HAL open science

On the stability of totally upwind schemes for the hyperbolic initial boundary value problem

Boutin Benjamin, Pierre Le Barbenchon, Seguin Nicolas

To cite this version:

Boutin Benjamin, Pierre Le Barbenchon, Seguin Nicolas. On the stability of totally upwind schemes for the hyperbolic initial boundary value problem. IMA Journal of Numerical Analysis, In press, 10.1093/imanum/drad040 . hal-03732720v3

HAL Id: hal-03732720

https://hal.science/hal-03732720v3

Submitted on 16 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

ON THE STABILITY OF TOTALLY UPWIND SCHEMES FOR THE HYPERBOLIC INITIAL BOUNDARY VALUE PROBLEM

BENJAMIN BOUTIN, PIERRE LE BARBENCHON, AND NICOLAS SEGUIN

Abstract. In this paper, we present a numerical strategy to check the strong stability (or GKS-stability) of one-step explicit totally upwind schemes in 1D with numerical boundary conditions. The underlying approximated continuous problem is the one-dimensional advection equation. The strong stability is studied using the Kreiss-Lopatinskii theory. We introduce a new tool, the intrinsic Kreiss-Lopatinskii determinant, which possesses remarkable regularity properties. By applying standard results of complex analysis, we are able to relate the strong stability of numerical schemes to the computation of a winding number, which is robust and cheap. The study is illustrated with the Beam-Warming scheme together with the simplified inverse Lax-Wendroff procedure at the boundary.

AMS classification: 65M12, 65M06
Keywords: boundary conditions, Kreiss-Lopatinskii determinant, GKS-stability, finite-difference methods, inverse Lax-Wendroff

1. Introduction

1.1. **Motivations.** The purpose of this work is to establish an efficient numerical strategy to determine whether a given finite difference method on the half line is stable or not. More precisely, the study is focused on a certain subclass of explicit one-step linear finite difference schemes, specified hereafter. We restrict our attention to the approximation of a rightgoing linear advection equation set on the positive real axis:

$$\begin{cases} \partial_t u + a\partial_x u = 0, & t \geqslant 0, x \geqslant 0, \\ u(t,0) = g(t), & t \geqslant 0, \\ u(0,x) = f(x), & x \geqslant 0, \end{cases} \tag{1}$$

where $u(t,x) \in \mathbb{R}$. The velocity is assumed to be positive $a > 0$ so that at the inflow boundary located at the point $x = 0$, a physical boundary datum g is prescribed.

Let us first recall some general ideas and historical context. As a central idea in numerical analysis, the Lax equivalence theorem (Lax & Richtmyer, 1956) asserts that a consistent scheme is convergent if and only if it is stable. Therefore, all along the paper only consistent numerical schemes are considered, and the discussion concentrates only on their stability issues. While the Cauchy-stability for the space-periodic problem is easily handled with the Fourier symbolic analysis and the so-called Von-Neumann stability analysis, the case with boundaries is significantly trickier. Indeed, the presence of (non-physical) numerical boundary conditions forms another kind of instabilities. The normal mode analysis, directly related to the work by Godunov & Ryabenkii (1963), is the classic way to comprehend those kinds of instabilities. Deepening this analysis with resolvent estimates and Laplace transform leads to the notion of *GKS-stability* (Gustafsson *et al.*, 1972) (sometimes called *strong stability*, see Definition 2 hereafter). This notion is actually the most robust one concerning the stability of initial boundary value numerical methods, since this stability property is stable by

Date: May 9, 2023.
This work has been partially supported by ANR project NABUCO, ANR-17-CE40-0025 and by Centre Henri Lebesgue, program ANR-11-LABX-0020-0.

perturbations and makes use of the same norms for the solution and for the data itself. These features make possible further extensions to more general cases (e.g. nonlinearities), as it is done for the initial boundary value problem in the case of partial differential equations (Benzoni-Gavage & Serre, 2007). In this setting, the Kreiss theorem (see Theorem 3 later) expresses a necessary and sufficient condition for strong stability by the use of the so-called Uniform Kreiss-Lopatinskii Condition. When this condition fails, the corresponding instabilities may be interpreted as numerical wave packets with exponential growth in time and/or bad group velocities; cf. Trefethen (1983, 1984). Some sketches of the strong stability theory will be unfolded later on, but we refer the interested reader to the monographs by Gustafsson (2008) and by Gustafsson *et al.* (2013) for a more complete overview of the GKS-stability theory.

The GKS-stability theory is not used so often in the numerical analysis literature. The reason is that the Uniform Kreiss-Lopatinskii Condition requires the search for the vanishing points of the Kreiss-Lopatinskii determinant, which is a complex-valued function defined on $\{|z| \geqslant 1\}$. Except for some particular numerical schemes and boundary conditions, this determinant is not known explicitly. Indeed, the complexity of the underlying algebra rapidly increases as the size of stencil increase. As an example, Thuné (1986) developed a software system for investigating the GKS-stability. Nevertheless, the method requires the numerical approximate computation of the roots of some parameterized characteristic polynomial equation, and may be expensive in terms of CPU time. In order to tackle the stability properties of the discrete initial boundary value problem, some other strategies are available in the literature. Among them, the most natural approach is based on the spectral properties of the operator corresponding to the time-iteration in the numerical scheme. For a large but finite grid of size J, it is represented by a matrix T_J of size J. It is a banded Toeplitz or a quasi-Toeplitz matrix depending on the boundary conditions under consideration. Beam & Warming (1993) study the asymptotic spectra of such matrices in the limit of large J. Roughly speaking, the stability properties are then related to the uniform boundedness of the powers of the matrix T_J, known as the Kreiss matrix Theorem (Trefethen & Embree, 2005, Chap 18). Nevertheless, the main difficulty is to also guarantee another uniform boundedness property, with respect to the dimension J. The uniform boundedness is not easy to characterize by spectral properties. Some specialized tools exist to that aim: resolvent estimates and ϵ-pseudospectrum. For a wide overview of the Kreiss matrix Theorem and its relationship with resolvent estimates and with the central notion of ϵ-pseudospectrum (Borovykh & Spijker, 2000; Spijker *et al.*, 2003), we refer the reader to the book by Trefethen & Embree (2005). Nonetheless, to our knowledge, the link between GKS-instabilities and the pseudospectrum of the family of quasi-Toeplitz matrices associated to a given scheme is still not completely understood. In the numerical analysis literature, a first attempt thus consists in considering only grids with a large but fixed size J. The postulate is that the asymptotic spectral properties are then already available. This strategy has been used by Dakin *et al.* (2018) for analyzing some specific boundary conditions that we will again consider with our own method in the present paper.

In the present work, the selected strategy is based on the Uniform Kreiss-Lopatinskii Condition and the search of the vanishing points of the corresponding Kreiss-Lopatinskii determinant, that is a function of the complex parameter z defined for $|z| \geqslant 1$. Instead of using the Kreiss-Lopatinskii determinant, we define the *intrinsic Kreiss-Lopatinskii determinant* that shares the same zeros with the Kreiss-Lopatinskii determinant. The main result of the paper (Theorem 13) yields an explicit formula for the intrinsic Kreiss-Lopatinskii determinant, showing that it is holomorphic on $\{|z| > 1\}$. Moreover, the formula does not require the numerical computation of the roots of the associated characteristic equation. Thus, this new theoretical result is particularly useful for numerical applications. Indeed, Corollary 15 presents a strategy to find the number of zeros of the intrinsic Kreiss-Lopatinskii determinant on the domain $\{|z| > 1\}$ using a numerical computation of winding numbers. Hence, this corollary enables the Method 19 to tackle the stability of the scheme. The

whole study in this paper is restricted to totally upwind schemes, so the consistency order is limited to 2 (Iserles & Strang, 1983). As typical examples, we therefore focus on the classic first-order upwind and Beam-Warming schemes, while the generality of the study comes from the fact that we can take any extrapolation boundary condition using some points of the domain (the precise form of the considered boundary conditions will be set later at equation (5)). In the paper, the numerical examples deal with the inverse Lax-Wendroff boundary condition, and the simplified variants of it, as introduced by Tan & Shu (2010); Vilar & Shu (2015) and used by Li *et al.* (2016, 2017, 2022) to solve advection and diffusion equations. These authors consider a stability analysis based either on the Godunov-Ryabenkii algebraic condition, or by the so-called eigenvalue spectrum visualization method. This last method again requires the use of a finite grid and the computation of the eigenvalues for a large banded matrix.

The outline of the paper is as follows. In the sequel of this introduction section, we describe the main assumptions and the notion of stability into play. In Section 2, we set up the main tool for our study that is the Kreiss-Lopatinskii determinant and the intrinsic Kreiss-Lopatinskii determinant, then we state our main results. In Section 3, we prove these results relying on linear algebra tools and complex analysis results. Section 4 gathers several examples and numerical experiments for illustrating the efficiency of the proposed strategy.

1.2. **Notations and assumptions.** Throughout this paper we denote $\mathbb{S} = \{z \in \mathbb{C},\ |z| = 1\}$ the unit circle, $\mathbb{D} = \{z \in \mathbb{C},\ |z| < 1\}$ the open unit disk, $\mathcal{U} = \{z \in \mathbb{C},\ |z| > 1\}$ the exterior domain and $\overline{\mathcal{U}} = \{z \in \mathbb{C},\ |z| \geqslant 1\}$ its closure. For $n < m$, the notation $[\![n : m]\!]$ is for the set $\{k \in \mathbb{N},\ n \leqslant k \leqslant m\}$.

At the discrete level, we consider explicit one-step finite difference methods of the form

$$U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{j+k}^n, \tag{2}$$

with integers $r, p \geqslant 0$. Here, the unknown of the scheme U_j^n is expected to approximate the quantity $u(n\Delta t, j\Delta x)$. The time step $\Delta t > 0$ and the space step $\Delta x > 0$ are usually chosen with respect to some CFL condition $\lambda = a\Delta t/\Delta x \leqslant \lambda_{\mathsf{CFL}}$ discussed later on.

The *symbol* associated to the scheme (2) is defined, for $\xi \in \mathbb{R}$, by

$$\gamma(\xi) = \sum_{k=-r}^{p} a_k e^{ik\xi}. \tag{3}$$

The common set of assumptions used hereafter is the following one.

Assumptions. The scheme (2) is

- (H0) *non-degenerate*, in the sense that $a_{-r} \neq 0$,
- (H1) *totally upwind*, in the sense that $p = 0$,
- (H2) *Cauchy-stable*, meaning that the symbol γ satisfies $|\gamma(\xi)| \leqslant 1$ for all $\xi \in \mathbb{R}$.
- (H3) *consistent* and at least first order, meaning that

$$\gamma(0) = \sum_{k=-r}^{p} a_k = 1 \text{ and } -i\gamma'(0) = \sum_{k=-r}^{p} k a_k = -\lambda.$$

When dealing with the discrete schemes set over the full line $j \in \mathbb{Z}$, the algebraic characterization of the Cauchy-stability classically follows from the Fourier analysis and makes use of the symbol γ. This method is known as the Von Neumann analysis (Courant *et al.*, 1928) and (Crank & Nicolson, 1947). In the scalar case, it reduces to a geometric property concerning the following closed complex curve.

Definition 1. The *symbol curve* Γ is the closed complex parameterized curve

$$\Gamma = \{\theta \in [0, 2\pi] \mapsto \gamma(\theta)\}.$$

This definition enables a geometric interpretation of the Cauchy-stability assumption (H2) reformulated equivalently as the inclusion $\Gamma \subset \overline{\mathbb{D}}$ (see later Figure 2 for the Beam-Warming scheme). In the same vein, the consistency assumption (H3) admits a geometric form through a first order tangency property of Γ to the vertical axis at the parameter point $\theta = 0$. The stability condition (H2) can be easily illustrated graphically in the complex plane. In some sense, our goal is to extend this kind of graphical study when including the numerical boundary conditions.

For solving the Initial Boundary Value Problem (IBVP) (1) with the discrete scheme (2), r additional ghost points are needed to take into account the left boundary condition and to fully define the discrete approximation. In the theoretical results of the paper, we assume that the values at these ghost points are obtained from a linear combination of the first values of the solution close to the boundary and at the same time step, as follows.

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{0} a_k U_{k+j}^n, & j \in \mathbb{N},\ n \in \mathbb{N}, \quad (4) \\ U_j^n = \displaystyle\sum_{k=0}^{m-1} b_{j,k} U_k^n + g_j^n, & j \in [\![-r : -1]\!],\ n \in \mathbb{N}, \quad (5) \\ U_j^0 = f_j, & j \in \mathbb{N}, \quad (6) \end{cases}$$

where m, r are integers, f_j are approximations of the initial condition $f(x_j)$ and g_j^n are numerical data related to the boundary datum g. With the vector notation $U = (U_{-r}^n \cdots U_{m-1}^n)^\top$ and $G = (g_{-r}^n \cdots g_{-1}^n)^\top$, the boundary equation (5) reads also equivalently as $BU = G$ with the following matrix

$$B \overset{def}{=} \begin{pmatrix} 1 & & 0 & -b_{-r,0} & \cdots & -b_{-r,m-1} \\ & \ddots & & \vdots & & \vdots \\ 0 & & 1 & -b_{-1,0} & \cdots & -b_{-1,m-1} \end{pmatrix} \in \mathcal{M}_{r,r+m}(\mathbb{C}). \quad (7)$$

This class of boundary conditions encompasses the Dirichlet and Neumann extrapolation procedures (Goldberg, 1977), but also the more general simplified inverse Lax-Wendroff procedure; cf. (Vilar & Shu, 2015), (Li *et al.*, 2017), (Dakin *et al.*, 2018) and Section 4.3. We will focus on these boundary conditions in our numerical examples. More specific treatments at the boundary exist, as for example absorbing boundary conditions (Engquist & Majda, 1977) and (Ehrhardt, 2010), or transparent boundary conditions (Arnold *et al.*, 2003) and (Coulombel, 2019), however, in general, they do not enter the present framework.

