $$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}| \leq C_4C_7\Delta x^{\min(\omega,\omega_1)+1}|\xi|^{\min(\omega,\omega_1)+1} = C_4C_7\Delta x^{\omega_1+1}|\xi|^{\omega_1+1}.$$

This shows that the order in Δx is given by $\Delta x^{\min(\omega,\omega_2,\omega_1+1)}$. We introduce

$$
\overline{\sigma} = \begin{cases}
\omega_2 + 1, & \text{if} \quad \min(\omega,\omega_2) < \min(\omega,\omega_1) + 1, \quad \omega_2 < \omega, & \text{(I)},\\
\omega + 1, & \text{if} \quad \min(\omega,\omega_2) < \min(\omega,\omega_1) + 1, \quad \omega_2 \geq \omega, & \text{(II)},\\
\omega_2 + 1, & \text{if} \quad \min(\omega,\omega_2) = \min(\omega,\omega_1) + 1, \quad \omega_2 < \omega, & \text{(III)},\\
\omega + 1, & \text{if} \quad \min(\omega,\omega_2) = \min(\omega,\omega_1) + 1, \quad \omega_2 \geq \omega, & \text{(IV)},\\
\omega_1 + 1, & \text{if} \quad \min(\omega,\omega_2) > \min(\omega,\omega_1) + 1, \quad \omega_1 < \omega - 1, & \text{(V)}.
\end{cases}
$$

Observe that $\overline{\sigma} \geq \min(\omega,\omega_2,\omega_1+1)$. By [Str04, Equation (10.1.10)], we have that $\int_{|\xi\Delta x|>\pi} |e^{-in\mathcal{C}\xi\Delta x}|^2|\hat{u}^\circ(\xi)|^2 \mathrm{d}\xi \leq C\Delta x^5|u^\circ|^2_{H^5}$. We gain

$$\|\mathcal{T}u(t^n) - \mathsf{w}^n\|_{\ell^2,\Delta x} \leq C(t^n, u^\circ, \omega, \omega_2, \omega_1)\Delta x^{\min(\omega,\omega_2,\omega_1+1)},$$

where, by indicating all the constants by $C > 0$ (each one is different):

$$C(t^n, u^\circ, \omega, \omega_2, \omega_1) = \begin{cases} Ct^n |u^\circ|_{H^\sigma}, & \text{(I)}, \\ Ct^n |u^\circ|_{H^\sigma}, & \text{(II)}, \\ (C(t^n)^2 |u^\circ|^2_{H^\sigma} + Ct^n |u^\circ|^2_{H^{\sigma-1/2}} + C|u^\circ|^2_{H^{\sigma-1}})^{1/2}, & \text{(III)}, \\ (C(t^n)^2 |u^\circ|^2_{H^\sigma} + Ct^n |u^\circ|^2_{H^{\sigma-1/2}} + C|u^\circ|^2_{H^{\sigma-1}})^{1/2}, & \text{(IV)}, \\ C|u^\circ|_{H^\sigma}, & \text{(V)}. \end{cases}$$

The overall claim comes adding (21) and (22), which are nevertheless negligible. □

2.3 Error behavior in time

The behavior of the error of actual numerical computations as time goes on can be studied using the modal decomposition (15): for every $|\xi\Delta x| \leq \pi$, we have that

$$\hat{\mathsf{g}}^{[n]}(\xi\Delta x) = \begin{cases} \hat{\sigma}_1(\xi\Delta x)\hat{\mathsf{g}}_1(\xi\Delta x)^n + \hat{\sigma}_2(\xi\Delta x)\hat{\mathsf{g}}_2(\xi\Delta x)^n + \hat{\sigma}_3(\xi\Delta x)\hat{\mathsf{g}}_3(\xi\Delta x)^n, & |\xi\Delta x| \neq 0, \\ \tilde{\sigma}_1 + \tilde{\sigma}_2(-1)^n + \tilde{\sigma}_3(-1)^n n, & |\xi\Delta x| = 0. \end{cases} \tag{24}$$

Here, the coefficients $\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3$, and $\tilde{\sigma}_1, \tilde{\sigma}_2, \tilde{\sigma}_3 \in \mathbb{R}$ are determined using the initialization schemes $\hat{\mathsf{g}}^{[2]}$ and $\hat{\mathsf{g}}^{[1]}$. For the initialization schemes that we have considered, $\tilde{\sigma}_1 = 1$ and $\tilde{\sigma}_2 = \tilde{\sigma}_3 = 0$, *cf.* (17). We remind that, contrarily to what this explicit formula (24) suggests, $\hat{\mathsf{g}}^{[n]}(\xi\Delta x)$ is a smooth function for every $|\xi\Delta x| \leq \pi$. In order to find its Taylor expansion for $|\xi\Delta x| \ll 1$, we can rely on its explicit representation for $|\xi\Delta x| \neq 0$. This gives the system, for $|\xi\Delta x| \neq 0$: $\hat{\mathbf{V}}(\xi\Delta x)(\hat{\sigma}_1(\xi\Delta x), \hat{\sigma}_2(\xi\Delta x), \hat{\sigma}_3(\xi\Delta x))^\mathsf{T} = (\hat{\mathsf{g}}^{[2]}(\xi\Delta x), \hat{\mathsf{g}}^{[1]}(\xi\Delta x), 1)^\mathsf{T}$. We can inverse the Vandermonde matrix to give $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\sigma}_3$. Let us study some given choice of initialization scheme.

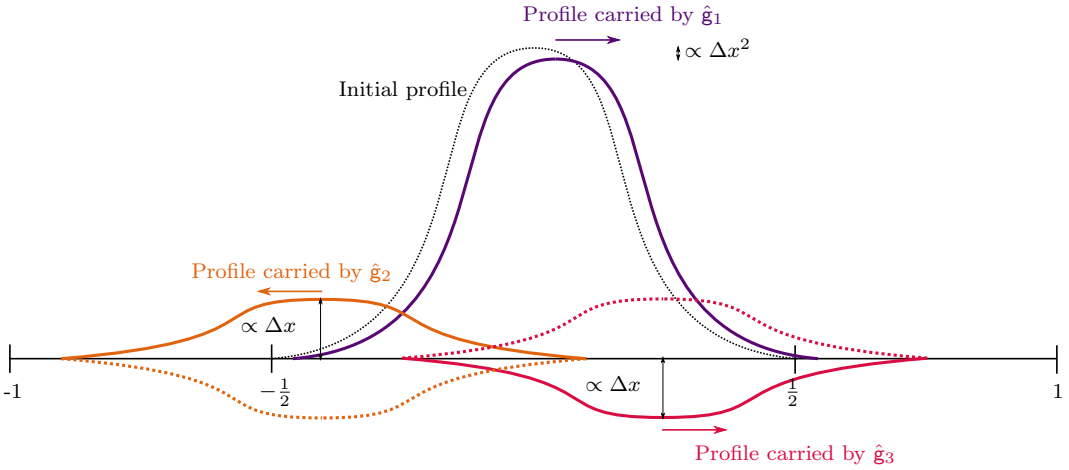

Figure 3: Packets propagated by different modes for the case $\langle 1, 2\rangle$. For the ones carried by $\hat{\mathsf{g}}_2$ and $\hat{\mathsf{g}}_3$, we draw an envelope-like shape to highlight the rapidly oscillating nature between odd and even time-steps.

- Initialization $\langle 1, 2\rangle$. We obtain, in the limit $|\xi\Delta x| \ll 1$, $\hat{\sigma}_1(\xi\Delta x) = 1 + O(|\xi\Delta x|^2)$, $\hat{\sigma}_2(\xi\Delta x) = O(|\xi\Delta x|)$, and $\hat{\sigma}_3(\xi\Delta x) = O(|\xi\Delta x|)$. This results in the decomposition of the discrete solution—see also Figure 3—as

$$\begin{aligned} \hat{\mathsf{g}}^{[n]}(\xi\Delta x) = {} & (1 + O(|\xi\Delta x|^2))e^{-iVt^n\xi(1+O(|\xi\Delta x|^4))} \\ & + \Big(\frac{i\sqrt{3}(1-\mathcal{C}^2)}{2\sqrt{8-5\mathcal{C}^2}}\xi\Delta x + O(|\xi\Delta x|^2)\Big)(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V+\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))} \\ & + \Big(-\frac{i\sqrt{3}(1-\mathcal{C}^2)}{2\sqrt{8-5\mathcal{C}^2}}\xi\Delta x + O(|\xi\Delta x|^2)\Big)(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V-\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))}. \end{aligned}$$

  The physical mode—which is accurate at order four—is present with a distortion of order two, see first row. What lowers the overall order to one are the rapidly oscillating spurious modes which have amplitude $O(|\xi\Delta x|)$ (second and third rows).

