



HAL
open science

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

Benjamin Boutin, Pierre Le Barbanchon, Nicolas Seguin

► **To cite this version:**

Benjamin Boutin, Pierre Le Barbanchon, Nicolas Seguin. On the stability of totally upwind schemes for the hyperbolic initial boundary value problem. 2022. hal-03732720v1

HAL Id: hal-03732720

<https://hal.science/hal-03732720v1>

Preprint submitted on 21 Jul 2022 (v1), last revised 16 Jun 2023 (v3)

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 scheme in 1D with numerical boundary conditions. The underlying approximated continuous problem is a hyperbolic partial differential equation. Our approach is based on the Uniform Kreiss-Lopatinskii Condition, using linear algebra and complex analysis to count the number of zeros of the associated determinant. 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 propose an efficient numerical strategy to decide 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 later on. We restrict our attention to the approximation of a rightgoing linear transport equation set on the positive real axis:

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

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 data g is prescribed.

Let us first remind some general ideas and historical context. As a central idea in numerical analysis, the Lax equivalence theorem [19] 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 handled with the Fourier symbolic analysis and the so-called Von-Neumann stability analysis, the case with boundaries is a little more tricky. The reason for that is the irruption of another kind of instability arising because of the presence of (unphysical) numerical boundary conditions. The normal mode analysis, as used in [13], is the classical way to comprehend this kind of instabilities. Deepening this analysis with resolvent estimates and Laplace transform leads to the notion of *GKS-stability* [17] (sometimes called *strong stability*, see Definition 2 later). This notion is actually the most robust one concerning the stability of initial boundary value numerical methods, since the stability property is stable by perturbations and makes use of the same norms for the solution and for the data itself, accessing for example to the

Date: July 21, 2022.

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.

possible handling of nonlinearities. The Kreiss theorem (Theorem 3) gives a necessary and sufficient condition for strong stability by the use of the Uniform Kreiss-Lopatinskii Condition. When this condition fails, the instabilities are interpreted thanks to numerical wave packets with bad group velocities and/or amplitudes, as done by Trefethen [26, 27]. Some sketches of the strong stability theory will be unfolded later on but we refer the interested reader to the monographs [15] and [16] for a more complete overview of the GKS-theory.

The GKS-theory is not used so often in the numerical analysis litterature. The reason for that is probably that the checking of the Uniform Kreiss-Lopatinskii Condition requires the search for the vanishing points of the Kreiss-Lopatinskii determinant in the complex domain $\{|z| \geq 1\}$. Except for some particular numerical schemes and boundary conditions, this determinant is not known explictely. Let us however mention the work by [25] for automatic GKS-stability investigation and some other works concerned with a very similar problematic that is the stability of viscous shock profiles [2] thanks to Evans functions.

Other tools are met. Among them, the matrix form of the considered numerical scheme, considered for a large but finite grid, gives rise to the computation of the corresponding spectral properties. For our case of study, the involved matrix T_J of size J has a Toeplitz or a Quasi-Toeplitz structure due to boundary conditions, making a specific analysis possible (see [3]). The question, to obtain a uniform bound for the set of powers of T_J , is related to the Kreiss matrix theorem [28, Chap 18]. Nevertheless, the corresponding bounds have to be also uniform with respect to the dimension J in order to obtain finally stability estimates. A first idea consists in considering only arbitrary fixed grids (with $J = 20$ or 30), expecting then that it gives the significant informations for the infinite dimensional case, see [9]. A more strong tool, directly related to resolvant estimates, is the ε -pseudospectrum [4, 23]. Nonetheless, the link between GKS-instabilities and the pseudospectrum of the matrix for a given scheme is still not completely understood, to our knowledge.

In the present work, the selected strategy is based on the Uniform Kreiss-Lopatinskii Condition, thanks to the Kreiss-Lopatinskii determinant, that is a function of the complex parameter z defined for $|z| \geq 1$. It consists in first finding the zeros of the function on the unit circle $\{|z| = 1\}$ and then counting the number of zeros on the domain $\{|z| > 1\}$, in order finally to conclude on stability with the Kreiss theorem. Our main point is to secure the numerical computation of winding numbers after a reshaping of the Kreiss-Lopatinskii determinant.