1.3. **Classic results about strong stability.** The GKS-stability theory (see the seminal paper by Gustafsson *et al.* (1972)) handles the discrete IBVP (4)-(5)-(6) with a zero initial data. We refer the reader to the work by Wu (1995) and Coulombel (2013) for more recent development on semigroup estimates. They extend a stability result for the discrete IBVP (4)-(5)-(6), available for zero initial data, to the case of non-zero initial data. The corresponding notions of stability for the boundary problem makes use of the following discrete norms:

$$\|U_j\|_{\Delta t}^2 = \sum_{n=0}^{+\infty} \Delta t |U_j^n|^2 \quad \text{and} \quad \|U\|_{\Delta x, \Delta t}^2 = \sum_{n=0}^{+\infty} \sum_{j=-r}^{+\infty} \Delta t \Delta x |U_j^n|^2.$$

The latter norm is associated with the space $\ell^2(\{-r, \ldots, -1\} \cup \mathbb{N})$, denoted shortly ℓ^2. We are now in position to define the so-called strong stability, for zero initial data.

Definition 2 (Strong stability)**.** The scheme (4)-(5)-(6) is strongly stable if, taking $(f_j) = 0$, there exist $C > 0$ and α_0, such that for all $\alpha > \alpha_0$, for all boundary data (g_j^n), for all $\Delta x > 0$, for all $n \in \mathbb{N}$, the approximate solution (U_j^n) satisfies

$$\sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2 + \left(\frac{\alpha - \alpha_0}{\alpha \Delta t + 1}\right) \|e^{-\alpha n \Delta t} U\|_{\Delta x, \Delta t}^2 \leqslant C \sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} g_j\|_{\Delta t}^2. \tag{8}$$

We warn the reader that $\|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2$ is an abuse of notation to describe $\sum_{n=0}^{+\infty} \Delta t e^{-2\alpha n \Delta t} |U_j^n|^2$, and similarly for $\|e^{-\alpha n \Delta t} U\|_{\Delta x, \Delta t}^2$.

The following Kreiss theorem provides two necessary and sufficient conditions for the strong stability. We provide hereafter a condensed formulation of this theorem, obtained from (Gustafsson *et al.*, 1972, Thm. 5.1) combined with (Gustafsson *et al.*, 2013, Lem. 13.1.4) or with (Gustafsson, 2008, Def. 2.23). It makes use of the notions of *eigenvalue* and *generalized eigenvalue* that will be defined later in Definition 16 and Definition 17.

Theorem 3 (Kreiss)**.** *The following statements are equivalent:*

(i) The scheme (4)-(5)-(6) is strongly stable in the sense of Definition 2.
(ii) The scheme (4)-(5)-(6) has neither eigenvalue nor generalized eigenvalue.
(iii) The Uniform Kreiss-Lopatinskii Condition is satisfied.

The Uniform Kreiss-Lopatinskii Condition corresponds to the absence of zeros for the so-called Kreiss-Lopatinskii determinant (see later Definition 11 and (Gustafsson *et al.*, 2013)). These zeros are identified to eigenvalues or to generalized eigenvalues in the sense of Definitions 16 and 17 and correspond to modal instabilities. Our numerical analysis of the strong stability of the discrete IBVP will be based on a geometrical study of the Kreiss-Lopatinskii determinant.

2. Kreiss-Lopatinskii determinants

In this section, we introduce the Kreiss-Lopatinskii determinant, define the intrinsic Kreiss-Lopatinskii determinant and construct an algebraic reformulation of it (see Theorem 13 later). This explicit formula shows that it is holomorphic on $\{|z| > 1\}$ and is independent of the roots of the associated characteristic equation. At last, by Corollary 15, a numerical procedure based on the Theorem 3 (Kreiss) gives a strategy to tackle the stability of the scheme.

2.1. Stable subspace and matrix representation.

First, we assume (H1) and study the solutions to the interior equation:

$$U_j^{n+1} = \sum_{k=-r}^{0} a_k U_{k+j}^n, \ j \in \mathbb{N}, \ n \in \mathbb{N}. \tag{9}$$

To study this equation, the $\mathcal{Z}$-transform (see (Gasquet & Witomski, 1999, Lesson 40)) is applied. This transformation is defined for $(x_n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$ such that $x_0 = 0$ and $z \in \mathcal{U}$ by $\widetilde{x}(z) = \sum_{n\geqslant 0} z^{-n} x_n$. The previous equation then becomes

$$z\widetilde{U}_j(z) = \sum_{k=-r}^{0} a_k \widetilde{U}_{j+k}(z), \ j \in \mathbb{N}, \ z \in \mathcal{U}. \tag{10}$$

To solve the linear recurrence equation (10), let us introduce the following characteristic equation where z plays the role of a parameter and κ is the indeterminate:

$$z\kappa^r = \sum_{k=-r}^{0} a_k \kappa^{r+k}. \tag{11}$$

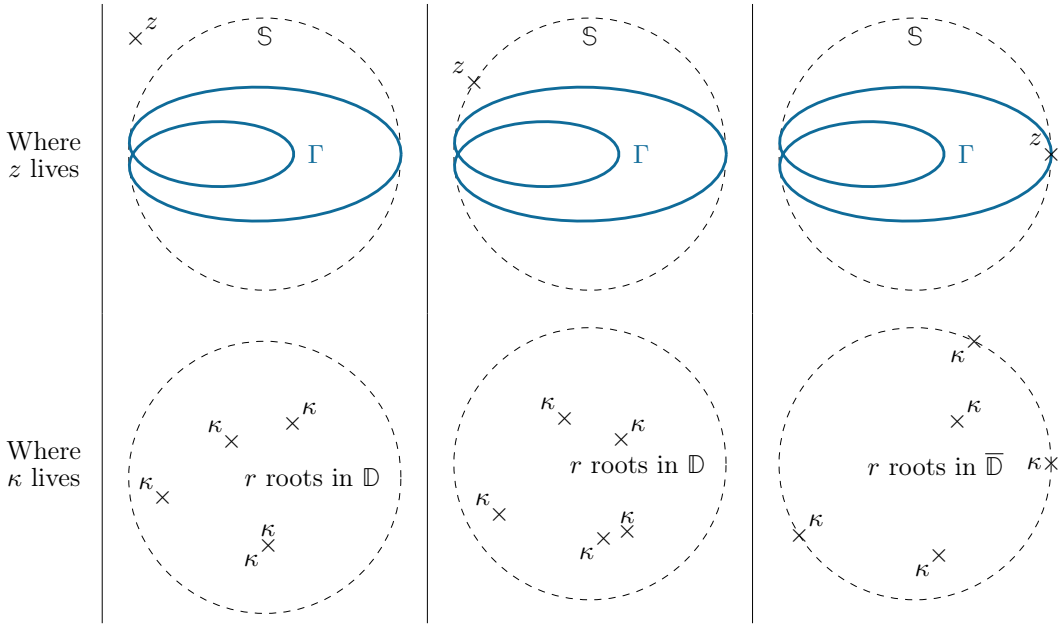

FIGURE 1. Illustration of Lemma 4: case $|z| > 1$ (first column), case $|z| = 1$ and $z \notin \Gamma$ (second column) and case $z \in \Gamma$ where Lemma 4 does not hold (third column).

This equation is nothing but the discrete dispersion relation of the finite difference scheme (9), with frequency parameter κ in space and z in time. It is formally obtained by looking for solutions to the interior equation (9) having the form $U_j^n = z^n \kappa^j$.

In the spirit of a classic result by Hersh (1963), the following lemma indicates a property of separation for the roots with respect to the unit circle.

Lemma 4 (Hersh)**.** *Assume (H0) and (H1). For z in the unbounded connected component of $\mathbb{C} \setminus \Gamma$, all the roots of the characteristic equation* (11) *are in $\mathbb{D}$.*

The proof of this result is omitted but may be found in (Hersh, 1963).

Remark 5. Under the Cauchy-stability assumption (H2), the inclusion $\Gamma \subset \overline{\mathbb{D}}$ is known. From there, it follows that the unbounded connected component of $\mathbb{C} \setminus \Gamma$ contains the whole set $\mathcal{U}$ so that a weaker form of the lemma is available for considering $z \in \mathcal{U}$ only. If in addition, the considered scheme is also *dissipative*, that is if its symbol γ satisfies

$$|\gamma(\xi)| \leqslant 1 - \delta|\xi|^{2s}, \quad \xi \in [-\pi, \pi],$$

for some $\delta > 0$ and an integer $s \in \mathbb{N}^*$ independent of ξ, then the same separation result is available for $z \in \overline{\mathcal{U}} \setminus \{1\}$. The reason for this property is that one has $\mathbb{S} \cap \Gamma = \{1\}$.

Lemma 4 (Hersh) is illustrated in Figure 1. The first two columns correspond to the Hersh lemma and the third one describes the possible configuration for $z \in \Gamma \cap \mathbb{S}$, typically not meeting the assumptions. This case will be the object of a subsequent discussion.

Remark 6. Setting the assumption (H1) aside, meaning with a non-zero number p of right points, the more general form of the Hersh lemma states that for any convenient value of z, there are exactly r roots (with multiplicity) inside the open unit disk, exactly p roots (with multiplicity) outside the unit disk and no root on the unit circle. The result can be proved by using Rouché's theorem.

For $|z| > 1$, we denote $\mathcal{E}^s(z)$ the linear subspace of solutions to (10) living in ℓ^2 (the ℓ^2 space with indices between $-r$ and $+\infty$). By Lemma 4 (Hersh), the space $\mathcal{E}^s(z)$ is generated by the following

r vectors:

$$\begin{pmatrix} \kappa_i^{-r} \\ \vdots \\ \kappa_i^{-1} \\ 1 \\ \kappa_i \\ \kappa_i^2 \\ \kappa_i^3 \\ \vdots \end{pmatrix}, \begin{pmatrix} -r\kappa_i^{-r} \\ \vdots \\ -\kappa_i^{-1} \\ 0 \\ \kappa_i \\ 2\kappa_i^2 \\ 3\kappa_i^3 \\ \vdots \end{pmatrix}, \dots, \begin{pmatrix} (-r)^{\beta_i-1}\kappa_i^{-r} \\ \vdots \\ (-1)^{\beta_i-1}\kappa_i^{-1} \\ 0 \\ \kappa_i \\ 2^{\beta_i-1}\kappa_i^2 \\ 3^{\beta_i-1}\kappa_i^3 \\ \vdots \end{pmatrix}, \quad i = 1, \dots, M \tag{12}$$

where $\kappa_1, \dots, \kappa_M$ of multiplicity $\beta_1, \dots, \beta_M$ are the solutions to (11), with $\beta_1 + \cdots + \beta_M = r$. (We omit the z-dependence of $\kappa(z)$ for the sake of readability.)

Notation. We denote $K_{i,j}(z) \in \mathcal{M}_{j-i+1,r}(\mathbb{C})$ the matrix where we put in columns the extraction of all the lines between i and j (included) of the previous vectors, where $-r \leqslant i \leqslant j$.

Remark 7. For $r = 2$, if the solutions to (11) are $\kappa_1(z) \neq \kappa_2(z)$, then there are exactly two roots with multiplicity 1. The solutions to (10) can be written $\widetilde{U}_j(z) = \alpha_1\kappa_1(z)^j + \alpha_2\kappa_2(z)^j$, and we have

$$K_{-2,2}(z) = \begin{pmatrix} \kappa_1(z)^{-2} & \kappa_2(z)^{-2} \\ \kappa_1(z)^{-1} & \kappa_2(z)^{-1} \\ 1 & 1 \\ \kappa_1(z) & \kappa_2(z) \\ \kappa_1(z)^2 & \kappa_2(z)^2 \end{pmatrix}.$$

Remark 8. Still for $r = 2$, if the solution to (11) now is $\kappa(z)$ with multiplicity 2, then the solutions to (10) can be written $\widetilde{U}_j(z) = (\alpha_1 + \alpha_2 j)\kappa(z)^j$, and we have

$$K_{0,3}(z) = \begin{pmatrix} 1 & 0 \\ \kappa(z) & \kappa(z) \\ \kappa(z)^2 & 2\kappa(z)^2 \\ \kappa(z)^3 & 3\kappa(z)^3 \end{pmatrix}.$$

We raise awareness of the dependence on z and of the continuity issues because the map $z \mapsto K_{i,j}(z)$ is not continuous whereas the set of roots of (11) is a continuous mapping with respect to z. Indeed, the root curves $(\kappa_j(z))_j$ can intersect, when a multiple root occurs. For example, for $r = 2$, if there is $(z_n)_{n\in\mathbb{N}} \subset \mathcal{U}$ with $\kappa_1(z_n) \neq \kappa_2(z_n)$ which converge to $z_\infty \in \mathcal{U}$ such that $\kappa_1(z_\infty) = \kappa_2(z_\infty)$ a double root, then we have, for $j = 1$ and $j = 2$,

$$\kappa_j(z_n) \xrightarrow[n\to\infty]{} \kappa_j(z_\infty)$$

but

$$K_{0,3}(z_n) = \begin{pmatrix} 1 & 1 \\ \kappa_1(z_n) & \kappa_2(z_n) \\ \kappa_1^2(z_n) & \kappa_2^2(z_n) \\ \kappa_1^3(z_n) & \kappa_2^3(z_n) \end{pmatrix} \not\xrightarrow[n\to\infty]{} K_{0,3}(z_\infty) = \begin{pmatrix} 1 & 0 \\ \kappa_1(z_\infty) & \kappa_1(z_\infty) \\ \kappa_1^2(z_\infty) & 2\kappa_1^2(z_\infty) \\ \kappa_1^3(z_\infty) & 3\kappa_1^3(z_\infty) \end{pmatrix}.$$

Consequently, the considered basis (12) of $\mathcal{E}^s(z)$ does not generally define a continuous mapping with respect to z. Nevertheless, $\mathcal{E}^s(z)$ is a continuous and even holomorphic vector bundle over $\mathcal{U}$ as it is discussed in (Coulombel, 2013, Thm. 4.3). This author proves in addition that this vector bundle $\mathcal{E}^s(z)$ can even be continuously extended over $\overline{\mathcal{U}}$, thus considering $z \in \mathbb{S}$ as well; cf. (Métivier & Zumbrun, 2004) for a similar property for the hyperbolic-parabolic PDE case. The main point therein is that for some $z_0 \in \mathbb{S}$, there may exist one (or several) root $\kappa_0(z_0)$ of (11) on $\mathbb{S}$, because Hersh lemma does not hold any more. This situation is depicted on the third column of Figure 1 and the different cases that may occur will be explained in Section 2.4.