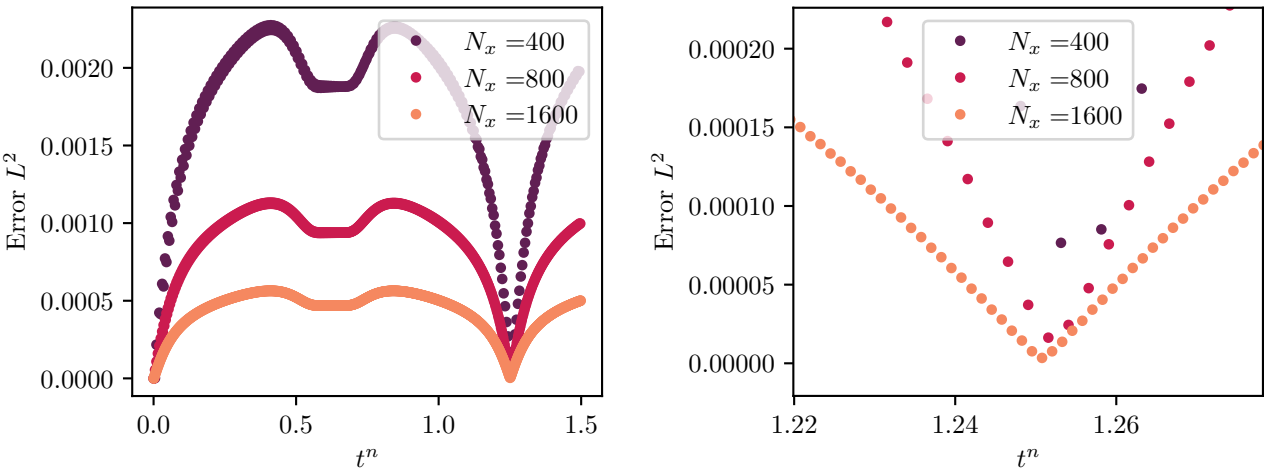

Figure 4: L^2 error in time for $\langle 1, 2\rangle$ using N_x grid points. A zoom around 1.25 is proposed on the right.

Looking at the error in time, see Figure 4, we see that for fixed times, the error is proportional to Δx. However, for a specific time around $t^n = 1.25$, we see that the error seems to be practically zero. Looking at the magnification in this area, we see that here convergence seems quadratic. We explain this spectacular decrease of the error—and enhanced convergence rate—by a destructive interference between the mode brought by $\hat{\mathrm{g}}_2$ and the one by $\hat{\mathrm{g}}_3$, which are those carrying the $O(\Delta x)$ part of the error, see Figure 3. At $\mathcal{C} = 1/4$, the dimensionless velocities for each mode are

$$\begin{aligned} \hat{\mathrm{g}}_1 \quad &\leftrightarrow \quad \mathcal{C}_1 = \mathcal{C} = 0.25, \qquad \hat{\mathrm{g}}_2 \quad \leftrightarrow \quad \mathcal{C}_2 = -\tfrac{\sqrt{3}}{6}(\sqrt{3}\mathcal{C} + \sqrt{8 - 5\mathcal{C}^2}) \approx -0.93, \\ &\qquad\qquad\qquad\quad \hat{\mathrm{g}}_3 \quad \leftrightarrow \quad \mathcal{C}_3 = -\tfrac{\sqrt{3}}{6}(\sqrt{3}\mathcal{C} - \sqrt{8 - 5\mathcal{C}^2}) \approx 0.68. \end{aligned}$$

The initial envelope has support $\mathrm{supp}(u^\circ) = [-1/2, 1/2]$. We consider that the maximal interference between this envelope transported by the mode $\hat{\mathrm{g}}_2$ and the one transported by $\hat{\mathrm{g}}_3$ takes place when the peaks, located at $x = 0$ at $t = 0$, meet again due to the periodic boundary conditions. In this case, the two symmetric packets coincide. The first peak to reach the boundary ($x = -1$) is the one transported by $\hat{\mathrm{g}}_2$, since $|\mathcal{C}_2| > |\mathcal{C}_3|$. The peaks met again at time

$$\frac{2}{|\mathcal{C}_2| + |\mathcal{C}_3|} \approx 1.25,$$

which is—unsurprisingly—the one where the error had its minimum in time. If we repeat the convergence test for different initializations selecting precisely $T = 1.25$, we obtain the result in Figure 5. Now we observe overall order $= \min(\omega, \omega_2 + 1, \omega_1 + 1)$, as in the genuinely stable framework, thanks to the fact that the parasitic modes, carrying the dominant part of the error, cancel out thanks to periodicity. When the packets are located at the same place thanks to periodicity, the sum of the modes yields a term $O(|\xi\Delta x|^2)\hat{\mathrm{u}}^0(\xi)$, which is second-order in Δx, giving the destructive interference. In order to interpret the time behavior of the error more closely, we observe that the time where the leftmost point of $\mathrm{supp}(u^\circ)$, namely $-1/2$, transported by $\hat{\mathrm{g}}_2$, and the rightmost point of $\mathrm{supp}(u^\circ)$, namely $1/2$, transported by $\hat{\mathrm{g}}_3$ merge is at time $t = 1/(|\mathcal{C}_2| + |\mathcal{C}_3|) \approx 0.62$, which is another remarkable time on Figure 4. In this figure, the articulate pattern of the error is made up of the interactions of the different waves/modes sustained by the numerical scheme also due to the periodic boundary conditions.

By writing the global truncation error coefficient in the low-frequency limit:

$$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathrm{g}}^{[n]}(\xi\Delta x)|^2 = \begin{cases} \left(\frac{n}{2}\right)^2(\mathcal{C}^2 - 1)^2(\xi\Delta x)^4 + O(|\xi\Delta x|^6), & n \text{ even}, \\ \left(\frac{n-1}{2}\right)^2(\mathcal{C}^2 - 1)^2(\xi\Delta x)^4 + O(|\xi\Delta x|^6), & n \text{ odd}. \end{cases}$$

we see that even and odd steps behave essentially in the same way, as shown in Figure 4. The fixed $-1/2$ term is what remains of the initialization schemes.

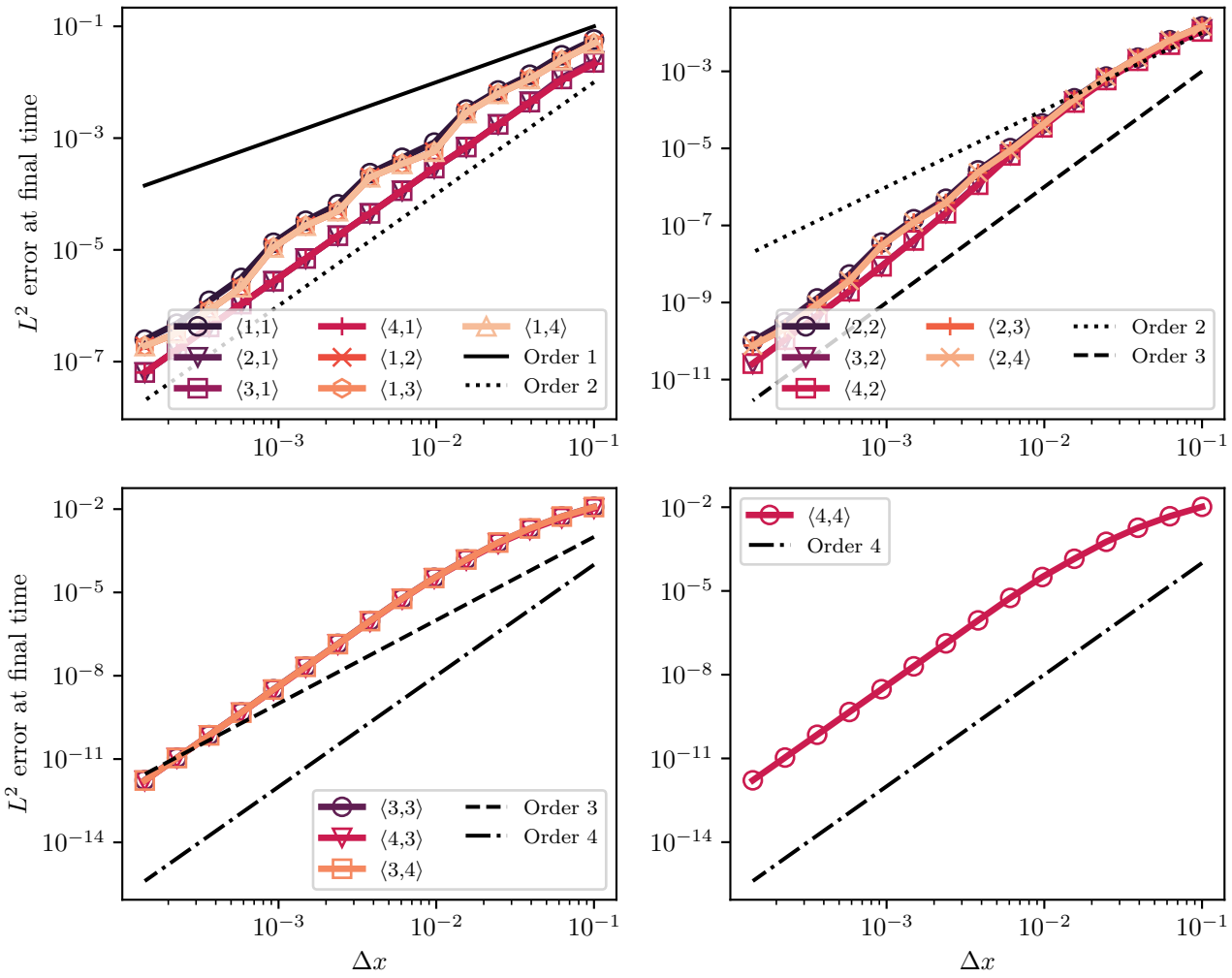

Figure 5: Error for (4) at final time $T = \frac{2}{|\mathcal{C}_2|+|\mathcal{C}_3|} \approx 1.25$ with different initialization schemes.