The outline of the paper is as follows. In this introductory 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 then state our main results. In Section 3, we prove these results thanks to an explicit formula for the Kreiss-Lopatinskii determinant. 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 $\bar{\mathcal{U}} = \{z \in \mathbb{C}, |z| \geq 1\}$ its closure.

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

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

with integers $r, p \geq 0$. 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 choosen with respect to some CFL condition $\lambda = a\Delta t/\Delta x \leq \lambda_{\text{CFL}}$ discussed later on.

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

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

The current set of assumptions on the finite difference scheme (2) 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 $|\gamma(\xi)| \leq 1$ for all $\xi \in \mathbb{R}$.
- (H3) *consistent* and at least of order 1, meaning that

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

When dealing with the discrete schemes set over the full line $j \in \mathbb{Z}$, the algebraic characterization of the Cauchy-stability follows classically from the Fourier analysis and makes use of the symbol γ . This method is known as the Von Neumann analysis (see [7] and [8]). In the scalar case, it reduces to a geometric property concerning the following closed complex curve of importance hereafter.

Definition 1. The *symbol curve* Γ is the closed complex parametrized 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 IBVP (1) with the discrete scheme (2), some r additional ghost points are needed to take into account the left boundary condition and to fully define the discrete approximation. There are several ways to design a numerical boundary condition, using Dirichlet or Neumann extrapolation procedures [14], absorbing boundary conditions [11] and [10], transparent boundary conditions [1] and [6]... In this paper, we will focus on the simplified inverse Lax-Wendroff procedure (see [29], [21], [9] and Section 4.2).

We are interested in schemes of the form

$$\begin{cases} U_j^{n+1} = \sum_{k=-r}^0 a_k U_{k+j}^n, & j \in \mathbb{N}, n \in \mathbb{N}, \\ U_j^n = \sum_{k=0}^{m-1} b_{j,k} U_k^n + g_j^n, & j \in \llbracket -r; -1 \rrbracket, n \in \mathbb{N}, \\ U_j^0 = f_j, & j \in \mathbb{N}. \end{cases} \quad (4)$$

$$\quad (5)$$

$$\quad (6)$$

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 data g . With the vector notation $U = (U_{-r}^n \cdots U_{m-1}^n)^T$ and $G =$

$(g_{-r}^n \cdots g_{-1}^n)^T$, the boundary equation (5) reads also equivalently as $BU = G$ with the following matrix

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

1.3. Classical results about strong stability. The GKS theory (see [17]) handles the discrete IBVP (4)-(5)-(6) with a zero initial data. We refer the reader to [30] and [5] for more recent development on semigroup estimates in order to deduce the stability of the discrete IBVP (4)-(5)-(6) with a non zero initial data from the one with a zero initial data.

The corresponding notion of stability for the boundary problem make 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 $\ell^2(\{-r, \dots, -1\} \cup \mathbb{N})$, denoted shortly ℓ^2 .

We are now in position to define the so-called strong stability.

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 solution satisfies

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

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

The Kreiss theorem gives us a necessary and sufficient condition to conclude on the stability. We provide hereafter a condensated formulation of this theorem, obtained from [17, Thm 5.1] combined with [16, Lem 13.1.4] or with [15, Def 2.23].

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 is related to the absence of zeros for the so-called Kreiss-Lopatinskii determinant (see later Definition 10 and [16]). These zeros are called eigenvalues or generalized eigenvalues (see later Definition 15 and Definition 16) and correspond to modal instabilities. Our study 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 will introduce two versions of the Kreiss-Lopatinskii determinant and construct an algebraic reformulation of them. Thanks to Theorem 3, a numerical procedure can thus be proposed to conclude.

2.1. Stable subspace $\mathcal{E}^s(z)$ and matrix representation. First, we assume (H1) and study the solutions to the interior equation:

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

To study this equation, the \mathcal{Z} -transform (see [12, Lesson 40]) is applied. This transformation is defined for $(x_n)_n \in \ell^2(\mathbb{N})$ such that $x_0 = 0$ and $z \in \mathcal{U}$ by $\tilde{x}(z) = \sum_{n \geq 0} z^{-n} x_n$. The previous equation then becomes

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

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:

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

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 injecting the solution $U_j^n = z^n \kappa^j$ in the interior equation (9).