In the case of a totally upwind scheme, it is easy to extend the space $\mathcal{E}^s(z)$ because it is the linear space generated by the r roots of (11) with polynomial terms for multiplicity. Indeed, $\kappa(z)$ can be defined for all $z \in \overline{\mathcal{U}}$ by continuity of $\kappa(z)$ for $z \in \mathcal{U}$. The space $\mathcal{E}^s(z)$ still is of dimension r, and we extend the notation $K_{i,j}(z)$ for z on $\mathbb{S}$. But the difficulty is to prove the continuity of $\mathcal{E}^s(z)$ after the extension, it follows from the existence of a K-symmetrizer and is obtained e.g. in (Coulombel, 2013, Thm. 4.3). As previously observed, $K_{i,j}(z)$ is generally not continuous with respect to z.

We can summarize the discussion in the following theorem.

Theorem 9 (Coulombel (2013))**.** *Under assumptions (H0), (H1) and (H2), the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over $\mathcal{U}$ and can be extended uniquely as a continuous vector bundle over $\overline{\mathcal{U}}$.*

Moreover, in the more general case where there are p right points, the extension of $\mathcal{E}^s(z)$ is not so easy to define because the r roots that come from the inside of the unit open disk must be selected. Indeed, if there is some $\kappa_0(z_0)$ on the unit circle, one has to know if the root comes from the outside or the inside of the unit disk when z tends to z_0 from the outside. Worse, it is possible to have a multiple root on the unit circle with some come from the inside of the unit disk and others from the outside.

2.2. **Intrinsic Kreiss-Lopatinskii determinant.** Now, let us consider the $\mathcal{Z}$-transformed version of the boundary condition (5), that is, for j between $-r$ and -1,

$$\widetilde{U}_j(z) = \sum_{k=0}^{m-1} b_{j,k}\widetilde{U}_k(z) + \widetilde{g}_j(z). \tag{13}$$

Injecting the fundamental solutions to $\mathcal{E}^s(z)$ into (13), we obtain a system of r equations with r scalar unknowns. They are the coefficients of the solution to (13) written in the basis (12) of $\mathcal{E}^s(z)$.

Remark 10. For $r = 2$ and a given value of z (we skip for convenience the dependence in z hereafter), if $\kappa_1 \neq \kappa_2$ so that the solution to (13) has the form $\alpha_1\kappa_1^j + \alpha_2\kappa_2^j$, then that solution is constrained by the following two scalar equations:

$$\begin{cases} \alpha_1\kappa_1^{-2} + \alpha_2\kappa_2^{-2} = \sum_{k=0}^{m-1} b_{-2,k}(\alpha_1\kappa_1^k + \alpha_2\kappa_2^k) + \widetilde{g}_{-2}, \\ \alpha_1\kappa_1^{-1} + \alpha_2\kappa_2^{-1} = \sum_{k=0}^{m-1} b_{-1,k}(\alpha_1\kappa_1^k + \alpha_2\kappa_2^k) + \widetilde{g}_{-1}. \end{cases}$$

The matricial form of that system reads

$$\underbrace{\begin{pmatrix} 1 & 0 & -b_{-2,0} & \cdots & -b_{-2,m-1} \\ 0 & 1 & -b_{-1,0} & \cdots & -b_{-1,m-1} \end{pmatrix}}_{B} \begin{pmatrix} \kappa_1^{-2} & \kappa_2^{-2} \\ \kappa_1^{-1} & \kappa_2^{-1} \\ 1 & 1 \\ \kappa_1 & \kappa_2 \\ \kappa_1^2 & \kappa_2^2 \\ \vdots & \vdots \\ \kappa_1^{m-1} & \kappa_2^{m-1} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \widetilde{g}_{-2} \\ \widetilde{g}_{-1} \end{pmatrix}.$$

The injectivity, whence invertibility, of the boundary condition is thus directly related to the property $\det BK_{-2,m-1}(z) \neq 0$, where $BK_{-2,m-1}(z) \in \mathcal{M}_{2,2}(\mathbb{C})$.

Definition 11 (Kreiss-Lopatinskii determinant)**.** The *Kreiss-Lopatinskii determinant* is the complex-valued function defined for $|z| \geqslant 1$ by

$$\Delta_{\mathrm{KL}}(z) \overset{def}{=} \det BK_{-r,m-1}(z).$$

Before giving the definition let us motivate the *intrinsic Kreiss-Lopatinskii determinant* Δ by the following informal discussion. The above Kreiss-Lopatinskii determinant is actually not well-defined until we order in some way the roots $(\kappa_j(z))_{j=1,\dots,r}$ of (11). There are two points to notice. The

first one is related to crossing roots, already discussed after Remark 8. The second one is that, outside crossing cases, being given any choice for the ordering of the roots (and thus of the vectors of the basis (12) for the vector bundle), there is in general no chance to obtain a holomorphicity property for the components of the matrix $K_{-r,m-1}(z)$ over $\mathcal{U}$. For example, even the roots of $X^2 - z$ are not holomorphic w.r.t z because of the logarithm determination. On the other side, any symmetric functions of the roots $(\kappa_j(z))_{j=1,\dots,r}$ however are holomorphic because they can be obtained directly in terms of the coefficients of the polynomial (11). Therefore, except for crossing roots, the same holds for the quantity $\Delta_{\mathrm{KL}}(z)$ since the matrix B is constant and the determinant itself is a symmetric function.

It is now known that the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over $\mathcal{U}$, continuous over $\overline{\mathcal{U}}$, and thus we should expect the same for Δ_{KL}. A very natural way to reach that property and go beyond the last difficulties consists in dividing Δ_{KL} by the quantity $\det K_{0,r-1}(z)$. In this manner, the same permutation or combination of the vectors of the basis (12) is involved for both computations.

Definition 12 (Intrinsic Kreiss-Lopatinskii determinant)**.** The *intrinsic Kreiss-Lopatinskii determinant* is the complex-valued function defined for $|z| \geqslant 1$ by:

$$\Delta(z) = \frac{\Delta_{\mathrm{KL}}(z)}{\det K_{0,r-1}(z)}. \tag{14}$$

To conclude with these definitions, let us state a little more about the *Uniform Kreiss-Lopatinskii Condition*. With the above notations and additionally to the invertibility of $BK_{-r,m-1}(z)$, it corresponds to the existence of a constant $C > 0$ such that for any $z \in \overline{\mathcal{U}}$, any $U \in \mathcal{E}^s(z)$ solution to (13) satisfies the uniform estimate

$$\|\widetilde{U}\| \leqslant C\|\widetilde{g}\|.$$

From the Parseval identity for the $\mathcal{Z}$-transform, this inequality yields directly the first necessary half-part of the strong stability estimate (8). We refer the reader to (Gustafsson *et al.*, 2013) for a more detailed presentation.

2.3. **Main results.** Theorem 13 is our main theoretical result. It yields an explicit formulation of the intrinsic Kreiss-Lopatinskii determinant and therefore describes its properties. Namely, as a function of z, this determinant Δ is holomorphic on $\mathcal{U}$, is continuous on $\overline{\mathcal{U}}$ and depends on $\mathcal{E}^s(z)$ but not on the choice of a basis (what justifies the *intrinsic* denomination of that quantity).

Theorem 13 (Explicit formula of the intrinsic Kreiss-Lopatinskii determinant)**.** *Assume (H0), (H1), (H2) and (H3). The intrinsic Kreiss-Lopatinskii determinant is given, for $z \in \overline{\mathcal{U}}$, by*

$$\Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a_{-r}}{a_0 - z}\right)^{m-r} \tag{15}$$

where $\det C(z)$ is a constructible polynomial of z depending only on the coefficients $(a_j)_{j=-r}^{0}$ and on the components of B.

By "constructible polynomial", we mean here that we establish a computable algorithm to get a matrix $C(z)$ and then the polynomial $\det C(z)$. This algorithm, based on a Gaussian elimination, is fully described in the proof of Lemma 21. In the proof of Theorem 13, we will explicitly see the holomorphic property of Δ. Another property, important for the forthcoming applications, lies in the next Corollary 15 and involves the following important geometrical object:

Definition 14. The *Kreiss-Lopatinskii curve* $\Delta(\mathbb{S})$ is the closed complex parameterized curve

$$\Delta(\mathbb{S}) = \{\theta \in [0, 2\pi] \mapsto \Delta(e^{i\theta})\}.$$

Corollary 15 (Number of zeros of the intrinsic Kreiss-Lopatinskii determinant)**.** *Assume (H0), (H1), (H2) and (H3). If $0 \notin \Delta(\mathbb{S})$ then the equation $\Delta(z) = 0$ has exactly $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ zeros in $\mathcal{U}$.*

Here above and in all the paper, $\mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ denotes the winding number of the origin with respect to the closed oriented curve $\Delta(\mathbb{S})$; cf. (Lang, 1999) for a definition of the winding number. This previous corollary is the fundamental piece to the following numerical procedure to tackle stability. Indeed, by the definition (14), the function Δ shares the same zeros with the Kreiss-Lopatinskii determinant Δ_{KL}, which in turn characterizes the stability with Theorem 3 (Kreiss).

2.4. **Numerical procedure.** As already seen in the Theorem 3 (Kreiss), the strong stability can be characterized by the notion of eigenvalue and generalized eigenvalue for the boundary problem. The definition of generalized eigenvalue is not universal, the following one comes from (Gustafsson *et al.*, 2013, Def. 12.2.2), but one can also find a slightly different one in (Gustafsson, 2008, Def. 2.2). The difference will be discussed afterwards.

Definition 16 (Eigenvalue)**.** Let z be a complex number. If $|z| \geqslant 1$, $\Delta(z) = 0$ and the solution $(\widetilde{U}_j(z))_j$ to (10) and (13) with $(\widetilde{g}_j(z)) = 0$ is in ℓ^2 then z is called an *eigenvalue*.

Definition 17 (Generalized eigenvalue)**.** Let z_0 be a complex number with $|z_0| = 1$. If $\Delta(z_0) = 0$ and the solution $(\widetilde{U}_j(z_0))_j$ to (10) and (13) with $(\widetilde{g}_j(z_0)) = 0$is not in ℓ^2 then z_0 is called a *generalized eigenvalue*.

If $|z| > 1$ and $\Delta(z) = 0$, it is not possible to have $(\widetilde{U}_j(z)) \notin \ell^2$, because by Lemma 4 (Hersh), the r roots of (11) that are used to construct $(\widetilde{U}_j(z))$ are in the open unit disk. That's why the definition of generalized eigenvalue concerns only complex values on the unit circle.

Therefore, we can split all cases in four types:

(i) z such that $\Delta(z) = 0$ and $|z| > 1$.
(ii) z such that $\Delta(z) = 0$, $|z| = 1$ and $z \notin \Gamma$.
(iii) z such that $\Delta(z) = 0$, $|z| = 1$, $z \in \Gamma$ and $(\widetilde{U}_j(z)) \in \ell^2$.
(iv) z such that $\Delta(z) = 0$, $|z| = 1$, $z \in \Gamma$ and $(\widetilde{U}_j(z)) \notin \ell^2$.

The types (i), (ii) and (iii) describe all the eigenvalues. Indeed, for type (i) and (ii), by Lemma 4 (Hersh), we have $(\widetilde{U}_j(z)) \in \ell^2$, because every root κ of (11) is in the open unit disk. Type (i) corresponds to the first column of Figure 1 and type (ii) corresponds to the second column.

Moreover, the non-existence of eigenvalue of type (i) is a necessary condition to have stability. It is called the *Godunov-Ryabenkii* condition, introduced in (Godunov & Ryabenkii, 1963) and described in (Trefethen, 1984).

If z is of type (iii) or (iv), there exists a $\kappa_0(z)$ root of (11) on the unit circle because $z \in \Gamma$. This is the situation depicted on the third column of Figure 1. The distinction between (iii) and (iv) is more subtle and comes from the expression of $(\widetilde{U}_j(z))$ in the basis of $\mathcal{E}^s(z)$, where the coefficient(s) in front of the vector(s) related to $\kappa_0(z)$ can be zero or not.

By the way, let us mention that our definition of generalized eigenvalue, from (Gustafsson *et al.*, 2013, Def.12.2.2), corresponds, as we already said, to type (iv) whereas the definition from (Gustafsson, 2008, Def. 2.2) combines type (iii) and (iv).

Now, Corollary 15 can be reformulated as follows.