- Initialization $\langle 2,1\rangle$. The modal decomposition for $|\xi\Delta x| \ll 1$ is:

$$\hat{\mathsf{g}}^{[n]}(\xi\Delta x) = (1 + \tfrac{1}{4}(\mathcal{C}^2-1)(\xi\Delta x)^2 + O(|\xi\Delta x|^3))e^{-iVt^n\xi(1+O(|\xi\Delta x|^4))}$$
$$+ (-i\sqrt{3}(\mathcal{C}^2-1)(19\mathcal{C}+\sqrt{3}\sqrt{8-5\mathcal{C}^2})(\xi\Delta x)^2 + O(|\xi\Delta x|^3))(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V+\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))}$$
$$+ (-i\sqrt{3}(\mathcal{C}^2-1)(-19\mathcal{C}+\sqrt{3}\sqrt{8-5\mathcal{C}^2})(\xi\Delta x)^2 + O(|\xi\Delta x|^3))(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V-\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))}.$$

The physical mode and the parasitic modes are all present with a distortion of order two. Now, the coefficients of the $O(|\xi\Delta x|^2)$ perturbation of the parasitic modes are not one the opposite of the other, hence do not totally cancel out when the modes meet again when periodic boundary conditions are imposed, see Figure 6 for $t^n \approx 1.25$. We also have

$$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathsf{g}}^{[n]}(\xi\Delta x)|^2 = \begin{cases} \frac{n^2}{36}\mathcal{C}^2(\mathcal{C}^2-1)^2(\xi\Delta x)^6 + O(|\xi\Delta x|^8), & n \text{ even}, \\ \frac{1}{4}(\mathcal{C}^2-1)^2(\xi\Delta x)^4 + \alpha_{\mathcal{C}}(n)(\xi\Delta x)^6 + O(|\xi\Delta x|^8), & n \text{ odd}, \end{cases}$$

where $\alpha_{\mathcal{C}}(n) = O(n^2)$ or more explicitly $\alpha_{\mathcal{C}}(n) = (\frac{n^2}{9} - \frac{17n}{72} + \frac{1}{9})\mathcal{C}^6 + (-\frac{11n^2}{36} + \frac{23n}{36} - \frac{25}{72})\mathcal{C}^4 + (\frac{5n^2}{18} - \frac{41n}{72} + \frac{13}{36})\mathcal{C}^2 + (-\frac{n^2}{12} + \frac{n}{6} - \frac{1}{8})$. Only the odd steps carry the fixed term being the remaining trace of the Lax-Friedrichs scheme for $n = 1$. Moreover, we see that the behavior of the error is radically different, as visible in Figure 6, between even and odd steps. The former typically carry smaller errors. The initial decreasing behavior of the error for odd steps can be understood by noticing that the function $\alpha_{1/4}(n)$ decreases in n.

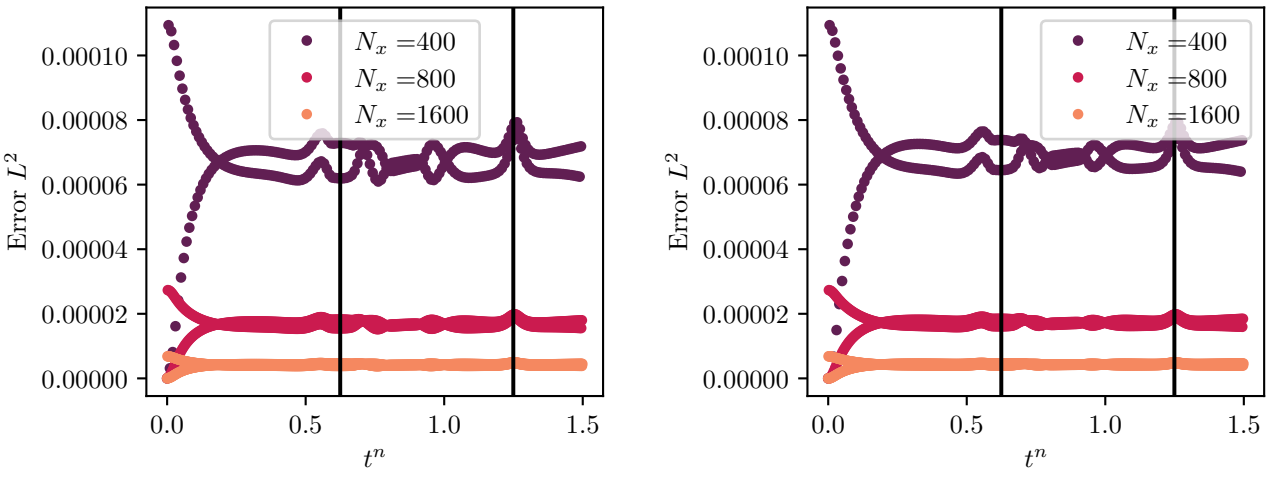

Figure 6: L^2 error in time for $\langle 2,1\rangle$ (left) and $\langle 3,1\rangle$ (right) using N_x grid points.

- Initialization $\langle 3,1\rangle$. The conclusions are the same as $\langle 2,1\rangle$. The expression for $|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathrm{g}}^{[n]}(\xi\Delta x)|^2$ in the even cases is even more different from the one in the odd cases compared to $\langle 1,2\rangle$.
- Initialization $\langle 3,4\rangle$. The modal decomposition is, for $|\xi\Delta x| \ll 1$:

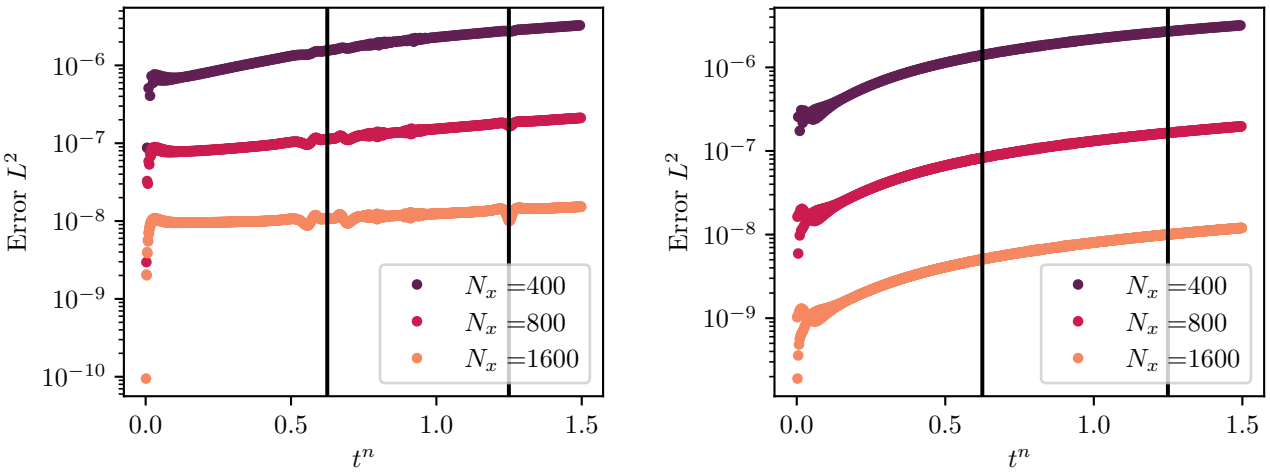

Figure 7: L^2 error in time for $\langle 3,4\rangle$ (left) and $\langle 4,3\rangle$ (right) using N_x grid points.

$$\begin{aligned}\hat{\mathrm{g}}^{[n]}(\xi\Delta x) = (1 - \tfrac{\mathcal{C}(\mathcal{C}^2-1)(\mathcal{C}-2)}{48}(\xi\Delta x)^4 + O(|\xi\Delta x|^5))e^{-iVt^n\xi(1+O(|\xi\Delta x|^4))}\\ + (\tfrac{i\mathcal{C}(\mathcal{C}^2-1)(\mathcal{C}-2)}{24\sqrt{8-5\mathcal{C}^2}}(\xi\Delta x)^3 + O(|\xi\Delta x|^4))(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V+\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))}\\ + (-\tfrac{i\mathcal{C}(\mathcal{C}^2-1)(\mathcal{C}-2)}{24\sqrt{8-5\mathcal{C}^2}}(\xi\Delta x)^3 + O(|\xi\Delta x|^4))(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V-\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))}.\end{aligned}$$

The physical mode is present with a distortion of order four: this is not what limits the order. What lowers the overall order to three is are the rapidly oscillating spurious modes which have amplitude $O(|\xi\Delta x|^3)$. However, for periodic boundary conditions, we see that the scheme converges at order four for a final time $T \approx 1.25$ (small basins on Figure 7), thanks to the cancellation of spurious modes. Also

$$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathrm{g}}^{[n]}(\xi\Delta x)|^2 = \begin{cases} \left(\frac{\mathcal{C}(\mathcal{C}^2-1)(\mathcal{C}-2)}{24}n\right)^2(\xi\Delta x)^8 + O(|\xi\Delta x|^{10}), & n \text{ even},\\ \left(\frac{\mathcal{C}(\mathcal{C}^2-1)(\mathcal{C}-2)}{24}(n-1)\right)^2(\xi\Delta x)^8 + O(|\xi\Delta x|^{10}), & n \text{ odd},\end{cases}$$

shows that—*cf.* Figure 7—even and odd steps behave essentially in the same way.