The following lemma is in the spirit of a classical result by Hersh [18] and is concerned with 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} .*

Let us remark that 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 even *dissipative*, that is if its symbol γ satisfies

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

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 that property is that in that case one has $\mathbb{S} \cap \Gamma = \{1\}$.

The lemma above is illustrated in Figure 1. The first two lines 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 5. Setting the assumption (H1) aside, meaning with a nonzero 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. The proof of this result and of Lemma 4 is omitted.

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 Hersh lemma, the space $\mathcal{E}^s(z)$ is generated by the following r vectors:

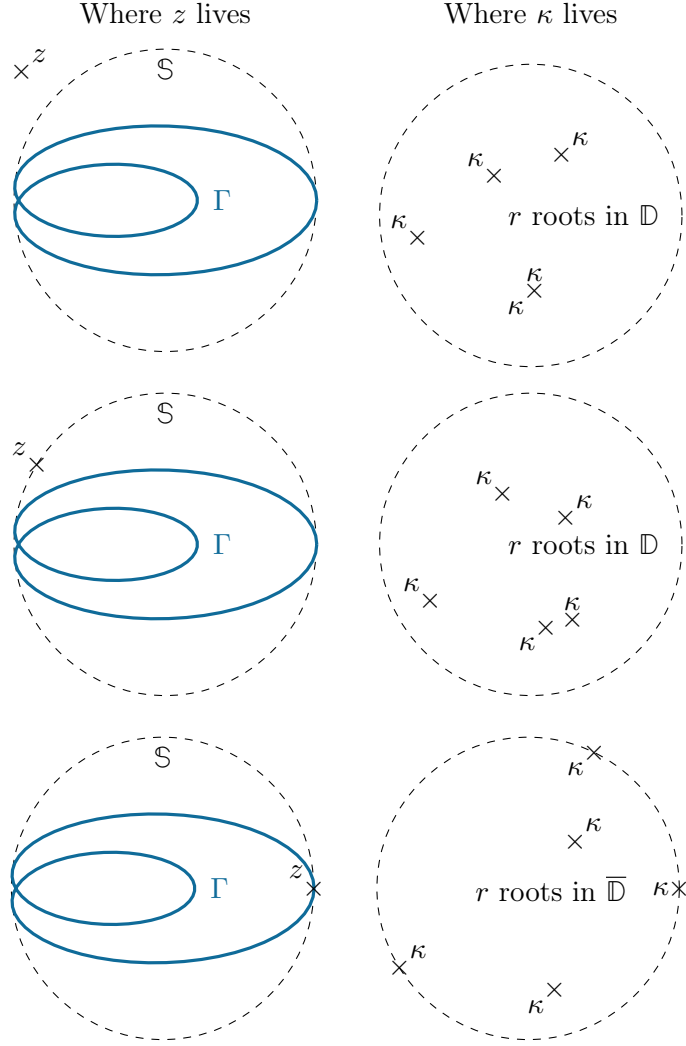


FIGURE 1. Illustration of Lemma 4: case $|z| > 1$ (first line), case $|z| = 1$ and $z \notin \Gamma$ (second line) and case $z \in \Gamma$ where Lemma 4 doesn't hold (third line).

$$\left\{ \begin{pmatrix} \kappa_1^{-r} \\ \vdots \\ \kappa_1^{-1} \\ 1 \\ \kappa_1 \\ \kappa_1^2 \\ \kappa_1^3 \\ \vdots \end{pmatrix}, \begin{pmatrix} -r\kappa_1^{-r} \\ \vdots \\ -\kappa_1^{-1} \\ 0 \\ \kappa_1 \\ 2\kappa_1^2 \\ 3\kappa_1^3 \\ \vdots \end{pmatrix}, \dots, \begin{pmatrix} (-r)^{\beta_1-1}\kappa_1^{-r} \\ \vdots \\ (-1)^{\beta_1-1}\kappa_1^{-1} \\ 0 \\ \kappa_1 \\ 2^{\beta_1-1}\kappa_1^2 \\ 3^{\beta_1-1}\kappa_1^3 \\ \vdots \end{pmatrix}, \begin{pmatrix} \kappa_2^{-r} \\ \vdots \\ \kappa_2^{-1} \\ 1 \\ \kappa_2 \\ \kappa_2^2 \\ \kappa_2^3 \\ \vdots \end{pmatrix}, \begin{pmatrix} -r\kappa_2^{-r} \\ \vdots \\ -\kappa_2^{-1} \\ 0 \\ \kappa_2 \\ 2\kappa_2^2 \\ 3\kappa_2^3 \\ \vdots \end{pmatrix}, \dots \right. \quad (12)$$