Corollary 18. *Assume (H0), (H1), (H2) and (H3). If* $0 \notin \Delta(\mathbb{S})$ *then the scheme has* $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ *eigenvalues in* $\mathcal{U}$ *(type (i)).*

This corollary enables us to establish an efficient and practical method to study the stability of a given IBVP through Theorem 3 (Kreiss). In particular, the low computational cost of the following procedure is very appealing for the study of parameterized IBVP's, see Section 4.

Method 19 (Uniform Kreiss-Lopatinskii Condition check)**.** *There are two different cases:*

- *If $0 \notin \Delta(\mathbb{S})$, there is neither generalized eigenvalue (type (iv)) nor eigenvalue on the unit circle (type (ii) and (iii)) and there are $r - Ind_{\Delta(\mathbb{S})}(0)$ zeros of Δ in $\mathcal{U}$ by Theorem 15 (type (i)). It follows that if the scheme has no eigenvalue in $\mathcal{U}$ then the scheme is stable. Otherwise, there exists an eigenvalue and the scheme is unstable.*
- *If $0 \in \Delta(\mathbb{S})$, then there exists $z_0 \in \mathbb{S}$ such that $\Delta(z_0) = 0$.*
  - *If $z_0 \in \Gamma$, then z_0 is a generalized eigenvalue (of type (iv)) or an eigenvalue of type (iii).*
  - *If $z_0 \notin \Gamma$, then there are two possibilities:*
    - ➢ *First, z_0 is in the unbounded connected component of $\mathbb{C} \setminus \Gamma$. By Lemma 4 (Hersh), there is no κ on the unit circle, so z_0 is an eigenvalue on the unit circle (type (ii)).*
    - ➢ *Second, z_0 is in a bounded connected component of $\mathbb{C} \setminus \Gamma$. Contradiction with the Cauchy-stability because $\Gamma \subset \overline{\mathbb{D}}$ and $z_0 \in \mathbb{S}$.*

This method does not distinguish between types (iii) and (iv). In fact, we only study the presence or absence of instabilities, we do not attempt to determine which type of instability mode is met; cf. Trefethen (1984).

In summary, by Theorem 3 (Kreiss), if $0 \in \Delta(\mathbb{S})$ then the scheme is not stable, and if $0 \notin \Delta(\mathbb{S})$, Theorem 15 can be used to conclude that the scheme is stable or not, depending on the value of $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$. Some illustrations for the Beam-Warming scheme follow in Section 4.

3. Proof of Theorem 13 and Corollary 15

In order to use the residue theorem, the holomorphy of the Kreiss-Lopatinskii determinant is needed. To this end, we want a nicer expression of $\det BK_{-r,m-1}(z)$ the Kreiss-Lopatinskii determinant. Clearly the multiplicativity of the determinant does not apply in the expression $\Delta_{\mathrm{KL}}(z) = \det BK_{-r,m-1}(z)$ since B and $K_{-r,m-1}(z)$ are non-square matrices. A first step consists of reducing the problem to a linear algebra formulation with square matrices to use the multiplicativity of the determinant. All along the current section, the assumptions (H0), (H1) and (H2), required to define the matrices $K_{i,j}$, the vector bundle $\mathcal{E}^s$ as well as its extension over $\overline{\mathcal{U}}$, are supposed to be fulfilled.

3.1. Reduction to a square formulation.

Let us fix $z \in \overline{\mathcal{U}}$. We recall that $\mathcal{E}^s(z)$ denotes the space of solutions $(\widetilde{U}_j(z))_{j \geqslant -r}$ to

$$z\widetilde{U}_j(z) = \sum_{k=-r}^{0} a_k \widetilde{U}_{j+k}(z),$$

for all $j \geqslant 0$ and with $a_{-r} \neq 0$.

Definition 20. Let E be a linear subspace of $\ell^2(\mathbb{N})$. Two matrices $B, D \in \mathcal{M}_{r,N}(\mathbb{C})$ (with $N \in \mathbb{N} \setminus \{0\}$ be any non-zero integer) are said to be equivalent, which we denote $B \sim_E D$, if and only if for all $U \in E$, one has $B\pi(U) = D\pi(U)$, where π is the canonical projection from ℓ^2 onto $\mathbb{C}^N$, keeping the N first components of U.

To act conveniently with elementary Gaussian operations, we use some specific notations in the following discussions. We denote $M[i : j, k : \ell]$ the matrix obtained by the extraction of the lines between i and j and the columns between k and ℓ of the matrix M (all indices are included). Similarly, we denote more shortly $M[k : \ell]$ for the entire columns between column k and column ℓ and $M[k]$ for the column k.

Lemma 21. *Let $N \geqslant r$ be an integer. Let $B \in \mathcal{M}_{r,N}(\mathbb{C})$ be a constant complex matrix such that $B[1 : r, 1 : r] \in \mathrm{GL}_r(\mathbb{C})$. Assume moreover that $|a_0| < 1$. For any $z \in \overline{\mathcal{U}}$, consider the associated linear subspace $\mathcal{E}^s(z)$. There exists a unique square matrix $C(z) \in \mathcal{M}_r(\mathbb{C})$ such that*

$$B \sim_{\mathcal{E}^s(z)} \begin{array}{c} \left(\begin{array}{ccc|c} 0 & \cdots & 0 & \\ \vdots & & \vdots & C(z) \\ 0 & \cdots & 0 & \end{array}\right) \begin{array}{l} \updownarrow r \\ \\ \\ \end{array} \\ \begin{array}{cc} \longleftrightarrow & \longleftrightarrow \\ N-r & r \end{array} \end{array}$$

Moreover, the components of $C(z)$ are polynomial functions of z and satisfy $\deg\det C(z) = N - r$.

Remark 22. Let us highlight our use of this lemma. Let $\ell \geqslant -r$ and $z \in \overline{\mathcal{U}}$ be fixed. From the basis (12), the columns of the matrix $K_{\ell,\ell+N-1}(z)$ take the form $\pi(U)$ for some $U \in \mathcal{E}^s(z)$ and π the canonical projection from $\ell^2(\mathbb{N})$ onto $\mathbb{C}^N$. Therefore, for any convenient matrices B and D with $B \sim_{\mathcal{E}^s(z)} D$, one has then $BK_{\ell,\ell+N-1}(z) = DK_{\ell,\ell+N-1}(z)$.
Now for the boundary matrix B defined in (7) and the matrix $D(z) = (0 \mid C(z))$ obtained by the lemma, the following computation by block is possible

$$\begin{aligned} \det(BK_{-r,m-1}(z)) &= \det(0K_{-r,m-r-1}(z) + C(z)K_{m-r,m-1}(z)) \\ &= \det(C(z)K_{m-r,m-1}(z)) \\ &= \det C(z) \det K_{m-r,m-1}(z). \end{aligned}$$

In other words, the product $BK_{-r,m-1}(z)$ is written as the product of two square matrices, so that the multiplicativity of the determinant can be applied.

Proof of Lemma 21.
Proof of existence: we proceed by induction for j going from 0 to $N - r$. At each step, we construct a matrix $B^{(j)}$ which satisfies the following induction hypotheses:

(a) $B \sim_{\mathcal{E}^s(z)} B^{(j)}$.
(b) the j first columns of $B^{(j)}$ are zero.
(c) every component of $B^{(j)}$ is polynomial of z.
(d) every component of $B^{(j)}[r+1+j : N]$ are independent of z.
(e) the degree of $\det B^{(j)}[j+1 : j+r]$ is j.

Initialization: we define $B^{(0)} \overset{def}{=} B$ which satisfies the five induction hypotheses. The induction hypotheses from (a) to (d) are trivially satisfied. The induction hypothesis (e) is satisfied because, $\det B[1 : r, 1 : r] \in \mathbb{C}^*$ which is a non zero constant polynomial.

Induction: we suppose true the induction hypotheses for some $j \in [\![0 : N - r - 1]\!]$ and we want to prove it for $j + 1$. Let us define $B^{(j+1)} \overset{def}{=} B^{(j)} - \widetilde{B}^{(j)}$ where

$$\widetilde{B}^{(j)} \overset{def}{=} \begin{pmatrix} B^{(j)}_{1,j+1} \\ \vdots \\ B^{(j)}_{r,j+1} \end{pmatrix} \Big(\underbrace{0 \cdots 0}_{j} \quad \underbrace{1 \quad \tfrac{a_{-r+1}}{a_{-r}} \quad \cdots \quad \tfrac{a_0 - z}{a_{-r}}}_{r+1} \quad \underbrace{0 \cdots \cdots 0}_{N-(r+1)-j} \Big).$$

By construction of $\widetilde{B}^{(j)}$, we have $\widetilde{B}^{(j)}U = 0$ for all $U \in \mathcal{E}^s(z)$, because the product of the previous row matrix and every vector $U \in \mathcal{E}^s(z)$ is equal to zero. Then, we have $B^{(j+1)} \sim_{\mathcal{E}^s(z)} B^{(j)}$ and by $(a)_j$, we have $(a)_{j+1}$. Moreover, by $(b)_j$, the first j columns of $B^{(j+1)}$ are zero because those columns in the construction of $B^{(j+1)}$ are unchanged and by construction of $B^{(j+1)}$, the $(j+1)$-th column is vanished. Then we have $(b)_{j+1}$. By construction, components of $B^{(j)}$ are added and multiplied by z or by real coefficients, then we have $(c)_{j+1}$. By $(d)_j$, the last $N-(r+1)-j$ columns of $B^{(j+1)}$ are independent of z because we do not take into account those columns in the construction of $B^{(j+1)}$, then we have $(d)_{j+1}$.

Finally, we have to find the degree of $\det B^{(j+1)}[j+2:j+1+r]$. We use the multilinearity and the alternating property of the determinant. We work on block matrices and find

$$\det B^{(j+1)}[j+2:j+1+r] = \det\left(B^{(j+1)}[j+2:j+r] \,\middle|\, B^{(j)}[j+1+r] - \frac{a_0-z}{a_{-r}} B^{(j)}[j+1]\right).$$

Since the matrix $B^{(j)}[j+1+r]$ is independent of z by hypothesis $(\mathrm{d})_j$, the degree of the polynomial $\frac{a_0-z}{a_{-r}}\det\left(B^{(j+1)}[j+2:j+r] \mid B^{(j)}[j+1]\right)$ is greater than the degree of $\det\left(B^{(j+1)}[j+2:j+r] \mid B^{(j)}[j+1+r]\right)$, then it is sufficient to find the degree of $\frac{a_0-z}{a_{-r}}\det\left(B^{(j+1)}[j+2:j+r] \mid B^{(j)}[j+1]\right)$. Moreover, the k-th column of $B^{(j+1)}[j+2:j+r]$ for $k \in [\![1:r-1]\!]$ is $B^{(j)}[j+1+k] - \frac{a_{-r+k}}{a_{-r}} B^{(j)}[j+1]$.

Then, by alternating property of the determinant, we have

$$\begin{aligned}
&-\frac{a_0-z}{a_{-r}}\det\left(B^{(j+1)}[j+2:j+r] \mid B^{(j)}[j+1]\right)\\
=&-\frac{a_0-z}{a_{-r}}\det\left(B^{(j)}[j+2:j+r] \mid B^{(j)}[j+1]\right)\\
=&-\frac{a_0-z}{a_{-r}}(-1)^{r+1}\det\left(B^{(j)}[j+1:j+r]\right).
\end{aligned}$$

By hypothesis $(\mathrm{e})_j$, we know that the polynomial $\det\left(B^{(j)}[j+1:j+r]\right)$ is of degree j, then the polynomial $-\frac{a_0-z}{a_{-r}}(-1)^{r+1}\det\left(B^{(j)}[j+1:j+r]\right)$ is of degree $j+1$ and $(\mathrm{e})_{j+1}$ follows.

Conclusion: the matrix $B^{(N-r)}$ gives the result, where

$$C(z) \overset{def}{=} B^{(N-r)}[1:r, N-r+1:N].$$

Proof of uniqueness: assume that C and C' are satisfying the lemma. Then

$$B \sim_{\mathcal{E}^s(z)} \underbrace{\left(\begin{array}{ccc|c} 0 & \cdots & 0 & \\ \vdots & & \vdots & C(z) \\ 0 & \cdots & 0 & \end{array}\right)}_{=D} \sim_{\mathcal{E}^s(z)} \underbrace{\left(\begin{array}{ccc|c} 0 & \cdots & 0 & \\ \vdots & & \vdots & C'(z) \\ 0 & \cdots & 0 & \end{array}\right)}_{=D'}.$$

On the one side, we have $(D-D')\pi_{|\mathcal{E}^s(z)} = 0$ and on the other side, because the $N-r$ first columns are zero, we have $(D-D')_{|\mathrm{Vect}(e_1,\dots,e_{N-r})} = 0$ where $e_1,\dots,e_N$ is the canonical basis of $\mathbb{C}^N$. Let us introduce the linear subspace $F \overset{def}{=} \ker A$ where

$$A \overset{def}{=} \begin{pmatrix} a_{-r} & \dots & (a_0-z) & & 0 \\ & \ddots & & \ddots & \\ 0 & & a_{-r} & \dots & (a_0-z) \end{pmatrix} \in \mathcal{M}_{N-r,N}(\mathbb{C}).$$

We have $F \cap \mathrm{Vect}(e_1,\dots,e_{N-r}) = \{0\}$. Indeed, if $x \in F \cap \mathrm{Vect}(e_1,\dots,e_{N-r})$, then $Ax = 0$ and $x = (x_1,\dots,x_{N-r},0,\dots,0)^\top$. By solving the triangular system

$$\begin{cases} a_{-r}x_1 + a_{-r+1}x_2 + \dots + a_{-1}x_r + (a_0-z)x_{r+1} = 0 \\ \vdots \\ a_{-r}x_{N-r-1} + a_{-r+1}x_{N-r} = 0 \\ a_{-r}x_{N-r} = 0, \end{cases}$$

we find $x = 0$. Moreover, $\dim F = r$ by rank-nullity theorem, then we have

$$F \oplus \mathrm{Vect}(e_1,\dots,e_{N-r}) = \mathbb{C}^N.$$

We want to show $(D-D')_{|F}=0$, and we know that $(D-D')\pi_{|\mathcal{E}^s}=0$. Let $x\in F$ and extend it to $\tilde{x}\in\mathcal{E}^s$. To that aim, it suffices to set recursively for all $j>N$,

$$\tilde{x}_j=\frac{1}{a_0-z}(-a_{-1}\tilde{x}_{j-1}-\cdots-a_{-r}\tilde{x}_{j-r}).$$

It follows that $(D-D')\pi(\tilde{x})=0$ and $(D-D')x=0$. Then, we have $(D-D')=0$ on $\mathbb{C}^N$. □

Remark 23. The uniqueness result is actually not needed for the next results.