- Initialization $\langle 4,3\rangle$. The modal decomposition is

$$\hat{\mathrm{g}}^{[n]}(\xi\Delta x) = (1 - \tfrac{\mathcal{C}(\mathcal{C}^2-1)(\mathcal{C}-2)}{48}(\xi\Delta x)^4 + O(|\xi\Delta x|^5))e^{-iVt^n\xi(1+O(|\xi\Delta x|^4))} + O(|\xi\Delta x|^4)(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V+\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))} + O(|\xi\Delta x|^4)(-1)^n e^{\frac{i\sqrt{3}t^n}{6}(\sqrt{3}V-\sqrt{8\lambda^2-5V^2})\xi(1+O(|\xi\Delta x|^2))},$$

where the coefficients of the spurious waves (not given for the sake of compactness) are not one the opposite of the other even at leading order. Also

$$|e^{-in\mathcal{C}\xi\Delta x} - \hat{\mathrm{g}}^{[n]}(\xi\Delta x)|^2 = \begin{cases} \alpha_{\mathcal{C}}(n)(\xi\Delta x)^{10} + O(|\xi\Delta x|^{12}), & n \text{ even}, \\ \left(\frac{\mathcal{C}(\mathcal{C}^2-1)(\mathcal{C}-2)}{24}\right)^2(\xi\Delta x)^8 + \beta_{\mathcal{C}}(n)(\xi\Delta x)^{10} + O(|\xi\Delta x|^{12}), & n \text{ odd}, \end{cases}$$

where $\alpha_{\mathcal{C}}(n) = O(n^2)$ and $\beta_{\mathcal{C}}(n) = O(n^2)$.

2.4 Convergence with non-smooth initial data

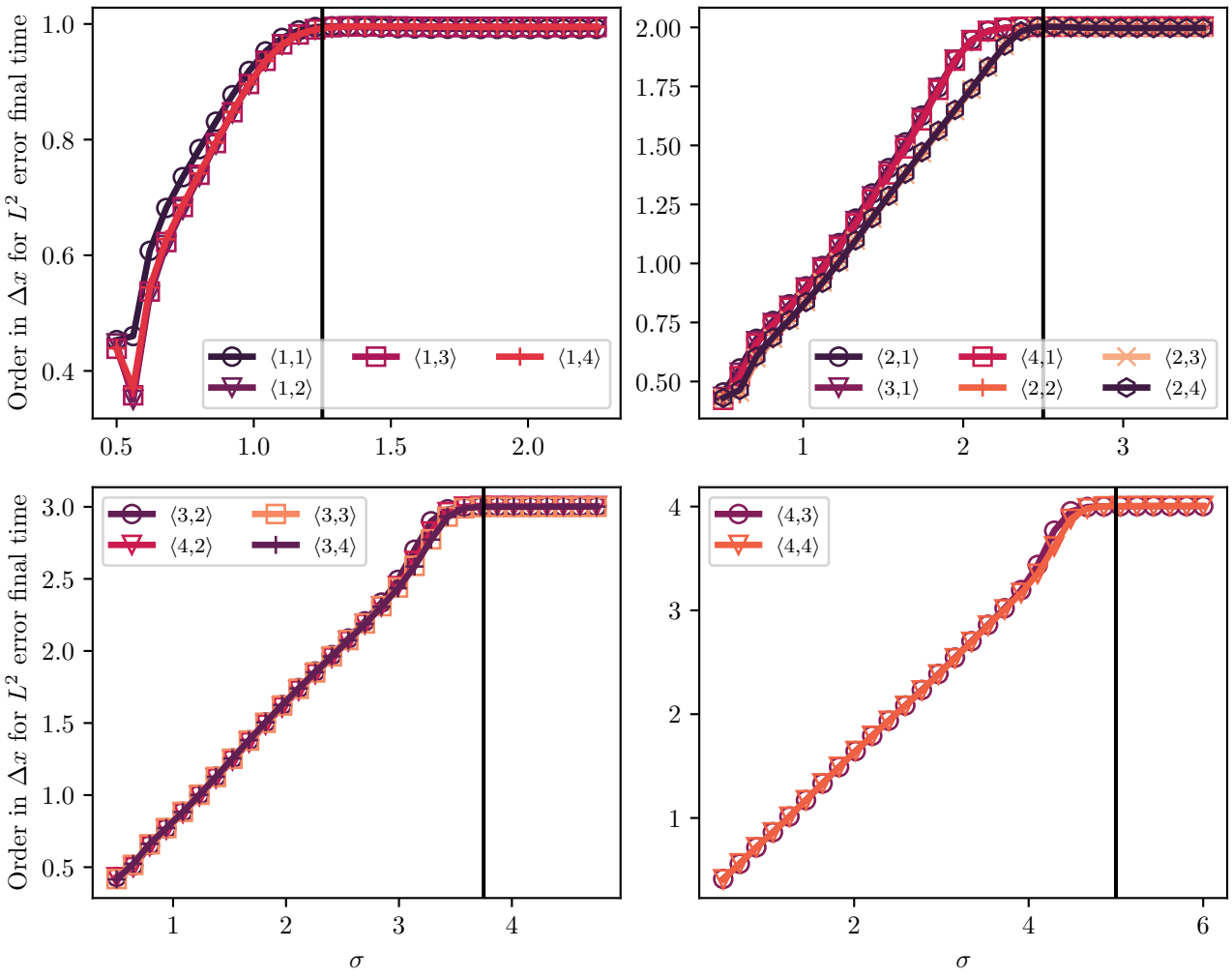

Figure 8: Convergence rates in Δx for the L^2 error at final time $T = 0.2$, as function of σ when $u^\circ \in H^{\sigma-}(\mathbb{R})$. The black vertical line corresponds to $\sigma = \frac{5}{4}\min(\omega, \omega_2, \omega_1 + 1)$.

In the past, orders of convergence for non-smooth initial data were investigated in [BT70, BTW75, Str04, Cou17] for genuinely stable schemes. As previously pointed out, the assumption that $u^\circ \in H^{\omega+1}(\mathbb{R})$ is often too strong—especially when the overall order $\min(\omega, \omega_2, \omega_1+1)$ is way smaller than ω—to actually observe the expected order. Indeed, this order can be observed for non-smooth initial data being less that H^5. We now perform a numerical simulation to verify this fact. The conditions are, as usual, domain $[-1,1]$ endowed with periodic boundary conditions, $\mathcal{C} = 1/4$, and $T = 0.2$. We consider the initial functions $u^\circ(x) = \cos(\pi x)^{\sigma-1/2}\mathbb{I}_{(-1,1)}(2x)$ for $\sigma \geq 1$, inspired by [BTW75, Example II, Section 2.4]. These are such that $u^\circ \in B^{\sigma,2}_\infty(\mathbb{R})$ in terms of Besov spaces or $u^\circ \in H^{\sigma-}(\mathbb{R})$ in terms of Sobolev spaces. We measure the convergence rate in Δx with respect to the L^2 norm of the error at final time, obtaining the results in Figure 8. When the overall scheme is less than fourth-order accurate (maximal order), the order saturates before the smoothness $u^\circ \in H^5(\mathbb{R})$ for the initial datum is reached. The value of σ from which the order saturates can be estimated using [Str04, Corollary 10.3.2] (notice that the error estimates in this result contain rarely observed logarithmic terms in Δx) by $\sigma \approx \frac{5}{4}\min(\omega, \omega_2, \omega_1+1)$. This is obtained by intersecting the line $\sigma \mapsto \frac{4}{5}\sigma$, associated with a fourth-order (bulk) scheme, with the fact that the order is limited by $\min(\omega, \omega_2, \omega_1+1)$. This is the intuitive best approximation of the minimal smoothness that we were able to find. Slight deviations from this behavior can be understood by looking at (18) or (23). For the selected final time, there might still be an important (yet difficult to describe) effect of the part of the solution associated with $\hat{\mathsf{G}}_2^{[n]}$ and $\hat{\mathsf{G}}_1^{[n]}$. If we take larger T (simulations not presented in the paper), the term brought by $\hat{\mathsf{g}}_1^n$ becomes dominant and the observed behavior adhere better and better to the estimate according to which the order is $\min(\min(\omega, \omega_2, \omega_1+1), \frac{4}{5}\sigma)$ for $u^\circ \in H^{\sigma-}(\mathbb{R})$.