$$\left(\begin{array}{c} (-r)^{\beta_2-1} \kappa_2^{-r} \\ \vdots \\ (-1)^{\beta_2-1} \kappa_2^{-1} \\ 0 \\ \kappa_2 \\ 2^{\beta_2-1} \kappa_2^2 \\ 3^{\beta_2-1} \kappa_2^3 \\ \vdots \end{array} \right), \dots, \left(\begin{array}{c} \kappa_M^{-r} \\ \vdots \\ \kappa_M^{-1} \\ 1 \\ \kappa_M \\ \kappa_M^2 \\ \kappa_M^3 \\ \vdots \end{array} \right), \dots, \left(\begin{array}{c} (-r)^{\beta_M-1} \kappa_M^{-r} \\ \vdots \\ (-1)^{\beta_M-1} \kappa_M^{-1} \\ 0 \\ \kappa_M \\ 2^{\beta_M-1} \kappa_M^2 \\ 3^{\beta_M-1} \kappa_M^3 \\ \vdots \end{array} \right)$$

where $\kappa_1, \dots, \kappa_M$ of multiplicity β_1, \dots, β_M are the solutions to (11), with $\beta_1 + \dots + \beta_M = r$. (We forget 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 $-r \leq i \leq j$.

Remark 6. 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 7. 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 in 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 $(z_n) \in \mathcal{U}^{\mathbb{N}}$ 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

$$\forall j \in \{1, 2\}, \quad \kappa_j(z_n) \xrightarrow{n \rightarrow \infty} \kappa_j(z_\infty)$$

but

$$K_{0,3}(z_n) = \begin{pmatrix} 1 & 1 \\ \kappa_1(z_n) & \kappa_2(z_n) \\ \kappa_1^2(z_n) & \kappa_2^2(z_n) \\ \kappa_1^3(z_n) & \kappa_2^3(z_n) \end{pmatrix} \xrightarrow{n \rightarrow \infty} K_{0,3}(z_\infty) = \begin{pmatrix} 1 & 0 \\ \kappa_1(z_\infty) & \kappa_1(z_\infty) \\ \kappa_1^2(z_\infty) & 2\kappa_1^2(z_\infty) \\ \kappa_1^3(z_\infty) & 3\kappa_1^3(z_\infty) \end{pmatrix}.$$

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 [5, Thm 4.3]. This author proves in addition that this vector bundle $\mathcal{E}^s(z)$ can even be continuously extended over $\widetilde{\mathcal{U}}$, thus considering $z \in \mathbb{D}$ as well (see also [22] 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} . At such points z_0 the Hersh lemma doesn't hold anymore. This situation is depicted on the third line 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 [5, 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 8. *Under assumptions (H0), (H1) and (H2), the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over \mathcal{U} and can be extended in a unique way 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 easy because the r roots that come from the inside of the unit open disk have to be selected. Indeed, if there is some $\kappa_0(z_0)$ on the unit circle, one have 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 boundary condition in order to study the solution in $\mathcal{E}^s(z)$ to the \mathcal{Z} -transformed version, that is

$$\forall j \in \llbracket -r; -1 \rrbracket, \quad \widetilde{U}_j(z) = \sum_{k=0}^{m-1} b_{j,k} \widetilde{U}_k(z) + \widetilde{g}_j(z). \quad (13)$$

While injecting the fundamental solutions to $\mathcal{E}^s(z)$ into (13), we obtain a system of r equations where the coefficients $(\alpha_i)_{i=1}^r$ are the scalar unknowns. They are the coefficients of the solution to (13) written in the basis (12) of $\mathcal{E}^s(z)$.