In Section 4.2 (resp. Section 4.5), we perform the explicit algorithmic computations described above for the classic first-order upwind scheme (21) (resp. Beam-Warming scheme (25)).

3.2. **Holomorphy.**

Lemma 24. *Assume $|a_0|<1$. For all $\ell\in\mathbb{N}$ and for all $z\in\overline{\mathcal{U}}$, we have*

$$\frac{\det K_{\ell,\ell+r-1}(z)}{\det K_{0,r-1}(z)}=(-1)^{\ell r}\left(\frac{a_{-r}}{a_0-z}\right)^{\ell}. \tag{16}$$

Proof.
Case with only one root κ of multiplicity β. By Lemma 4 (Hersh), we know that $\beta=r$, but let keep β because it will be useful for the next step. We recall that

$$\det K_{0,r-1}=\begin{vmatrix}1 & 0 & \cdots & 0\\ \kappa & \kappa & \cdots & \kappa\\ \kappa^2 & 2\kappa^2 & \cdots & 2^{\beta-1}\kappa^2\\ \vdots & & & \vdots\\ \kappa^{r-1} & (r-1)\kappa^{r-1} & \cdots & (r-1)^{\beta-1}\kappa^{r-1}\end{vmatrix}. \tag{17}$$

We want to work on

$$\begin{aligned}\det K_{\ell,\ell+r-1}&=\begin{vmatrix}\kappa^{\ell} & \ell\kappa^{\ell} & \cdots & \ell^{\beta-1}\kappa^{\ell}\\ \kappa^{\ell+1} & (\ell+1)\kappa^{\ell+1} & \cdots & (\ell+1)^{\beta-1}\kappa^{\ell+1}\\ \vdots & & & \vdots\\ \kappa^{\ell+r-1} & (\ell+r-1)\kappa^{\ell+r-1} & \cdots & (\ell+r-1)^{\beta-1}\kappa^{\ell+r-1}\end{vmatrix}\\ &=\kappa^{\ell\beta}\begin{vmatrix}1 & \ell & \cdots & \ell^{\beta-1}\\ \kappa & (\ell+1)\kappa & \cdots & (\ell+1)^{\beta-1}\kappa\\ \vdots & & & \vdots\\ \kappa^{r-1} & (\ell+r-1)\kappa^{r-1} & \cdots & (\ell+r-1)^{\beta-1}\kappa^{r-1}\end{vmatrix}.\end{aligned}$$

We do some operations on columns to recover (17). For n from $\beta-1$ to 0, we replace the column C_n by $\sum_{k=0}^{n}(-\ell)^{n-k}\binom{n}{k}C_k$. After the transformation, the component in position (i,n), with $i\in[\![0:r-1]\!]$

and $n \in [\![0 : \beta - 1]\!]$, is

$$\begin{aligned}\sum_{k=0}^{n}(-\ell)^{n-k}\binom{n}{k}(\ell+i)^k\kappa^i &= \sum_{k=0}^{n}(-\ell)^{n-k}\binom{n}{k}\sum_{s=0}^{k}\binom{k}{s}\ell^{k-s}i^s\kappa^i \\ &= \sum_{k=0}^{n}\sum_{s=0}^{k}(-\ell)^{n-k}\binom{n}{s}\binom{n-s}{k-s}\ell^{k-s}i^s\kappa^i \\ &= \sum_{s=0}^{n}\binom{n}{s}i^s\kappa^i\sum_{k=s}^{n}(-\ell)^{n-k}\binom{n-s}{k-s}\ell^{k-s} \\ &= \sum_{s=0}^{n}\binom{n}{s}i^s\kappa^i\underbrace{\sum_{\tilde{k}=0}^{n-s}(-\ell)^{n-s-\tilde{k}}\binom{n-s}{\tilde{k}}\ell^{\tilde{k}}}_{=\delta_{n,s}} = i^n\kappa^i.\end{aligned}$$

This is exactly the component in (i, n) of the matrix $K_{0,r-1}(z)$.

General case. We can do the same operation on columns for each root. We take out $\kappa_1^{\ell\beta_1}\cdots\kappa_M^{\ell\beta_M}$, and for each root κ_j with $j \in [\![1 : M]\!]$, we vary n_{κ_j} from $\beta_j - 1$ to 0 and modify columns linked to κ_j. We regain matrix $K_{0,r-1}(z)$.

Conclusion. We proved

$$\frac{\det K_{\ell,\ell+r-1}}{\det K_{0,r-1}} = \kappa_1^{\ell\beta_1}\cdots\kappa_M^{\ell\beta_M}.$$

Observe that $a_0 - z \neq 0$ because $z \in \overline{\mathcal{U}}$ and $a_0 \in \mathbb{D}$. Therefore, by Vieta's formulas for the polynomial (11), we finally have

$$\kappa_1^{\beta_1}\cdots\kappa_M^{\beta_M} = (-1)^r\frac{a_{-r}}{a_0 - z}.$$ □

Lemma 24 implies the holomorphy on $\mathcal{U}$ and continuity on $\overline{\mathcal{U}}$ of the function in (16).

For the sake of completeness in the forthcoming proofs, we state hereafter two elementary lemmas. Both are easily deduced from classic properties of the winding number in complex analysis; cf. (Lang, 1999).

Lemma 25. *Let P and Q be two polynomials with $\deg P > \deg Q$. If the function $z \mapsto P(z)Q(z)^{-1}$ is holomorphic on $\mathcal{U}$ then $z \mapsto P(1/z)Q(1/z)^{-1}$ is meromorphic on $\mathbb{D}$ with only one pole, at the origin of order $\deg P - \deg Q$.*

Lemma 26. *Let f be a holomorphic function on $\mathcal{U}$ and continuous on $\overline{\mathcal{U}}$ and g be the function defined on $\overline{\mathbb{D}}^*$ by $g : z \mapsto f(1/z)$. Then, one has $\mathrm{Ind}_{g(\mathbb{S})}(0) = -\mathrm{Ind}_{f(\mathbb{S})}(0)$.*

3.3. **Explicit form of the intrinsic Kreiss-Lopatinskii determinant.** In the previous Lemmas 21 and 24, the assumption $|a_0| < 1$ is made. This is not a restriction since this is a consequence of the supplemented consistency assumption.

Lemma 27. *Let the scheme (4) be Cauchy-stable (H2) and consistent (H3), then $|a_0| < 1$.*

Proof. We have

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi}\sum_{k=-r}^{p}a_k e^{ik\xi}d\xi.$$

Integrating on the unit circle, by triangle inequality and Cauchy-stability, we have

$$|a_0| \leqslant \frac{1}{2\pi}\int_0^{2\pi}\underbrace{\left|\sum_{k=-r}^{p}a_k e^{ik\xi}\right|}_{\leqslant 1}d\xi \leqslant 1.$$

Let us assume now the identity $|a_0| = 1$, so that the equality occurs within the previous triangle inequality. Therefore, there exists a real-valued function g and a complex α such that for all $\xi \in \mathbb{R}$, we have $\sum_{k=-r}^{p} a_k e^{ik\xi} = \alpha g(\xi)$. Now at the point $\xi = 0$, we obtain $1 = \sum_{k=-r}^{p} a_k = \alpha g(0)$. Therefore, α is real, as well as the symbol $\gamma(\xi)$. Using the complex conjugate we deduce $\sum_{k=-r}^{p} a_k e^{ik\xi} - \sum_{k=-p}^{r} a_{-k} e^{ik\xi} = 0$ and then by the injectivity of the Fourier coefficients, it follows that $p = r$ and $a_k = a_{-k}$ for all $k \in \{1, \dots, r\}$. Finally, using now the consistency assumption, one has:

$$0 = \sum_{k=-r}^{p} k a_k = -\lambda \neq 0.$$

By this contradiction, the proof is complete. □

Now every piece can be put together to prove Theorem 13.

Proof of Theorem 13. Let us recall the function

$$\Delta : z \in \overline{\mathcal{U}} \mapsto \frac{\det BK_{-r,m-1}(z)}{\det K_{0,r-1}(z)} \in \mathbb{C}.$$

By Lemma 27, we will be able to use Lemma 21 and Lemma 24. With Lemma 21 and Remark 22, we express Δ as

$$\Delta(z) = \frac{\det C(z) \det K_{m-r,m-1}(z)}{\det K_{0,r-1}(z)} \tag{18}$$

where $C(z)$ is polynomial with respect to z. By Lemma 24, we have

$$\frac{\det K_{m-r,m-1}(z)}{\det K_{0,r-1}(z)} = (-1)^{r(m-r)} \left(\frac{a_{-r}}{a_0 - z} \right)^{m-r}. \tag{19}$$

By Lemma 27, a_0 cannot be a pole of Δ, then the function Δ can be written, for $z \in \overline{\mathcal{U}}$, as

$$\Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a_{-r}}{a_0 - z} \right)^{m-r}, \tag{20}$$

where $\det C(z)$ is a polynomial of z and $(a_0 - z)$ does not vanish because $z \in \overline{\mathcal{U}}$ and $a_0 \in \mathbb{D}$. □

The proof of Theorem 15 relies on the residue theorem to count the zeros of a holomorphic function.

Proof of Theorem 15. By Theorem 13, the function Δ is holomorphic on $\mathcal{U}$ and continuous on $\overline{\mathcal{U}}$. Let take the function

$$\widetilde{\Delta} : z \in \mathbb{D}^* \mapsto \Delta(1/z) \in \mathbb{C}.$$

The function $\widetilde{\Delta}$ is meromorphic on $\mathbb{D}$ with a pole in 0 of order r. By Lemma 21, we have $\deg \det C(z) = m$ because the r first columns of B form the identity matrix of size r which is invertible. By Lemma 25 with $P = \det C(z)$ and $Q = (-1)^{r(m-r)} \frac{(a_0 - z)^{m-r}}{a_{-r}^{m-r}}$, the only pole of $\widetilde{\Delta}$ is in 0 and of order

$$\deg \det C(z) - (m - r) = m - (m - r) = r.$$

Residue theorem on $\widetilde{\Delta}$. The function Δ is continuous on $\overline{\mathcal{U}}$, then the function $\widetilde{\Delta}$ is continuous on $\overline{\mathbb{D}}^*$. We can use the residue theorem on $\widetilde{\Delta}$ with the unit circle $\mathbb{S}$ as loop around 0. Then we have

$$\mathrm{Ind}_{\widetilde{\Delta}(\mathbb{S})}(0) = \#\mathrm{zeros}_{\widetilde{\Delta}}(\mathbb{D}) - \#\mathrm{poles}_{\widetilde{\Delta}}(\mathbb{D}).$$

Conclusion. We have $\#\mathrm{zeros}_{\Delta}(\mathcal{U}) = \#\mathrm{zeros}_{\widetilde{\Delta}}(\mathbb{D})$ and, by Lemma 26, we have $\mathrm{Ind}_{\widetilde{\Delta}(\mathbb{S})}(0) = -\mathrm{Ind}_{\Delta(\mathbb{S})}(0)$. It follows that

$$\#\mathrm{zeros}_{\Delta}(\mathcal{U}) = \underbrace{\#\mathrm{poles}_{\widetilde{\Delta}}(\mathbb{D})}_{r} - \mathrm{Ind}_{\Delta(\mathbb{S})}(0).$$

This concludes the proof. □

4. Numerical results

In this section, we first explain the numerical computation of the winding number of the origin in order to use Corollary 15 and Method 19. The simplest first order upwind scheme is then quickly treated, but for a general three-points boundary condition. Next, a main class of high-order boundary conditions, known as the simplified Lax-Wendroff procedure, is presented. They will be used together with the Beam-Warming scheme. After introducing the Beam-Warming scheme, we present computations of Kreiss-Lopatinskii determinant and numerical illustrations. Finally, we study the stability of discretizations where the physical boundaries are not aligned with the mesh.

4.1. **Computation of the winding number.** In the forthcoming numerical illustrations, the interest of Method 19 is showcased. Indeed, Theorem 15 makes the link between the number of zeros of a holomorphic function and the winding number of a curve which is easy to compute. In fact, as an integer is expected, the approximation of the winding number is generally more reliable, contrary to a real or complex computation because of machine precision.

When the origin is not on the curve, there are different ways to compute the winding number of the origin with respect to the curve. Either we can apply the definition and compute approximately a complex integral, or we can count the number of paths around the origin by using a polygonal approximation of the curve. The second approach is studied by García Zapata & Díaz Martín (2012, 2014) with a careful numerical treatment that consists in detecting the possible proximity of the curve to the origin. To that aim, the discretization of the curve is locally refined by a so-called "insertion procedure with control of singularity". Indeed, by the explicit formula (15) of the intrinsic Kreiss-Lopatinskii determinant, the curve $\Delta(\mathbb{S})$ is clearly parameterized by the Lipschitz function Δ, thus satisfies the required assumptions from the result in (García Zapata & Díaz Martín, 2012, 2014).