3 Role of the round-off errors

The discussion that has been developed hitherto has been conducted as if actual numerical simulations were done at arbitrary machine precision, without being affected by round-off errors. However, the instability cancellation elucidated in Section 2.2.2 could not actually take place in floating-point arithmetic, due to round-off errors. As pointed out by [Tre92], this does however not question the interest of the previous discussions. Quite the opposite, this gives one more reason why, besides the order of consistency/convergence, one generally avoids weakly unstable schemes: they could lead unpredictable behaviors in presence of floating point numbers.

We perform the same test as we did in Section 1.2 except for the fact that we use a longer final time of $T = 16$ and fix the number of points in the domain to $N_x = 200$. Computations are carried out using unusually low floating-point precisions, ranging from 10 to 13 bits, to be compared with the usual 64 bits double precision in `Python`. We employ the stable lattice Boltzmann scheme (5)/(6) with (8) using $\delta = 0$, its weakly unstable corresponding Finite Difference scheme (4) with initialization (8) using $\delta = 0$, and the one-step stable OS4 scheme. Notice that the first two procedures are perfectly equivalent "on paper". In Figure 9, we show the solution both at the final and penultimate time step, and its L^2 norm as function of time, which we take as a measure of the instability when applicable.

We observe that for the weakly unstable scheme (4) with initialization (8) with $\delta = 0$, especially for very low floating point accuracy, the solution is totally shifted either upward or downward, according to the parity of the time step n. This is the reason why we decided to show both the solution at the final and penultimate time step. The fact that the solution is approximately shifted by a constant proportional to $(-1)^n n$ stems from the following features of the scheme. The constant nature of the shift is dictated by the unstable mode corresponding to the frequency zero, and thus corresponding to an unstable mode in the physical space under the form of a constant function: by inverse Fourier transform $(2\pi)^{-1/2}\int_{|\xi|\leq\pi} e^{ik\xi}\delta_0(\xi)\mathrm{d}\xi = (2\pi)^{-1/2}$ for every $k \in \mathbb{Z}$. The oscillation trending like $(-1)^n$ comes from the fact that the unstable modes are associated with the spurious roots, which are such that $\hat{\mathsf{g}}_2(0)^n = \hat{\mathsf{g}}_3(0)^n = (-1)^n$. The growth behaving roughly proportionally to n comes from the instability due to the multiple nature of the roots at frequency zero. For the stable OS4 scheme, the simulation remains perfectly controlled and the norm of the solution does not increase even using a very rough floating point arithmetic. The solution for the lattice Boltzmann scheme (5)/(6) with (8) and $\delta = 0$ is completely different from the one of its corresponding Finite Difference scheme and remains totally stable, as it should be. This indicates that the former implementation has a better backward stability compared to the latter. Small oscillations, compared to the OS4 scheme, come from the lack of dissipation of the lattice Boltzmann scheme on the whole spectrum of frequencies.

Remark 5 (On the enhanced stability of the lattice Boltzmann algorithm)**.** *To understand the enhanced stability of* (5)/(6)/(8) *compared to* (4)/(8)*, we come back to the way of turning lattice Boltzmann schemes into Finite Difference ones. In the entire paper, the problematic frequency has been* $\xi = 0$*, where the matrix giving the lattice*

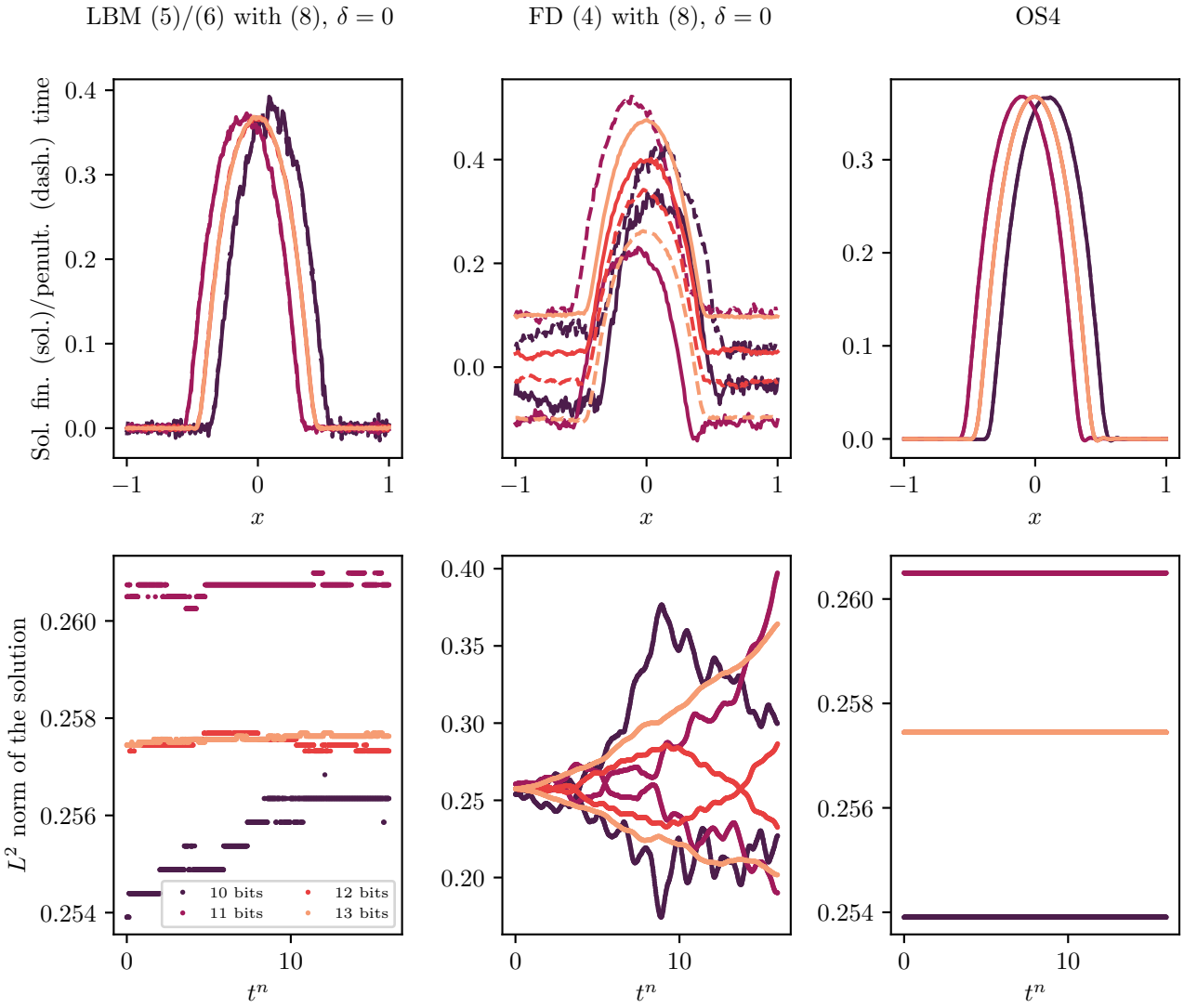

Figure 9: Test varying the floating point precision.

Boltzmann scheme becomes

$$\hat{\mathbf{E}}(0) = \begin{bmatrix} 1 & 0 & 0 \\ 2\mathcal{C} & -1 & 0 \\ 2(\mathcal{C}^2-1) & 0 & -1 \end{bmatrix}, \qquad \text{hence} \qquad \hat{\mathbf{E}}(0)^n = \begin{bmatrix} 1 & 0 & 0 \\ \mathcal{C}(1-(-1)^n) & (-1)^n & 0 \\ (\mathcal{C}^2-1)(1-(-1)^n) & 0 & (-1)^n \end{bmatrix},$$

thus is uniformly power bounded and thus the original lattice Boltzmann scheme genuinely stable, since

$$\hat{\mathsf{u}}^n(0) = \boldsymbol{e}_1^{\mathsf{T}} \hat{\mathbf{E}}(0)^n (\hat{\mathsf{u}}^0(0), \hat{\mathsf{v}}^0(0), \hat{\mathsf{w}}^0(0))^{\mathsf{T}} = \hat{\mathsf{u}}^0(0). \tag{25}$$

Observe that $\hat{\mathbf{E}}(0)^2 = \boldsymbol{I}$*, thus the polynomial* $\mathsf{z}^2 - 1$ *dividing* $\det(\mathsf{z}\boldsymbol{I} - \hat{\mathbf{E}}(0))$ *annihilates* $\hat{\mathbf{E}}(0)$*: it is its minimal polynomial. This difference between characteristic and minimal polynomials is coherent with the fact that* $\hat{\mathbf{E}}(0)$ *is not similar to its companion matrix* $\hat{\mathbf{C}}(0)$*, with the former being diagonalisable and the latter no. As pointed out in [BGM22], the fact that a polynomial annihilates the scheme matrix allows to recover the corresponding Finite Difference scheme through it. However, this should be true for all frequency* ξ*, which is not the case here for* $\mathsf{z}^2 - 1$*. Nevertheless, for the sole unstable frequency* $\xi = 0$ *(or equivalently, constant solutions), the dynamics of the lattice Boltzmann scheme* (5)/(6)*, as far as* u *is concerned, can be rewritten using something simpler (and more stable) than* (4)*, which reads* $\hat{\mathsf{u}}^{n+1}(0) = \hat{\mathsf{u}}^{n-1}(0)$*. This scheme is genuinely stable, since it has two eigenvalues* ± 1 *on the unit circle being distinct. It has to be interpreted as a constraint fulfilled by the solution* u *of* (5)/(6)*. The constraint given by* (4) *is also satisfied, but it yields an overconstrained mechanism, more prone to instabilities when*

the exact compensation resulting in overall stability could not take place, due to floating-point numbers. Pushing this way of rasoning even further, we have observed in [BGM22] that if a polynomial annihilates (again for all ξ) the first row of the scheme matrix, hence it divides its minimal polynomial, it yields a corresponding Finite Difference scheme as well. In our case, for $\hat{\mathbf{E}}(0)$, the polynomial is simply $\mathsf{z}-1$, resulting in $\hat{\mathsf{u}}^{n+1}(0)=\hat{\mathsf{u}}^n(0)$, which is perfectly compatible with (25) and extremely stable.