Remark 9. 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:

$$\forall j \in \{-2, -1\}, \quad \alpha_1 \kappa_1^j + \alpha_2 \kappa_2^j = \sum_{k=0}^{m-1} b_{j,k} (\alpha_1 \kappa_1^k + \alpha_2 \kappa_2^k) + \widetilde{g}_j.$$

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 10 (Kreiss-Lopatinskii determinant). The *Kreiss-Lopatinskii determinant* is the complex-valued function defined for $|z| \geq 1$ by:

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

Let us introduce the *intrinsic Kreiss-Lopatinskii determinant* Δ by the following informal discussion. The above Kreiss-Lopatinskii determinant is actually not entirely 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 and already discussed after Remark 7. 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). So except for crossing roots, the same holds for the quantity $\Delta_{\text{KL}}(z)$ since the matrix B is constant and the determinant itself is a symmetric function. Now it is known that the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over \mathcal{U} , continuous over $\bar{\mathcal{U}}$, and so we thus 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 $\det K_{0,r-1}(z)$. By this way, the same permutation or combination of the vectors of the basis (12) is involved in both computations.

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

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

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 $z \in \bar{\mathcal{U}}$, any $U \in \mathcal{E}^s(z)$ solution to (13) satisfies the uniform estimate

$$\|\tilde{U}\| \leq C\|\tilde{g}\|.$$

From the Parseval identity for the \mathcal{Z} -transform, this inequality gives directly the first necessary half-part of the strong stability estimate (8). We refer the reader to [16] for a more detailed presentation.

2.3. Main result. The main result of the present paper is the next theorem.

Theorem 12 (Main theorem). *Assume (H0), (H1), (H2) and (H3). The intrinsic Kreiss-Lopatinskii determinant is given by*

$$\forall z \in \bar{\mathcal{U}}, \quad \Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a_{-r}}{a_0 - z} \right)^{m-r} \quad (15)$$

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

Some important properties of Δ follow from the main theorem and from its proof. Namely, let us highlight right now that Δ is holomorphic on \mathcal{U} , is continuous on $\bar{\mathcal{U}}$, depends on $\mathcal{E}^s(z)$ but not on the choice of a basis (what justifies the *intrinsic* denomination of that quantity). By the definition (14), it also appears that the function Δ shares the same zeros as Δ_{KL} .

Another property, important for the forthcoming applications, lies in the next corollary 14 and involves the following important geometrical object:

Definition 13. 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 14 (Main corollary). *Assume (H0), (H1), (H2) and (H3). If $0 \notin \Delta(\mathbb{S})$ then the equation $\Delta(z) = 0$ has exactly $r - \text{Ind}_{\Delta(\mathbb{S})}(0)$ zeros in \mathcal{U} .*

Here above $\text{Ind}_{\Delta(\mathbb{S})}(0)$ denotes the winding number of the complex point 0 with respect to the closed oriented curve $\Delta(\mathbb{S})$. This previous corollary is the fundamental piece to the following numerical procedure to conclude on stability.

2.4. Numerical procedure. As already seen in the Kreiss Theorem 3, the strong stability can be characterized thanks to the notion of eigenvalue and generalized eigenvalue for the boundary problem. The definition of generalized eigenvalue is not universal, the following one comes from [16, Def.12.2.2] but one can find a slightly different one in [15, Def 2.2]. The difference will be discussed afterwards.

Definition 15 (Eigenvalue). Let z be a complex number. If $|z| \geq 1$, $\Delta(z) = 0$ and $(\widetilde{U}_j(z)) \in \ell^2$ then z is called an *eigenvalue*.

Definition 16 (Generalized eigenvalue). Let z_0 be a complex number with $|z_0| = 1$. If $\Delta(z_0) = 0$ and $(\widetilde{U}_j(z_0)) \notin \ell^2$ 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 Hersh Lemma, 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 is restricted to the unit circle and not the outside.

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) regroup all eigenvalues. Indeed, for type (i) and (ii), by Hersh lemma 4, 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 line of Figure 1 and type (ii) corresponds to the second line.

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 [13] and described in [27].

If z is of type (iii) or (iv), there exists a $\kappa_0(z)$ root of (11) on the unit circle because $z \in \Gamma$. It corresponds to the third line of Figure 1. The distinction between (iii) and (iv) comes from the expression of $(\widetilde{U}_j(z))$ where the coefficient in front of $\kappa_0(z)$ can be zero or not.

Our definition of generalized eigenvalue from [16, Def.12.2.2] corresponds, as we already said, to type (iv) whereas the definition from [15, Def 2.2] regroup type (iii) and (iv).

Now, the main corollary 14 can be reformulated as the following.