4.2. **Upwind scheme.** The easiest example of totally upwind scheme is the usual first-order upwind scheme defined, for $j \in \mathbb{N}$ and $n \in \mathbb{N}$, by

$$U_j^{n+1} = \lambda U_{j-1}^n + (1-\lambda) U_j^n \tag{21}$$

and supplemented at the boundary, for example, by

$$U_{-1}^n = b_0 U_0^n + b_1 U_1^n + b_2 U_2^n \tag{22}$$

with arbitrary coefficients b_0, b_1 and b_2. For that scheme, we have $r = 1$, $m = 3$ and the characteristic equation (11) reads

$$z\kappa(z) = \lambda + (1-\lambda)\kappa(z). \tag{23}$$

The scheme is Cauchy-stable for $\lambda \in]0,1]$ and one can check that for $0 < \lambda \leqslant 1$ and for $|z| > 1$, the root $\kappa(z)$ of (23) is in $\mathbb{D}$ by Lemma 4 (Hersh).

Let us now execute the computation of the intrinsic Kreiss-Lopatinskii determinant, as presented in Lemma 21 (here $N = 4$):

$$\begin{aligned}
B^{(0)} &= \begin{pmatrix} 1 & -b_0 & -b_1 & -b_2 \end{pmatrix} \\
\rightsquigarrow B^{(1)} &= \begin{pmatrix} 0 & -b_0 - \frac{1-\lambda-z}{\lambda} & -b_1 & -b_2 \end{pmatrix} \\
\rightsquigarrow B^{(2)} &= \begin{pmatrix} 0 & 0 & -b_1 + (b_0 + \frac{1-\lambda-z}{\lambda})\frac{1-\lambda-z}{\lambda} & -b_2 \end{pmatrix} \\
\rightsquigarrow B^{(3)} &= \begin{pmatrix} 0 & 0 & 0 & -b_2 + (b_1 - (b_0 + \frac{1-\lambda-z}{\lambda})\frac{1-\lambda-z}{\lambda})\frac{1-\lambda-z}{\lambda} \end{pmatrix}
\end{aligned}$$

It follows that $\det C(z) = -b_2 + (b_1 - (b_0 + \frac{1-\lambda-z}{\lambda})\frac{1-\lambda-z}{\lambda})\frac{1-\lambda-z}{\lambda}$.

Hence, the explicit formula (15) reads as follows:

$$\Delta(z) = (-1)^2 \det C(z) \left(\frac{a_{-r}}{a_0 - z}\right)^2 = -\frac{b_2\lambda^2}{(1-\lambda-z)^2} + \frac{b_1\lambda}{1-\lambda-z} - b_0 - \frac{1-\lambda-z}{\lambda}.$$

A similar computation can be achieved for boundary conditions with larger m and/or for totally upwind schemes with a larger stencil (see below for the Beam-Warming scheme).

4.3. **Simplified inverse Lax-Wendroff procedure.** As explained in (Tan & Shu, 2010) and (Vilar & Shu, 2015), the inverse Lax-Wendroff procedure is used to improve the consistency at the boundary by using the PDE to transform space derivative into time derivative. Namely, for the advection equation (1), the following relation holds, for $k \in \mathbb{N}^*$,

$$\frac{\partial^k u}{\partial x^k} = \frac{(-1)^k}{a^k}\frac{\partial^k u}{\partial t^k}.$$

By a Taylor expansion at order d to approximate $u(n\Delta t, j\Delta x)$ for $n \in \mathbb{N}$ and $j \in [\![-r : -1]\!]$, one can then define the ghost points used in the boundary condition (5) by

$$U_j^n = \sum_{k=0}^{d-1} \frac{(j\Delta x)^k}{k!}\frac{\partial^k u}{\partial x^k}(n\Delta t, 0) = \sum_{k=0}^{d-1} \frac{(j\Delta x)^k}{k!}(-1)^k \frac{g^{(k)}(n\Delta t)}{a^k}.$$

However, many derivatives of the datum g are required to obtain a high order approximation and the complexity then severely increases for multidimensional situations. As explained in (Vilar & Shu, 2015), the simplified inverse Lax-Wendroff procedure of order d with simplified order k_d that we call "Sk_dILWd" may be used when derivatives of g are not known. Therefore, the first $k_d - 1$ derivatives of g are considered and then for the next terms between order k_d and d, an extrapolation procedure is used.

Finally, the general formula is, for $j \in [\![-r : -1]\!]$, the following one

$$U_j^n = \sum_{k=0}^{k_d-1} \frac{(-j\Delta x)^k}{k!}\frac{g^{(k)}(n\Delta t)}{a^k} + \sum_{k=k_d}^{d-1} \frac{j^k}{k!}\sum_{s=0}^{d-1} p_{k,s}^{(d)} U_s^n, \tag{24}$$

where $\sum_{s=0}^{d-1} p_{k,s}^{(d)} U_s^n$ is an approximation of $\Delta x^k \frac{\partial^k u}{\partial x^k}(n\Delta t, 0)$ of order d.

4.4. **Beam-Warming scheme.** The Beam-Warming scheme with simplified inverse Lax-Wendroff of order 3 and simplified order 2 reads

$$\begin{cases} U_j^{n+1} = \dfrac{\lambda(\lambda-1)}{2}U_{j-2}^n + \lambda(2-\lambda)U_{j-1}^n + \dfrac{(\lambda-1)(\lambda-2)}{2}U_j^n, \\ U_{-1}^n = g(t^n) + \dfrac{\Delta x g'(t^n)}{a} + \dfrac{1}{2}(U_2^n - 2U_1^n + U_0^n), \\ U_{-2}^n = g(t^n) + \dfrac{2\Delta x g'(t^n)}{a} + 2(U_2^n - 2U_1^n + U_0^n), \\ U_j^0 = 0. \end{cases} \tag{25}$$

This scheme satisfies Assumptions (H1) and (H3). To have the Cauchy-stability assumption (H2), we study the symbol with respect to the CFL condition λ. From (25), the symbol is

$$\gamma(\xi) = \frac{\lambda(\lambda-1)}{2}e^{-2i\xi} + \lambda(2-\lambda)e^{-i\xi} + \frac{(\lambda-1)(\lambda-2)}{2}.$$

In the Figure 2, this symbol is represented for $\lambda = 1.8$.

Proposition 28. *The Beam-Warming scheme is Cauchy-stable if and only if* $0 < \lambda \leqslant 2$.

Even if it is a classic result, we recall the outline of the proof.

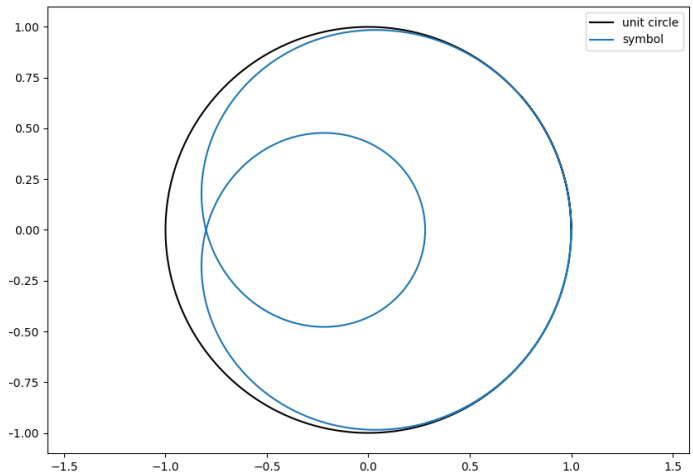

FIGURE 2. Symbol of Beam-Warming scheme for $\lambda = 1.8$.

Proof. While computing the symbol, we have, for all $\xi \in \mathbb{R}$,

$$\gamma(\xi) = \tfrac{(\lambda-1)(\lambda-2)}{2} + \lambda(2-\lambda)e^{-i\xi} + \tfrac{\lambda(\lambda-1)}{2}e^{-2i\xi} = e^{-i\xi}\left(\lambda(\lambda-1)\cos\xi + \lambda(2-\lambda) - (\lambda-1)e^{i\xi}\right).$$

Thus, the modulus of the symbol is after some easy computations

$$|\gamma(\xi)|^2 = 1 - \lambda(2-\lambda)(\lambda-1)^2(1-\cos\xi)^2.$$

To be Cauchy-stable, we must have $|\gamma(\xi)|^2 \leqslant 1$, so we want to have $\lambda(2-\lambda)(\lambda-1)^2(1-\cos\xi)^2 \geqslant 0$. Because $\lambda > 0$, then the condition is $\lambda \leqslant 2$. □

The non-degeneracy assumption (H0) is related to the value $r = 2$ for $\lambda \in]0,2] \setminus \{1\}$ and to the value $r = 1$ for $\lambda = 1$. This example will be useful to illustrate the theory, especially in the following subsection.

4.5. **Kreiss-Lopatinskii determinant computation for Beam-Warming scheme.** First, we compute the Kreiss-Lopatinskii determinant Δ_{KL} from Definition 11 for the Beam-Warming scheme with S2ILW3 boundary condition as in (25). Assuming that the roots of (11) are distinct for a given $|z| \geqslant 1$, we have

$$\Delta_{\mathrm{KL}}(z) = \det\begin{pmatrix} \kappa_1^{-2} - 2 + 4\kappa_1 - 2\kappa_1^2 & \kappa_2^{-2} - 2 + 4\kappa_2 - 2\kappa_2^2 \\ \kappa_1^{-1} - \frac{1}{2} + \kappa_1 - \frac{\kappa_1^2}{2} & \kappa_2^{-1} - \frac{1}{2} + \kappa_2 - \frac{\kappa_2^2}{2} \end{pmatrix}.$$

If there is one single root with multiplicity 2, then we have

$$\Delta_{\mathrm{KL}}(z) = \det\begin{pmatrix} \kappa_1^{-2} - 2 + 4\kappa_1 - 2\kappa_1^2 & -2\kappa_1^{-2} + 4\kappa_1 - 4\kappa_1^2 \\ \kappa_1^{-1} - \frac{1}{2} + \kappa_1 - \frac{\kappa_1^2}{2} & -\kappa_1^{-1} + \kappa_1 - \kappa_1^2 \end{pmatrix}.$$

In the rest of this section, we continue the example of the Beam-Warming scheme (25) so as to illustrate practically the algebraic transformation set up in Lemma 21.

For that scheme, the corresponding $\mathcal{Z}$-transformed equation (10) is, for $j \in \mathbb{N}$,

$$z\widetilde{U}_j(z) = a_{-2}\widetilde{U}_{j-2}(z) + a_{-1}\widetilde{U}_{j-1}(z) + a_0\widetilde{U}_j(z),$$

involving the coefficients $a_0 = \frac{(\lambda-1)(\lambda-2)}{2}$, $a_{-1} = \lambda(2-\lambda)$ and $a_{-2} = \frac{\lambda(\lambda-1)}{2}$.

Let us denote in the following lines $\alpha \overset{def}{=} \frac{-a_{-1}}{a_{-2}}$ and $\beta \overset{def}{=} \frac{z-a_0}{a_{-2}}$ so that the linear recurrence relation has now, for $j \in \mathbb{N}$, the form below:

$$\widetilde{U}_{j-2}(z) = \alpha\widetilde{U}_{j-1}(z) + \beta\widetilde{U}_j(z). \tag{26}$$

The considered boundary condition involves the following matrix:

$$B = \begin{pmatrix} 1 & 0 & -2 & 4 & -2 \\ 0 & 1 & -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix}$$

with dimensions $r = 2$ and $N = 5$. With the notations in the proof of Lemma 21, let us now construct the matrix $C(z) = B^{(3)}[1:2, 4:5]$. To that aim, we transform successively the matrix B so as to keep unchanged the vector $B(\widetilde{U}_{j-2}(z)\,\widetilde{U}_{j-1}(z)\,\widetilde{U}_j(z)\,\widetilde{U}_{j+1}(z)\,\widetilde{U}_{j+2}(z))^\top$ thanks to the recurrence relation (26). Hereafter are the steps:

$$\begin{aligned}
B^{(0)} &= \begin{pmatrix} 1 & 0 & -2 & 4 & -2 \\ 0 & 1 & -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix} \\
\rightsquigarrow B^{(1)} &= \begin{pmatrix} 0 & \alpha & -2+\beta & 4 & -2 \\ 0 & 1 & -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix} \\
\rightsquigarrow B^{(2)} &= \begin{pmatrix} 0 & 0 & -2+\beta+\alpha^2 & 4+\alpha\beta & -2 \\ 0 & 0 & -\frac{1}{2}+\alpha & 1+\beta & -\frac{1}{2} \end{pmatrix} \\
\rightsquigarrow B^{(3)} &= \begin{pmatrix} 0 & 0 & 0 & 4+\alpha\beta+\alpha(-2+\beta+\alpha^2) & -2+\beta(-2+\beta+\alpha^2) \\ 0 & 0 & 0 & 1+\beta+\alpha(-\frac{1}{2}+\alpha) & -\frac{1}{2}+\beta(-\frac{1}{2}+\alpha) \end{pmatrix}
\end{aligned}$$

From there, it follows that

$$C(z) = \begin{pmatrix} 4+\alpha\beta+\alpha(-2+\beta+\alpha^2) & -2+\beta(-2+\beta+\alpha^2) \\ 1+\beta+\alpha(-\frac{1}{2}+\alpha) & -\frac{1}{2}+\beta(-\frac{1}{2}+\alpha) \end{pmatrix},$$

and thus

$$\begin{aligned}
\det C(z) &= (4+\alpha\beta+\alpha(-2+\beta+\alpha^2))(-\tfrac{1}{2}+\beta(-\tfrac{1}{2}+\alpha)) \\
&\qquad - (1+\beta+\alpha(-\tfrac{1}{2}+\alpha))(-2+\beta(-2+\beta+\alpha^2)) \\
&= -\beta^3+\beta^2+2\beta-\alpha\beta^2/2+3\alpha\beta-\alpha^2\beta-2\alpha^2-\alpha^3/2.
\end{aligned}$$

The intrinsic Kreiss-Lopatinskii determinant explicit formula (20) (with here $m = 3$ and $r = 2$) is the following:

$$\Delta(z) = \frac{-1}{\beta}(-\beta^3+\beta^2+2\beta-\alpha\beta^2/2+3\alpha\beta-\alpha^2\beta-2\alpha^2-\alpha^3/2).$$

On Figure 3, the curve $\Delta(\mathbb{S})$ is represented successively for different values of the CFL parameter λ. The goal is to compute the winding number of 0, concerned with Corollary 15 in order to tackle stability thanks to the Theorem 3 (Kreiss). A premultiplication of the quantity Δ by a_{-2}^2 may reduce the order of magnitude of the curves, without changing the winding number. The left and right figures correspond to the case with or without rescaling.