4 Application: "nested" kinetic schemes for non-linear systems

We now showcase how a fourth-order scheme for the linear transport equation can be used, utilizing a Jin-Xin relaxation system [JX95], as the the basic brick to approximate the solution of a non-linear system of conservation laws [ADN00]. This application is strongly inspired by the work of [CFH$^+$19]. We call the schemes "nested" because we use an external kinetic scheme based on the Jin-Xin relaxation system to deal with the non-linearity and then inner lattice Boltzmann schemes—a special type of kinetic schemes—to solve the transport equations appearing in the Jin-Xin system.

4.1 Jin-Xin relaxation system

We tackle the solution of the system on $\boldsymbol{w}:\mathbb{R}\times\mathbb{R}\to\mathbb{R}^M$ which reads

$$\partial_t\boldsymbol{w}+\partial_x\boldsymbol{\varphi}(\boldsymbol{w})=0, \tag{26}$$

where $\boldsymbol{\varphi}:\mathbb{R}^M\to\mathbb{R}^M$ is a smooth and possibly non-linear flux. To give a simple example, taking $M=1$ and $\varphi(w)=\frac{1}{2}w^2$ gives the inviscid Burgers equation. In order to isolate the non-linearity of the problem into a local relaxation term which is easily tractable, we consider the associated Jin-Xin relaxation system [JX95]

$$\begin{cases}\partial_t\boldsymbol{w}+\partial_x\boldsymbol{z}=0,\\ \partial_t\boldsymbol{z}+V^2\partial_x\boldsymbol{w}=-\frac{1}{\epsilon}(\boldsymbol{z}-\boldsymbol{\varphi}(\boldsymbol{w})),\end{cases} \tag{27}$$

with $V>0$ a kinetic velocity and $\epsilon\geq 0$ a relaxation time, whose formal limit for $\epsilon\to 0$ gives back (26). Using the change of basis $\boldsymbol{w}=\boldsymbol{f}^+ +\boldsymbol{f}^-$ and $\boldsymbol{z}=V(\boldsymbol{f}^+-\boldsymbol{f}^-)$, the relaxation system (27) can be recast into its kinetic form

$$\partial_t\boldsymbol{f}^\pm\pm V\partial_x\boldsymbol{f}^\pm=-\frac{1}{\epsilon}(\boldsymbol{f}^\pm-\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w})),\qquad\text{where}\quad \boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w})=\frac{1}{2}\boldsymbol{w}\pm\frac{1}{2V}\boldsymbol{\varphi}(\boldsymbol{w}). \tag{28}$$

4.2 Splitting and numerical schemes

Notice that the left hand side of (28) is nothing but a linear transport equation at velocities $\pm V$, which can therefore be approximated using the scheme (4) introduced in the first part of the paper or, for efficient computations taking advantage of the peculiar structure of lattice Boltzmann methods, using (5) and (6). We then split (28) into its transport (T) part $\partial_t\boldsymbol{f}^\pm\pm V\partial_x\boldsymbol{f}^\pm=0$, and relaxation part (R) $\partial_t\boldsymbol{f}^\pm=-\frac{1}{\epsilon}(\boldsymbol{f}^\pm-\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}))$. The transport part (T) shall be solved using $\tilde{N}\geq 3$ steps of the fourth-order lattice Boltzmann scheme for the transport equation or its corresponding Finite Difference scheme. The reason to perform more than three steps to reach the desired final time $\Delta t>0$ is dictated by the fact that we want the multi-step method to produce fourth-order accurate approximations. Therefore, the time-step δt for this sub-routine will be given by $\delta t=\Delta t/\tilde{N}$. The associated discrete operator is denoted $\mathsf{T}_{\tilde{N}}(\Delta t)$. The relaxation part (R) is solved using the following trapezoidal quadrature, see [Del13]:

$$\begin{aligned}\int_0^{\Delta t}\partial_t\boldsymbol{f}^\pm\mathrm{d}t=\boldsymbol{f}^\pm(\Delta t)-\boldsymbol{f}^\pm(0)&=-\frac{1}{\epsilon}\int_0^{\Delta t}\partial_t(\boldsymbol{f}^\pm-\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}))\mathrm{d}t\\&=-\frac{\Delta t}{2\epsilon}((\boldsymbol{f}^\pm-\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}))(\Delta t)+(\boldsymbol{f}^\pm-\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}))(0))+O(\Delta t^2).\end{aligned}$$

Using the fact that the relaxation phase conserves $\boldsymbol{w}$ and thus $\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}(\Delta t))=\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}(0))$, the algorithm can be kept explicit and thus reads

$$\boldsymbol{f}^\pm(\Delta t)=\frac{2\epsilon-\Delta t}{2\epsilon+\Delta t}\boldsymbol{f}^\pm(0)+\frac{2\Delta t}{2\epsilon+\Delta t}\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}(0))\xrightarrow[\epsilon\to 0]{}-\boldsymbol{f}^\pm(0)+2\boldsymbol{f}^{\pm,\mathrm{eq}}(\boldsymbol{w}(0)).$$

This relaxation limit $\epsilon\to 0$ is unsurprisingly similar to (5) for the non-conserved moments when the relaxation parameters s_v and s_w equal two. Notice that the operator associated with this relaxation step in the relaxation

limit $\epsilon \to 0$ does not depend on Δt. It shall be denoted by $\mathbf{R}$ and is an involution, meaning that $\mathbf{R}^2 = \mathbf{I}$. Following [CFH+19], we construct the basic brick of a splitting procedure, called $\mathbf{B}_{\tilde{N}}$, given by

$$\mathbf{B}_{\tilde{N}}(\Delta t) = \mathbf{T}_{\tilde{N}}\Big(\frac{\Delta t}{4}\Big)\mathbf{R}\mathbf{T}_{\tilde{N}}\Big(\frac{\Delta t}{2}\Big)\mathbf{R}\mathbf{T}_{\tilde{N}}\Big(\frac{\Delta t}{4}\Big). \tag{29}$$

This operator is time-symmetric up to order four: the transformation $\Delta t \mapsto -\Delta t$ makes it its own inverse operator. The final operator to be applied is constructed using a fourth-order five-stages symmetric Suzuki splitting [MQ02] given by

$$\mathbf{O}_{\tilde{N}}(\Delta t) = \mathbf{B}_{\tilde{N}}\Big(\frac{1}{4-4^{1/3}}\Delta t\Big)^2\mathbf{B}_{\tilde{N}}\Big(-\frac{4^{1/3}}{4-4^{1/3}}\Delta t\Big)\mathbf{B}_{\tilde{N}}\Big(\frac{1}{4-4^{1/3}}\Delta t\Big)^2.$$

Notice that the central stage features a negative time-step. We handle this point using the fact that the relaxation $\mathbf{R}$ appearing in (29) does not depend on the time-step and that the linear transport equation solved by $\mathbf{T}_{\tilde{N}}$ is reversible in time, hence we just switch $\pm V \mapsto \mp V$.

4.3 Numerical experiments

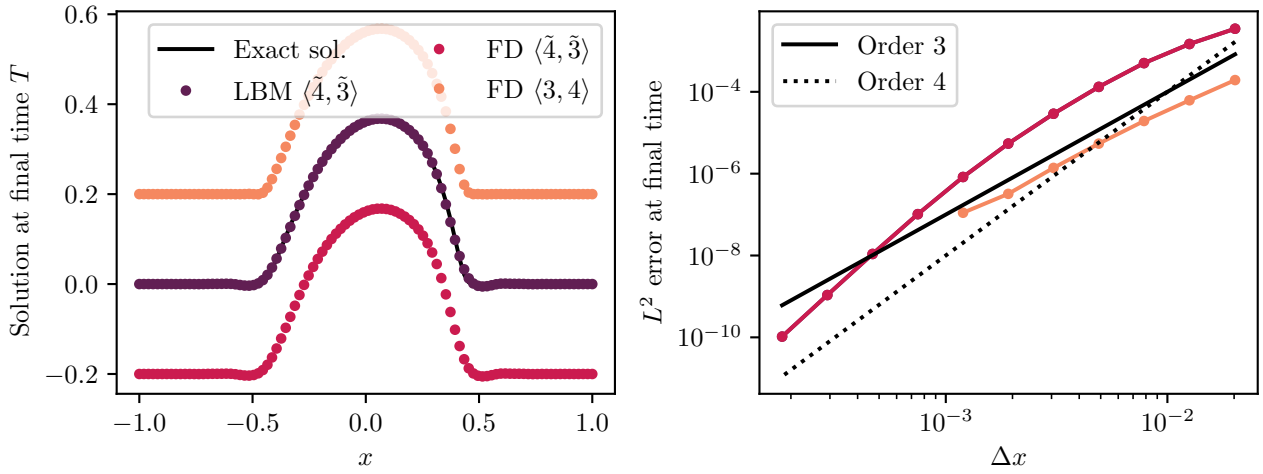

Figure 10: Left: solution of the Burgers equation at final time $T = 0.2$ using $N_x = 100$ grid points for different choices of initialization using the lattice Boltzmann algorithm and the corresponding Finite Difference scheme. The notation $\langle\tilde{4},\tilde{3}\rangle$ indicates (8) with $\delta = 0$. Curves are shifted in order to distinguish between them. Right: convergence of the L^2 error at final time for the Burgers equation.