Corollary 17 (Main corollary – eigenvalue formulation). *Assume (H0), (H1), (H2) and (H3). If $0 \notin \Delta(\mathbb{S})$ then the scheme has $r - \text{Ind}_{\Delta(\mathbb{S})}(0)$ eigenvalues in \mathcal{U} (type (i)).*

In practice, this corollary is really helpful to conclude on stability while using the Kreiss Theorem 3. Indeed, while drawing the curve $\Delta(\mathbb{S})$, we have the following case disjunction:

- 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 $r - \text{Ind}_{\Delta(\mathbb{S})}(0)$ zeros of Δ in \mathcal{U} by Corollary 14 (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 Hersh Lemma, 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 $z_0 \in \mathbb{S}$.

In summary, by Theorem 3 (Kreiss), if $0 \in \Delta(\mathbb{S})$ then the scheme is not stable, and if $0 \notin \Delta(\mathbb{S})$, Corollary 14 can be used to conclude. Some illustrations for the Beam-Warming scheme follow in Section 4.

3. PROOF OF MAIN THEOREM AND MAIN COROLLARY

In order to use the residue theorem, the holomorphy of the Kreiss-Lopatinskii determinant is mandatory. Clearly the multiplicativity of the determinant used for appearing in the definition $\Delta_{\text{KL}}(z) = \det BK_{-r,m-1}(z)$ does not apply since B and $K_{-r,m-1}(z)$ are rectangle matrices. A first step consists thus to reduce the problem to a linear algebra squared formulation. All along the current section, the assumptions (H0), (H1) and (H2), required to define the matrices $K_{i,j}(z)$, the vector bundle \mathcal{E}^s as well as its extension over \bar{U} , are met.

3.1. Reduction to a squared formulation. Let us fix $z \in \bar{U}$. We remind that $\mathcal{E}^s(z)$ denotes the space of solutions $(U_j)_{j \geq -r}$ of

$$zU_j = \sum_{k=-r}^0 a_k U_{j+k} \quad \forall j \geq 0 \quad \text{with } a_{-r} \neq 0.$$

Definition 18. 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 nonzero 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 $\ell^2(\mathbb{N})$ 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 19. Let $N \geq r$ be an integer. Let $B \in \mathcal{M}_{r,N}(\mathbb{C})$ be a constant complex matrix such that $B[1 : r, 1 : r] \in \text{GL}_r(\mathbb{C})$. Assume moreover that $|a_0| < 1$.

For any $z \in \bar{U}$, consider the associated linear subspace $\mathcal{E}^s(z)$. There exists a unique matrix $C(z) \in \mathcal{M}_r(\mathbb{C})$ such that

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

$$\begin{array}{c} \leftarrow \quad \quad \quad \rightarrow \\ N-r \quad \quad \quad r \end{array}$$

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

Remark 20. Let us now highlight the possible use of the previous lemma. Let $\ell \geq -r$ and $z \in \bar{U}$ 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 thanks to the lemma, the following computation by block is possible

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

Proof of Lemma 19.

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 $B^{(0)} \stackrel{\text{def}}{=} B$ 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 $j \in \llbracket 0; N-r-1 \rrbracket$ and we want to prove it for $j+1$.

Let us define $B^{(j+1)} \stackrel{\text{def}}{=} B^{(j)} - \widetilde{B^{(j)}}$ where

$$\widetilde{B^{(j)}} \stackrel{\text{def}}{=} \begin{pmatrix} B_{1,j+1}^{(j)} \\ \vdots \\ B_{r,j+1}^{(j)} \end{pmatrix} \left(\underbrace{0 \cdots 0}_j \underbrace{1 \frac{a_{-r+1}}{a_{-r}} \cdots \frac{a_0-z}{a_{-r}}}_{r+1} \underbrace{0 \cdots 0}_{N-(r+1)-j} \right).$$

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 don't touch 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

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

Since $B^{(j)}[j+1+r]$ is independent of z by (d) _{j} , the degree of $\frac{a_0-z}{a_{-r}} \det (B^{(j+1)}[j+2 : j+r] \mid B^{(j)}[j+1])$ is greater than the degree of $\det (B^{(j+1)}[j+2 : j+r] \mid B^{(j)}[j+1+r])$, then it is sufficient to find the degree of $\frac{a_0-z}{a_{-r}} \det (B^{(j+1)}[j+2 : j+r] \mid B^{(j)}[j+1])$.