By Theorem 15, we have $\#\text{zeros}_\Delta = r - \text{Ind}_{\Delta(\mathbb{S})}(0)$ but after dividing Δ by z^r:

$$\mathring{\Delta} : z \mapsto \Delta(z)/z^r,$$

we obtain $\#\text{zeros}_\Delta = -\text{Ind}_{\mathring{\Delta}(\mathbb{S})}(0)$, because $\text{Ind}_{\mathring{\Delta}(\mathbb{S})}(0) = \text{Ind}_{\Delta(\mathbb{S})}(0) - r$, see Figure 4.

A particular situation occurs for $\lambda = 1$, since $a_{-2} = 0$ and assumption (H0) fails if we consider $r = 2$. In that case, the equation (10) reads $\widetilde{U}_{j-1}(z) = z\widetilde{U}_j(z)$ which is the Beam-Warming scheme for $\lambda = 1$ after $\mathcal{Z}$-transform. Finally, in that case, we find $\det C(z) = (\frac{1}{2} + z(-1 + z(\frac{1}{2} - z)))$ that we must multiply by $\frac{1}{\beta^2} = \frac{1}{z^2}$ to find the Kreiss-Lopatinskii determinant (because $m = 3$, $r = 1$ and $\beta = \frac{z-a_0}{a_{-1}} = z$).

All these computations can be done for different boundary conditions and after drawing the curves, the winding number can be computed, as explained in Section 4.1, to tackle stability and that the purpose of the following subsection.

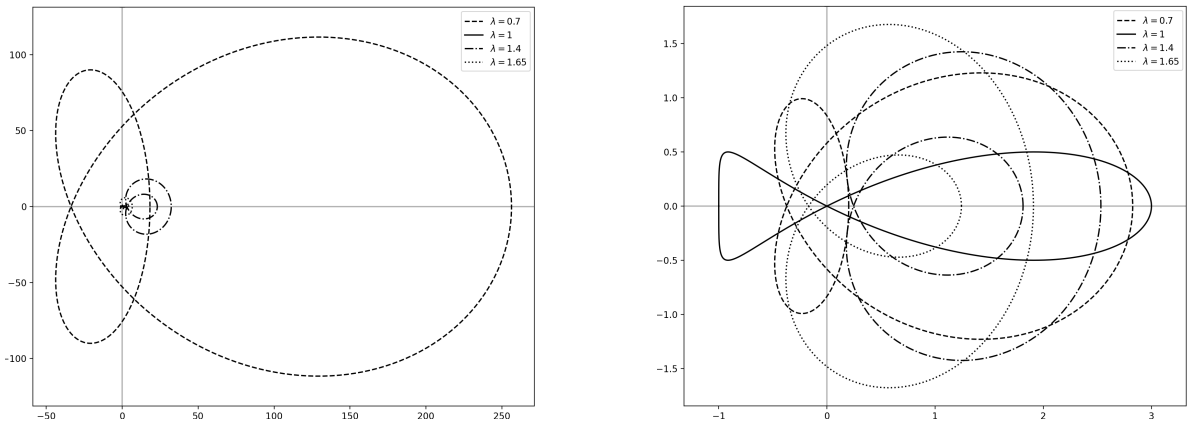

FIGURE 3. Kreiss-Lopatinskii Determinant Δ when z is on $\mathbb{S}$ for scheme (25) for $\lambda \in \{0.7, 1, 1.4, 1.65\}$ (left) and the rescaled one $a_{-2}^2\Delta$ (right).

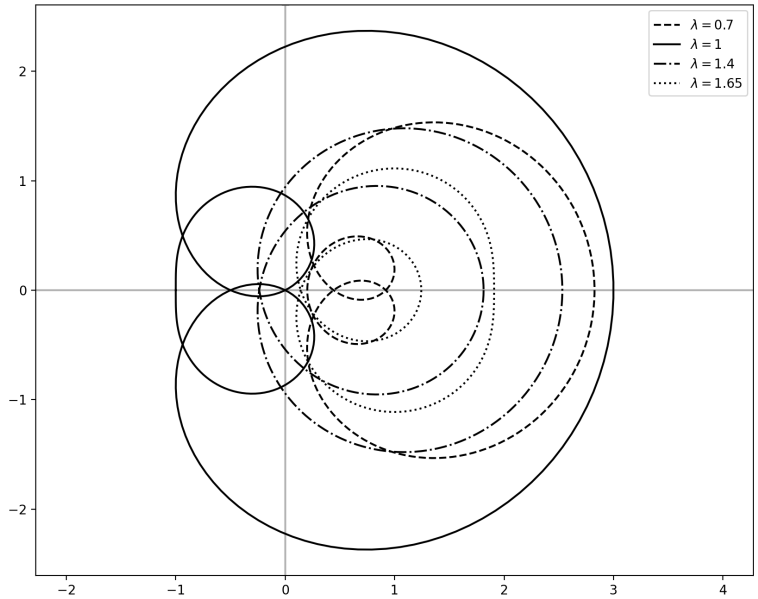

FIGURE 4. Rescaled Kreiss-Lopatinskii determinant $\frac{a_{-2}^2\Delta}{z^2}$ for z in $\mathbb{S}$.

4.6. **Numerical illustration.** Figure 4 may help to tackle the stability of the scheme (25) as we said in Section 2.4, indeed, as we said in Section 4.1, one can compute the winding number using a numerical procedure (García Zapata & Díaz Martín, 2012) and draw the winding number with respect to λ, as seen in Figure 5 for the case S2ILW3. It simplifies the observation of the number of zeros of the Kreiss-Lopatinskii determinant. Hence, the numerical experiments indicate that the scheme (25) is strongly stable for $\lambda \in]0, 1[$ but also for $\lambda \in]1.52, 1, 78[$ approximately, but is unstable outside these domains.

Moreover, instead of taking the Y-axis to represent the number of zeros of the Kreiss-Lopatinskii determinant and having a step function, one can draw areas and compute it for other simplified inverse Lax-Wendroff boundary conditions (defined by the equation (24)) as done in Figure 6. Note that the stability domain contains a full interval of the form $]0, \lambda_\star[$ (except for the S1ILW4 scheme), but also another disjoint interval included in $]1, 2[$ (except for the S2ILW4 scheme). This property may be used to increase the speed of the computations.

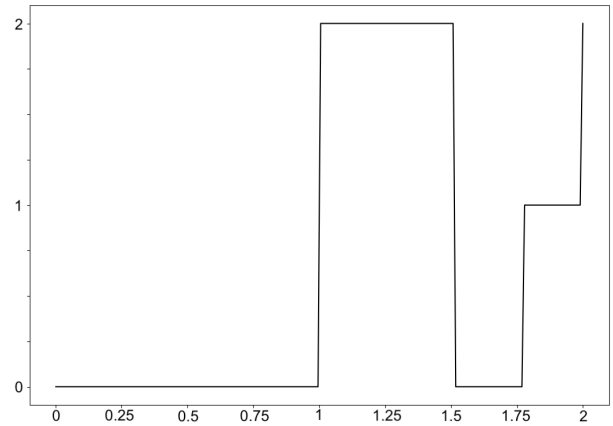

FIGURE 5. Number of zeros of Kreiss-Lopatinskii determinant with respect to λ for Beam-Warming scheme (25) with S2ILW3 boundary condition.

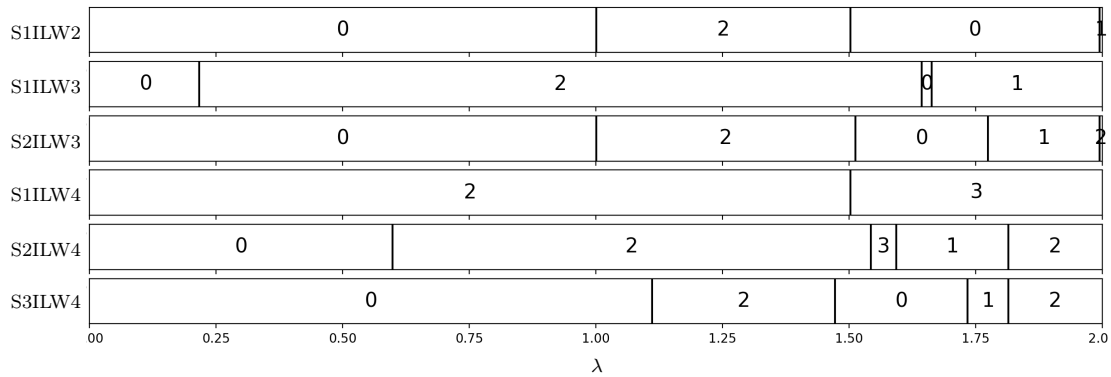

FIGURE 6. Number of zeros of Kreiss-Lopatinskii determinant for Beam-Warming scheme with different simplified inverse Lax-Wendroff boundary with respect to λ.

All the figures can be easily reproduced using the code available in the GitHub repository (Le Barbenchon, 2023), it can be used as a Python library, following the link : https://doi.org/10.5281/zenodo.7773742. The algorithm is really quick (less than one minute of computation achieved on a standard laptop). Moreover, our procedure provides sharp results, directly available on ℓ^2. In particular, contrary to numerical investigations of stability which are based on the computation of the spectral radius, no arbitrary truncation of (quasi-)Toeplitz matrices is needed.

4.7. **Misalignment between boundaries and grid points.** Motivated for example by solving multidimensional problems discretized on a Cartesian grid, or of one-dimensional problems with moving boundaries as well, a usual idea consists in extrapolating the physical boundary condition to the first boundary points. This idea may be combined with the inverse Lax-Wendroff procedure in order to improve the accuracy at the boundary; cf. (Dakin *et al.*, 2018), (Vilar & Shu, 2015) and (Li *et al.*, 2016). As an archetype for such a situation, we consider hereafter a simple misalignment between the left physical boundary and the first numerical grid point. The advection equation (1) is set on the space domain $[x_\sigma, 1]$:

$$\begin{cases} \partial_t u + a\partial_x u = 0, & t \geqslant 0, x \in [x_\sigma, 1], \\ u(t, x_\sigma) = g(t), & t \geqslant 0, \\ u(0, x) = f(x), & x \in [x_\sigma, 1]. \end{cases} \tag{27}$$

The space discretization $j\Delta x$ for $j \in \mathbb{Z}$, does not take into account the point x_σ, so that one may write $x_\sigma = (j_\sigma + \sigma)\Delta x$ for some integer $j_\sigma \in \mathbb{Z}$ and the gap (generally non-zero) $\sigma \in [-\frac{1}{2}, \frac{1}{2}[$. The scheme (4)-(5)-(6) is then implemented for $j \geqslant j_\sigma$ only, but with r ghost points at $j_\sigma - 1, \ldots, j_\sigma - r$. For simplicity in the presentation and by translational invariance, we assume from now on that $j_\sigma = 0$. We obtain the discretization represented on Figure 7.

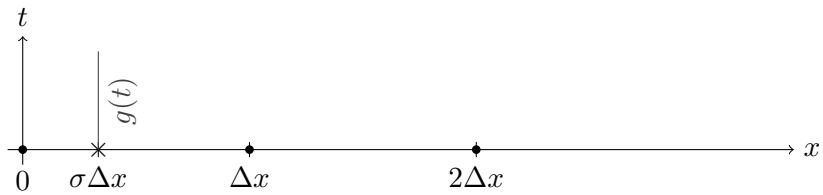

FIGURE 7. Representation of the mesh.

As explained above, because of the misalignment between the mesh and the boundary position, the simplified inverse Lax-Wendroff procedure (24) presented above has to be slightly adapted; cf. (Vilar & Shu, 2015). The numerical boundary condition reads

$$U_j^n = \sum_{k=0}^{k_d-1} \frac{(-(j+\sigma)\Delta x)^k}{k!} \frac{g^{(k)}(n\Delta t)}{a^k} + \sum_{k=k_d}^{d-1} \frac{(j+\sigma)^k}{k!} \sum_{s=0}^{d-1} p_{k,s}^{(d)} U_s^n, \quad j \in [\![-r : -1]\!].$$

We perform the stability analysis of the above scheme, according to the values of both the CFL parameter λ and the gap parameter σ. For example, with the Beam-Warming scheme (25) supplemented with the numerical boundary condition S2ILW3 at the point x_σ, the procedure based on the Kreiss-Lopatinskii determinant counts the number of zeros of the Kreiss-Lopatinskii determinant. The corresponding results are represented on Figure 8. Of course, on the line $\sigma = 0$, we recover the results obtained on Figure 5.