To test the previously described numerical procedure and—in particular—the fact that it ensures an overall fourth-order method for systems of non-linear PDEs, we consider the inviscid Burgers equation using the same setting of Section 1.2 as far computational domain, initial datum u°, and final time T are concerned. We perform numerical simulations considering $\Delta t = \Delta x$, $\tilde{N} = 6$, and $V = 1$. This choice respects the CFL condition, *cf.* Proposition 3, for the splitting stage with the largest step, given by

$$\frac{|V|4^{1/3}}{2\tilde{N}(4-4^{1/3})} < \frac{1}{2}.$$

The results are presented in Figure 10. We simulate using the original lattice Boltzmann scheme (5)/(6) with initialization (8) and $\delta = 0$, its corresponding Finite Difference scheme (4) with initialization (8) and $\delta = 0$ (which are indeed equivalent), and (4) with initialization $\langle 3,4\rangle$. We observe order four when using (8) with $\delta = 0$ and order three when using $\langle 3,4\rangle$, as expected from the previously conducted analyses. However, we see that $\langle 3,4\rangle$ eventually leads to instabilities when Δx decreases. This is probably due to the fact that (4) is non-dissipative for every mode (all its eigenvalue are on the unit circle for every frequency), both physical and parasitic, and that performing a large number of operations, due to the quite complex splitting procedure, triggers the instability of (4) due to round-off errors. Moreover, we have previously observed that the choice $\langle 3,4\rangle$ excites the unstable parasitic modes more (*i.e.* at order $O(\Delta x^3)$) than $\langle 4,3\rangle$, $\langle 4,4\rangle$ or (8) (*i.e.* at order $O(\Delta x^4)$).

These numerical experiments make the interest of developing high-order lattice Boltzmann schemes for the linear transport equation clear, for this allows to tackle non-linear equations in an efficient manner, thanks to a Jin-Xin relaxation. However, these experiments also show that genuine stability, and possibly some dissipation, are highly desirable property when working with floating point arithmetic.

Acknowledgements

The author would like to deeply thank C. Courtès (University of Strasbourg) for the discussion on this paper, L. François (ONERA) for pointing out that the features investigated in the paper come from the lack of genuine stability, V. Michel-Dansac (INRIA) for pointing out Remark 2, P. Helluy and L. Navoret (University of Strasbourg) for the help with Suzuki splittings, and D. Stantejsky (McMaster University) for his prompt advice on Sobolev spaces and Fourier analysis. The author's post-doc position is funded by IRMIA++ from University of Strasbourg.

References

[ADN00] Denise Aregba-Driollet and Roberto Natalini. Discrete kinetic schemes for multidimensional systems of conservation laws. *SIAM Journal on Numerical Analysis*, 37(6):1973–2004, 2000.

[Bel23a] Thomas Bellotti. Initialisation from lattice Boltzmann to multi-step Finite Difference methods: modified equations and discrete observability. *arXiv preprint arXiv:2302.07558*, 2023.

[Bel23b] Thomas Bellotti. Truncation errors and modified equations for the lattice Boltzmann method via the corresponding Finite Difference schemes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 57(3):1225–1255, 2023.

[BGM22] Thomas Bellotti, Benjamin Graille, and Marc Massot. Finite Difference formulation of any lattice Boltzmann scheme. *Numerische Mathematik*, 152(1):1–40, 2022.

[BT70] Philip Brenner and Vidar Thomée. Stability and convergence rates in L^p for certain difference schemes. *Mathematica Scandinavica*, 27(1):5–23, 1970.

[BTW75] Philip Brenner, Vidar Thomée, and Lars B Wahlbin. *Besov spaces and applications to difference methods for initial value problems*, volume 434. Springer, 1975.

[CF22] Jean-François Coulombel and Grégory Faye. Generalized gaussian bounds for discrete convolution powers. *Revista Matemática Iberoamericana*, 38(5):1553–1604, 2022.

[CFH+19] David Coulette, Emmanuel Franck, Philippe Helluy, Michel Mehrenberger, and Laurent Navoret. High-order implicit palindromic discontinuous Galerkin method for kinetic-relaxation approximation. *Computers & Fluids*, 190:485–502, 2019.

[CL99] Sui Sun Cheng and Yi-Feng Lu. General solutions of a three-level partial difference equation. *Computers & Mathematics with Applications*, 38(7-8):65–79, 1999.

[CL20] Jean-François Coulombel and Frédéric Lagoutière. The Neumann numerical boundary condition for transport equations. *Kinetic and Related Models*, 13(1):1–32, 2020.

[Cou17] Clémentine Courtès. *Analyse numérique de systèmes hyperboliques-dispersifs*. Theses, Université Paris-Saclay, November 2017.

[Cou21] Jean-François Coulombel. The Leray-Garding method for finite difference schemes. II. Smooth crossing modes. *North-W. Eur. J. of Math.*, 7:161–184, 2021.

[Cou22] Jean-François Coulombel. The Green's function of the Lax–Wendroff and Beam–Warming schemes. *Annales mathématiques Blaise Pascal*, 29(2):247–294, 2022.

[Del13] Paul J Dellar. An interpretation and derivation of the lattice Boltzmann method using Strang splitting. *Computers & Mathematics with Applications*, 65(2):129–141, 2013.

[Del23] Paul J Dellar. A magic two-relaxation-time lattice Boltzmann algorithm for magnetohydrodynamics. *Discrete and Continuous Dynamical Systems-S*, pages 0–0, 2023.

[DGR20] François Dubois, Benjamin Graille, and SV Raghurama Rao. A notion of non-negativity preserving relaxation for a mono-dimensional three velocities scheme with relative velocity. *Journal of Computational Science*, 47:101181, 2020.

[DL09] François Dubois and Pierre Lallemand. Towards higher order lattice Boltzmann schemes. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(06):P06006, 2009.

[DT04] Virginie Daru and Christian Tenaud. High order one-step monotonicity-preserving schemes for unsteady compressible flow calculations. *Journal of Computational Physics*, 193(2):563–594, 2004.

[Dub13] François Dubois. Stable lattice Boltzmann schemes with a dual entropy approach for monodimensional nonlinear waves. *Computers & Mathematics with Applications*, 65(2):142–159, 2013.

[Dub22] François Dubois. Nonlinear fourth order Taylor expansion of lattice Boltzmann schemes. *Asymptotic Analysis*, 127(4):297–337, 2022.

[Fév14] Tony Février. *Extension et analyse des schémas de Boltzmann sur réseau: les schémas à vitesse relative.* PhD thesis, Paris 11, 2014.

[FS21] Radek Fučík and Robert Straka. Equivalent finite difference and partial differential equations for the lattice Boltzmann method. *Computers & Mathematics with Applications*, 90:96–103, 2021.

[GKO95] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. *Time dependent problems and difference methods.* John Wiley & Sons, 1995.

[HRS03] Willem Hundsdorfer, Steven J Ruuth, and Raymond J Spiteri. Monotonicity-preserving linear multistep methods. *SIAM Journal on Numerical Analysis*, 41(2):605–623, 2003.

[JX95] Shi Jin and Zhouping Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Communications on Pure and Applied Mathematics*, 48(3):235–276, 1995.

[KP08] Seon-Hong Kim and Chang Woo Park. On the zeros of certain self-reciprocal polynomials. *Journal of Mathematical Analysis and Applications*, 339(1):240–247, 2008.

[LT84] Randall J LeVeque and Lloyd N Trefethen. On the resolvent condition in the Kreiss matrix theorem. *BIT Numerical Mathematics*, 24(4):584–591, 1984.

[Mar66] Morris Marden. *Geometry of polynomials.* Number 3. American Mathematical Soc., 1966.

[Mil71] John JH Miller. On the location of zeros of certain classes of polynomials with applications to numerical analysis. *IMA Journal of Applied Mathematics*, 8(3):397–406, 1971.

[MQ02] Robert I McLachlan and G Reinout W Quispel. Splitting methods. *Acta Numerica*, 11:341–434, 2002.