Moreover, the k -th column of $B^{(j+1)}[j+2 : j+r]$ for $k \in \llbracket 1; r-1 \rrbracket$ 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

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

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

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

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

Proof of uniqueness: assume that C and C' satisfy 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 $(D - D')|_{\text{Vect}(e_1, \dots, e_{N-r})} = 0$ where e_1, \dots, e_N is the canonical basis of \mathbb{C}^N .

Let us introduce the linear subspace $F \stackrel{\text{def}}{=} \ker A$ where

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

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

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

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

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

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 . \square

Remark 21. The uniqueness result is actually not mandatory for the next results.

In Section 4.4, this lemma is made explicit on the Beam-Warming scheme (22).

3.2. Holomorphy.

Lemma 22. Assume $|a_0| < 1$. For all $\ell \in \mathbb{N}$ and for all $z \in \bar{U}$, we have

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

Proof. Step 1: Case with only one root κ of multiplicity β .

By Hersh lemma, we know that $\beta = r$, but let keep β because it will be useful for the next step. We recall that

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

We want to work on

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

We do some operations on columns to meet (17) again. 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 $i \in \llbracket 0; r-1 \rrbracket$ and $n \in \llbracket 0; \beta-1 \rrbracket$, 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 $K_{0,r-1}(z)$.

Step 2: 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 $j \in \llbracket 1; M \rrbracket$, we vary n_{κ_j} from $\beta_j - 1$ to 0 and modify columns linked to κ_j . We regain matrix $K_{0,r-1}(z)$.

Step 3: 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 $z \in \bar{\mathcal{U}}$ 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 22 implies the holomorphy on \mathcal{U} and continuity on $\bar{\mathcal{U}}$ of the function in (16).

For the sake of simplicity, we recall hereafter the two following classical lemmas, useful later on.

Lemma 23. *Let P and Q be two polynomials with $\deg P > \deg Q$. If the function $z \mapsto \frac{P(z)}{Q(z)}$ is holomorphic on \mathcal{U} then $z \mapsto \frac{P(1/z)}{Q(1/z)}$ is meromorphic on \mathbb{D} with one pole in 0 of order $\deg P - \deg Q$.*

Lemma 24. *Let f be a holomorphic function on \mathcal{U} and continuous on $\bar{\mathcal{U}}$. We define $g : z \mapsto f(1/z)$ on $\bar{\mathbb{D}}^*$. Then, we have $\text{Ind}_g(\mathbb{S})(0) = -\text{Ind}_f(\mathbb{S})(0)$.*

3.3. Explicit form of the intrinsic Kreiss-Lopatinskii determinant. In the previous Lemmas 19 and 22, the assumption $|a_0| < 1$ is made. Actually this is not a restriction since this is a consequence of the supplemented consistency assumption.

Lemma 25. *Let the scheme (4) be Cauchy-stable (H2) and consistent (H3), then $|a_0| < 1$.*

Proof. We have

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

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

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

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 $1 = \sum_{k=-r}^p a_k = \alpha g(0)$. So 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:

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

Contradiction. □

Now every piece can be put together to prove the main theorem.

Proof of Theorem 12. Let us recall the function

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

By Lemma 25, we will be able to use Lemma 19 and Lemma 22. With Lemma 19 and Remark 20, we express Δ as

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

where $C(z)$ is polynomial with respect to z .

By Lemma 22, we have

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

By Lemma 25, a_0 cannot be a pole of Δ , then the function Δ can be written as

$$\forall z \in \bar{\mathcal{U}}, \quad \Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a-r}{a_0-z} \right)^{m-r}. \quad (20)$$

where $\det C(z)$ is a polynomial of z and $(a_0 - z)$ doesn't vanish because $z \in \bar{\mathcal{U}}$ and $a_0 \in \mathbb{D}$. \square

The proof of Corollary 14 relies on the residue theorem to count the zeros of a holomorphic function.