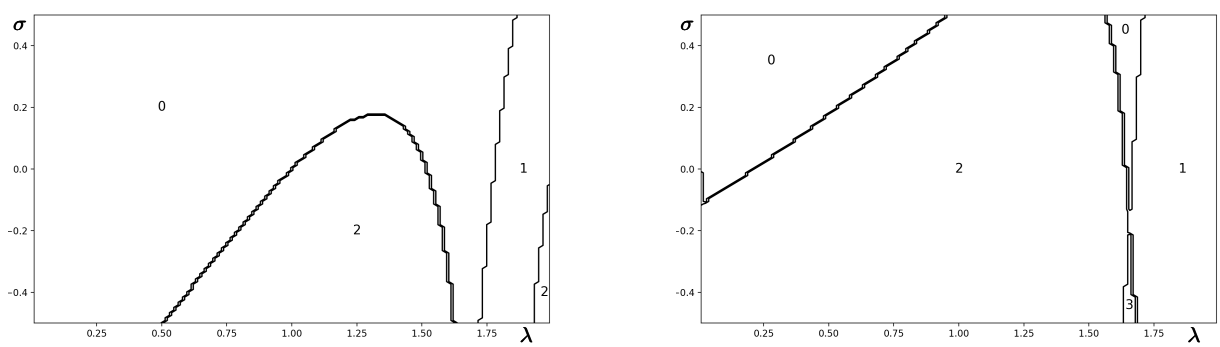

FIGURE 8. Stability of the Beam-Warming (25) with S2ILW3 boundary condition (left) and with S1ILW3 boundary condition (right).

Let us now consider a very simple application of the above results, considering the advection equation in 2D on a parallelogram domain (specified later) with a velocity field aligned with the x axis. Using a Cartesian grid in both directions x and y, the numerical boundary condition will generally not coincide exactly at the grid points and the use of (S)ILW method may appear useful to maintain the order of the scheme. However, it is then mandatory to retain a CFL number for which any of the considered values for the parameter σ along the boundary belong to the stability condition. Following the same lines of discussion as for the one-dimensional case, we consider

hereafter the next problem where the direction y coincide (artificially) with the parameter σ and where the first reference grid cell is again $x_\sigma = 0$ ($j_\sigma = 0$).

$$\begin{cases} \partial_t u(t,x,y) + a\partial_x u(t,x,y) = 0, & t \geqslant 0, y \in [-1,1], x \in [y\Delta x, +\infty[, \\ u(t,y,y) = g(t,y) & t \geqslant 0, y \in [-1,1], \\ u(0,x,y) = 0 & y \in [-1,1], x \in [y\Delta x, +\infty[. \end{cases}$$

In the simulations, the velocity is $a = 1$, the boundary condition is $g(t,y) = e^{-200(t-0.25)^2}$ and the initial datum is $f \equiv 0$. The numerical solution is computed at time $T = 0.3$ using the Beam-Warming scheme with S2ILW3, and with $N = 1000$ grid points in the (truncated) x-direction. The Figure 9 represents the amplitude of the numerical solution with respect to the space variable x and to the gap $\sigma = y$, the discrete solution being truncated beyond the value 1 so that unstable boundary oscillations appear as white areas. The two black lines represents the computational domain of Figure 8 to confront the left image of Figure 8 and the images of Figure 9. We observe a good agreement between the corresponding stable/unstable values of σ in Figure 8 and Figure 9.

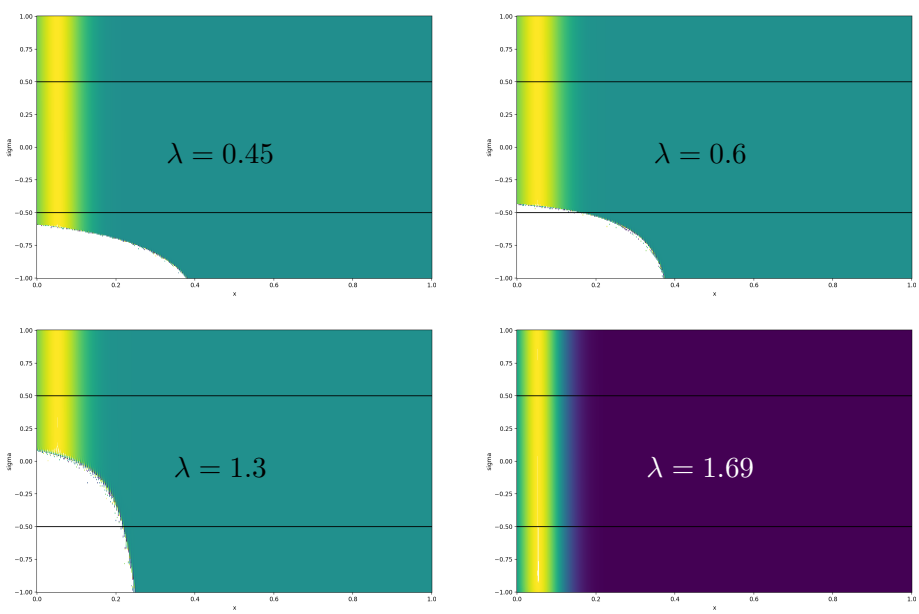

FIGURE 9. Numerical simulation of Beam-Warming scheme with S2ILW3 for CFL number $\lambda \in \{0.45, 0.6, 1.3, 1.69\}$.

5. Future directions

The main drawback of the present theoretical and numerical results is the restriction to the class of totally upwind schemes. This assumption enables a specific simple analysis of the Kreiss-Lopatinskii determinant, using the explicit formula (20), and a numerical strategy to conclude to the existence of eigenvalue or generalized eigenvalue. In this way, it answers the stability issue. This is only an initial effort on the method of designing efficient and automatic numerical tools for stability analysis based on the Kreiss-Lopatinskii determinant. A first extension of the present work is the extension to the case of one-time step explicit schemes without the totally upwind assumption that limits the application of our approach to second-order schemes, see Iserles & Strang (1983). Such an extension is natural but not straightforward because of the loss of Lemma 21: the intrinsic Kreiss-Lopatinskii determinant cannot be reduced easily into a formulation involving square matrices.

Another challenging issue is the treatment of multistep schemes and multistep boundary conditions as well. In this direction, explicit schemes may be the most practicable case because many theoretical tools remain available (Hersh, Kreiss...). The difference is the dependence on z in the coefficients of the characteristic equation (11). Indeed, each coefficient is a polynomial in z of degree s where s is the number of time steps. Hence, an explicit formula of a Kreiss-Lopatinskii is more difficult to compute. In another direction, for implicit schemes or for more general boundary conditions, such as absorbing boundary conditions (Engquist & Majda, 1977) and (Ehrhardt, 2010) or transparent boundary conditions (Arnold *et al.*, 2003) and (Coulombel, 2019), it seems to be more challenging to obtain such a user-friendly theory.

Acknowledgments

This work has been partially supported by ANR project NABUCO, ANR-17-CE40-0025 and by Centre Henri Lebesgue, program ANR-11-LABX-0020-0.

References

Arnold, A., Ehrhardt, M., & Sofronov, I. (2003) Discrete transparent boundary conditions for the Schrödinger equation: fast calculation, approximation, and stability. *Commun. Math. Sci.*, **1**, 501–556.

Beam, R. M., & Warming, R. F. (1993) The Asymptotic Spectra of Banded Toeplitz and Quasi-Toeplitz Matrices. *SIAM J. Sci. Comput.*, **14**, 971–1006.

Benzoni-Gavage, S., & Serre, D. (2007) *Multidimensional hyperbolic partial differential equations.* Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press.

Borovykh, N., & Spijker, M. N. (2000) Resolvent conditions and bounds on the powers of matrices, with relevance to numerical stability of initial value problems. *J. Comput. Appl. Math.*, **125**, 41–56.

Coulombel, J.-F. (2013) Stability of finite difference schemes for hyperbolic initial boundary value problems. *Page 146 of: HCDTE lecture notes. Part I. Nonlinear hyperbolic PDEs, dispersive and transport equations.* AIMS Ser. Appl. Math., vol. 6. Am. Inst. Math. Sci. (AIMS), Springfield, MO.

Coulombel, J.-F. (2019) Transparent numerical boundary conditions for evolution equations: derivation and stability analysis. *Ann. Fac. Sci. Toulouse Math.*, **6**, 259–327.

Courant, R., Friedrichs, K., & Lewy, H. (1928) Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, **100**, 32–74.

Crank, J., & Nicolson, P. (1947) A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Math. Proc. Cambridge Philos. Soc.*, **43**, 50–67.

Dakin, G., Després, B., & Jaouen, S (2018) Inverse Lax–Wendroff boundary treatment for compressible Lagrange-remap hydrodynamics on cartesian grids. *J. Comput. Phys.*, **353**, 228–257.

Ehrhardt, M. (2010) Absorbing boundary conditions for hyperbolic systems. *Numer. Math., Theory Methods Appl.*, **3**, 295–337.

Engquist, B., & Majda, A. (1977) Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, **31**, 629–651.

García Zapata, J. L., & Díaz Martín, J. C. (2012) A geometric algorithm for winding number computation with complexity analysis. *J. Complexity*, **28**, 320–345.

García Zapata, J. L., & Díaz Martín, J. C. (2014) Finding the number of roots of a polynomial in a plane region using the winding number. *Comput. Math. Appl.*, **67**, 555–568.

Gasquet, C., & Witomski, P. (1999) *Fourier Analysis and Applications: Filtering, Numerical Computation, Wavelets.* Texts in Applied Mathematics, vol. 30. Springer-Verlag, New York.

Godunov, S. K., & Ryabenkii, V. S. (1963) Spectral stability criteria for boundary-value problems for non-self-adjoint difference equations. *Russian Mathematical Surveys*, **18**, 1–12.

Goldberg, M. (1977) On a boundary extrapolation theorem by Kreiss. *Math. Comp.*, **31**, 469–477.

Gustafsson, B. (2008) *High Order Difference Methods for Time Dependent PDE*. Springer series in computational mathematics, no. 38. Springer-Verlag, Berlin.

Gustafsson, B., Kreiss, H.-O., & Sundström, A. (1972) Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems. II. *Math. Comp.*, **26**, 649–649.

Gustafsson, B., Kreiss, H.-O., & Oliger, J. (2013) *Time-dependent problems and difference methods*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. John Wiley & Sons, Inc., Hoboken, NJ.

Hersh, R. (1963) Mixed problems in several variables. *J. Math. Mech.*, **12**, 317–334.

Iserles, A., & Strang, G. (1983) The optimal accuracy of difference schemes. *Trans. Amer. Math. Soc.*, **277**, 779–803.

Lang, S. (1999) Complex analysis. *Graduate Texts in Mathematics,* Fourth edition, vol. **103**. Springer-Verlag, New York.

Lax, P. D., & Richtmyer, R. D. (1956) Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, **9**, 267–293.

Le Barbenchon, P. (2023) *Boundaryscheme: Python package for numerical scheme with boundary.* https://doi.org/10.5281/zenodo.7773742.

Li, T., Shu, C.-W., & Zhang, M. (2016) Stability analysis of the inverse Lax–Wendroff boundary treatment for high order upwind-biased finite difference schemes. *J. Comput. Appl. Math.*, **299**, 140–158.

Li, T., Shu, C.-W., & Zhang, M. (2017) Stability analysis of the inverse Lax–Wendroff boundary treatment for high order central difference schemes for diffusion equations. *J. Sci. Comput.*, **70**, 576–607.

Li, T., Lu, J., & Shu, C.-W. (2022) Stability analysis of inverse Lax–Wendroff boundary treatment of high order compact difference schemes for parabolic equations. *J. Comput. Appl. Math.*, **400**, 113711.

Métivier, G., & Zumbrun, K. (2004) Symmetrizers and continuity of stable subspaces for parabolic-hyperbolic boundary value problems. *Discrete Contin. Dyn. Syst.*, **11**, 205–220.

Spijker, M. N., Tracogna, S., & Welfert, B. D. (2003) About the sharpness of the stability estimates in the Kreiss matrix theorem. *Math. Comp.*, **72**, 697–714.

Tan, S., & Shu, C.-W. (2010) Inverse Lax-Wendroff procedure for numerical boundary conditions of conservation laws. *J. Comput. Phys.*, **229**, 8144 – 8166.

Thuné, M. (1986) Automatic GKS stability analysis. *SIAM J. Sci. Statist. Comput.*, **7**, 959–977.

Trefethen, L. N. (1983) Group velocity interpretation of the stability theory of Gustafsson, Kreiss, and Sundström. *J. Comput. Phys.*, **49**, 199–217.

Trefethen, L. N. (1984) Instability of difference models for hyperbolic initial boundary value problems. *Comm. Pure Appl. Math.*, **37**, 329–367.

Trefethen, L. N., & Embree, M. (2005) *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton, N.J: Princeton University Press.

Vilar, F., & Shu, C.-W. (2015) Development and stability analysis of the inverse Lax Wendroff boundary treatment for central compact schemes. *ESAIM Math. Model. Numer. Anal.*, **49**, 39–67.

Wu, L. (1995) The semigroup stability of the difference approximations for initial-boundary value problems. *Math. Comp.*, **64**, 71–88.

Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France.
Email address: benjamin.boutin@univ-rennes.fr
Email address: pierre.lebarbenchon@univ-rennes.fr

IMAG, Inria, Univ Montpellier, CNRS, Montpellier, France
Email address: nicolas.seguin@inria.fr