[MRM94] Gradimir V Milovanovic, Themistocles M Rassias, and DS Mitrinovic. *Topics in polynomials: extremal problems, inequalities, zeros.* World Scientific, 1994.

[MZ88] Guy R. McNamara and Gianluigi Zanetti. Use of the Boltzmann Equation to Simulate Lattice-Gas Automata. *Phys. Rev. Lett.*, 61:2332–2335, Nov 1988.

[SM03] Endre Süli and David F Mayers. *An Introduction to Numerical Analysis.* Cambridge University Press, 2003.

[Str62] Gilbert Strang. Trigonometric polynomials and difference methods of maximum accuracy. *Journal of Mathematics and Physics*, 41(1-4):147–154, 1962.

[Str04] John C Strikwerda. *Finite difference schemes and partial differential equations.* SIAM, 2004.

[Tre92] Lloyd N Trefethen. The definition of numerical analysis. Technical report, Cornell University, 1992.

[Vie19] Ricardo Vieira. Polynomials with symmetric zeros. In *Polynomials-Theory and Application.* IntechOpen, 2019.

[WH74] Robert F Warming and BJ Hyett. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of Computational Physics*, 14(2):159–179, 1974.

A Proof of Proposition 3

We use the iterative procedure by [Mil71] to determine if $\hat{\Phi}(\xi\Delta x, \mathsf{z})$ is a simple *von Neumann* polynomial, namely it has roots inside the closed unit disk and those on the unit circle are simple. Let us set $\varphi_3(\mathsf{z}) = \hat{\Phi}(\xi\Delta x, \mathsf{z})$ by (7). We obtain that $\varphi_3^\dagger(\mathsf{z}) = -\mathsf{z}^3 - \eta(\xi\Delta x)\mathsf{z}^2 + \overline{\eta}(\xi\Delta x)\mathsf{z} + 1$. This results in $\varphi_2(\mathsf{z}) = \mathsf{z}^{-1}(\varphi_3^\dagger(0)\varphi_3(\mathsf{z}) - \varphi_3(0)\varphi_3^\dagger(\mathsf{z})) \equiv 0$, hence we have to show that

$$\psi_2(\mathsf{z}) := \mathrm{d}_\mathsf{z}\varphi_3(\mathsf{z}) = 3\mathsf{z}^2 + 2\eta(\xi\Delta x)\mathsf{z} - \overline{\eta}(\xi\Delta x) \tag{30}$$

is a *Schur* polynomial, namely its roots belong to the open unit disk. We have $\psi_2^\dagger(\mathsf{z}) = -\eta(\xi\Delta x)\mathsf{z}^2 + 2\overline{\eta}(\xi\Delta x)\mathsf{z} + 3$. A first condition to check is $|\psi_2(\mathsf{z})| < |\psi_2^\dagger(0)|$, giving $|\eta(\xi\Delta x)| < 3$. It is simpler to work with the square of the modulus, which provides

$$(4\mathcal{C}^4 - 17\mathcal{C}^2 + 4)\cos^2(\xi\Delta x) + 2(-4\mathcal{C}^4 + 5\mathcal{C}^2 - 1)\cos(\xi\Delta x) + (4\mathcal{C}^4 + 7\mathcal{C}^2 - 20) < 0. \tag{31}$$

Calling $\mu := \cos(\xi\Delta x) \in [-1, 1]$, the previous equation gives a quadratic inequality on μ to be satisfied on $[-1, 1]$. Let us study it according to the sign of the leading term.

- $4\mathcal{C}^4 - 17\mathcal{C}^2 + 4 = (2\mathcal{C} - 1)(2\mathcal{C} + 1)(\mathcal{C} - 2)(\mathcal{C} + 2) > 0$, which is equivalent to $|\mathcal{C}| < 1/2$ or $|\mathcal{C}| > 2$. Under this condition, the maximum of the left-hand side in (31) is reached on the boundary of $[-1, 1]$. Taking $\mu = -1$ provides $16\mathcal{C}^4 - 20\mathcal{C}^2 - 14 < 0$, which solution is $|\mathcal{C}| < \sqrt{7}/2 \approx 1.3229$. Considering $\mu = 1$ trivially gives $-18 < 0$. In this case, the overall condition is $|\mathcal{C}| < 1/2$.
- $4\mathcal{C}^4 - 17\mathcal{C}^2 + 4 = (2\mathcal{C} - 1)(2\mathcal{C} + 1)(\mathcal{C} - 2)(\mathcal{C} + 2) \leq 0$, which is equivalent to $1/2 \leq |\mathcal{C}| \leq 2$. In this case, the maximum of the left-hand side of (31) is reached inside $[-1, 1]$. The value of the inequality (31) on the maximum is given by $4(-4\mathcal{C}^4 + 5\mathcal{C}^2 - 1)^2 - 4(4\mathcal{C}^4 - 17\mathcal{C}^2 + 4)(4\mathcal{C}^4 + 7\mathcal{C}^2 - 20) < 0$. The solutions are $1/2 < |\mathcal{C}| < \sqrt{3/2} \approx 1.2247$. Overall, we the condition is $1/2 < |\mathcal{C}| < \sqrt{3/2}$.

We see that for $|\mathcal{C}| = 1/2$, the inequality (31) is also verified. Therefore, from the previous discussion, the condition we obtain is $|\mathcal{C}| < \sqrt{3/2}$. The next step in the process is to check that $\psi_1(\mathsf{z}) = (9 - |\eta(\xi\Delta x)|^2)\mathsf{z} + 6\eta(\xi\Delta x) + 2\overline{\eta}(\xi\Delta x)^2$ is a *Schur* polynomial as well. This condition reads $|6\eta(\xi\Delta x) + 2\overline{\eta}(\xi\Delta x)^2|^2 - (9 - |\eta(\xi\Delta x)|^2)^2 < 0$. Observe that this does not hold for $\xi = 0$ (or $\mu = 1$), because here we have multiple roots on the unit circle. For $-1 \leq \mu < 1$:

$$\begin{aligned}(16\mathcal{C}^8 - 136\mathcal{C}^6 + 321\mathcal{C}^4 - 136\mathcal{C}^2 + 16)\mu^4 - 4(16\mathcal{C}^8 - 64\mathcal{C}^6 + 195\mathcal{C}^4 - 127\mathcal{C}^2 - 20)\mu^3 \\ + 3(32\mathcal{C}^8 + 16\mathcal{C}^6 + 30\mathcal{C}^4 - 137\mathcal{C}^2 + 32)\mu^2 - 4(16\mathcal{C}^8 + 80\mathcal{C}^6 - 219\mathcal{C}^4 + 107\mathcal{C}^2 + 16)\mu \\ + 16\mathcal{C}^8 + 152\mathcal{C}^6 - 507\mathcal{C}^4 + 467\mathcal{C}^2 - 128 < 0.\end{aligned}$$

As previously discussed, $\mu = 1$ is a zero of the left-hand side, thus we can factorize it out to yield

$$\begin{aligned}(16\mathcal{C}^8 - 136\mathcal{C}^6 + 321\mathcal{C}^4 - 136\mathcal{C}^2 + 16)\mu^3 - 3(16\mathcal{C}^8 - 40\mathcal{C}^6 + 153\mathcal{C}^4 - 124\mathcal{C}^2 - 32)\mu^2 \\ + 3(16\mathcal{C}^8 + 56\mathcal{C}^6 - 123\mathcal{C}^4 - 13\mathcal{C}^2 + 64)\mu - 16\mathcal{C}^8 - 152\mathcal{C}^6 + 507\mathcal{C}^4 - 467\mathcal{C}^2 + 128 > 0.\end{aligned} \tag{32}$$

Assume, without loss of generality, that $0 \leq \mathcal{C} \leq 1$. Differentiating the left-hand side in μ, we obtain a second-order equation for the extremal point of this expression. The one we are interested in is a minimum explicitly given by

$$\mu_{\min} = \frac{-32 - 124\mathcal{C}^2 + 153\mathcal{C}^4 - 40\mathcal{C}^6 + 16\mathcal{C}^8 + 3\sqrt{1872\mathcal{C}^2 - 1640\mathcal{C}^4 - 4459\mathcal{C}^6 + 8484\mathcal{C}^8 - 5392\mathcal{C}^{10} + 1216\mathcal{C}^{12}}}{16 - 136\mathcal{C}^2 + 321\mathcal{C}^4 - 136\mathcal{C}^6 + 16\mathcal{C}^8}.$$

By symbolic computations, this point $\mu_{\min}$ is in $[-1, 1]$ for $\mathcal{C} > 0.206...$. Below this threshold, the extremal point is on the boundary of $[-1, 1]$: for $\mu = -1$, (32) becomes $-16(\mathcal{C}^2 - 1)(2\mathcal{C}^3 + 1)^3 > 0$, which is fulfilled. For $\mu = 1$, which indeed we do not care about, (32) reads $432 - 270\mathcal{C}^2 > 0$, so $|\mathcal{C}| < \sqrt{8/5}$, which is true. Whenever $\mathcal{C} > 0.206...$, we plug $\mu_{\min}$ into (32) and solving the associated inequality in $\mathcal{C}$ using computer algebra provides the condition $|\mathcal{C}| < 1/2$. This achieves the proof.