Proof of Corollary 14. By Theorem 12, the function Δ is holomorphic on \mathcal{U} and continuous on $\bar{\mathcal{U}}$. Let take the function

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

Step 1: *The function $\tilde{\Delta}$ is meromorphic on \mathbb{D} with a pole in 0 of order r .*

By Lemma 19, 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 23 with $P = \det C(z)$ and $Q = (-1)^{r(m-r)} \frac{(a_0-z)^{m-r}}{a-r}$, the only pole of $\tilde{\Delta}$ is in 0 and of order $\deg \det C(z) - (m-r) = m - (m-r) = r$.

Step 2: *Residue theorem on $\tilde{\Delta}$*

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

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

Step 3: *Conclusion*

We have $\#\text{zeros}_{\Delta}(\mathcal{U}) = \#\text{zeros}_{\tilde{\Delta}}(\mathbb{D})$ and, by Lemma 24, we have $\text{Ind}_{\tilde{\Delta}(\mathbb{S})}(0) = -\text{Ind}_{\Delta(\mathbb{S})}(0)$. It follows that

$$\#\text{zeros}_{\Delta}(\mathcal{U}) = \underbrace{\#\text{poles}_{\tilde{\Delta}}(\mathbb{D})}_r - \text{Ind}_{\Delta(\mathbb{S})}(0).$$

This concludes the proof. \square

4. NUMERICAL RESULTS

In the section, we first introduce how to compute numerically the winding number of 0 in order to use the Corollary 14. Secondly, we present the SILW procedure used in the examples to define the boundary condition. Then, we introduce the Beam-Warming scheme, 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 numerical illustrations, the interest of Corollary 14 is showcased, indeed on the one hand the number of zeros of a function and the winding number of a curve are linked and on the other hand, numerically, it is easier to compute a winding number because it is an integer. In fact, if an integer is expected, the approximation of the computer is generally convincing contrary to a real or complex computation that can't tell if 10^{-16} is the same than zero.

When the origin is not on the curve, there are different ways to compute the winding number of the origin with respect to a curve. Neither we can apply the definition and compute approximately a complex integral, or we can count the number of path around the origin by using a polygonal approximation of the curve. In [31] and [32], they use the second approach and can detect if the origin is too close to the curve. Thereafter, they refine the discretization of the curve to be sure that the computation gives the correct result, they call it ‘‘insertion procedure with control of singularity for the number of roots’’. Indeed, thank to the explicit formula (15) of the intrinsic Kreiss-Lopatinskii determinant, we can easily prove that Δ is parametrized by a Lipschitz function in order to apply the robust results in [31] and [32].

4.2. Simplified inverse Lax-Wendroff procedure. As explained in [24] and [29], 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

$$\forall k \in \mathbb{N}^*, \quad \frac{\partial^k u}{\partial x^k} = \frac{(-1)^k}{a^k} \frac{\partial^k u}{\partial t^k}.$$

By a Taylor expansion at order d to approximate $u(n\Delta t, j\Delta x)$ for $n \in \mathbb{N}$ and $j \in \llbracket -r; -1 \rrbracket$, one can then consider the following numerical boundary condition

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

However, many derivatives of the data g are required to get high order approximation and the complexity then severely increases for multidimensional situations. As explained in [29], the simplified inverse Lax-Wendroff procedure of order d with simplified order k_d that we call ‘‘Sk_dILW d ’’ 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 the following one

$$\forall j \in \llbracket -r; -1 \rrbracket, \quad U_j^n = \sum_{k=0}^{k_d-1} \frac{(-j\Delta x)^k}{k!} \frac{g^{(k)}(n\Delta t)}{a^k} + \sum_{k=k_d}^{d-1} \frac{j^k}{k!} \sum_{s=0}^k \binom{k}{s} (-1)^{k-s} U_s^n. \quad (21)$$

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

$$\begin{cases} U_j^{n+1} = \frac{\lambda(\lambda-1)}{2} U_{j-2}^n + \lambda(2-\lambda) U_{j-1}^n + \frac{(\lambda-1)(\lambda-2)}{2} U_j^n, \\ U_{-1}^n = g(t^n) + \frac{\Delta x g'(t^n)}{a} + \frac{1}{2}(U_2^n - 2U_1^n + U_0^n), \\ U_{-2}^n = g(t^n) + \frac{2\Delta x g'(t^n)}{a} + 2(U_2^n - 2U_1^n + U_0^n), \\ U_j^0 = 0. \end{cases} \quad (22)$$

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 (22), 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 $\lambda = 1.8$.

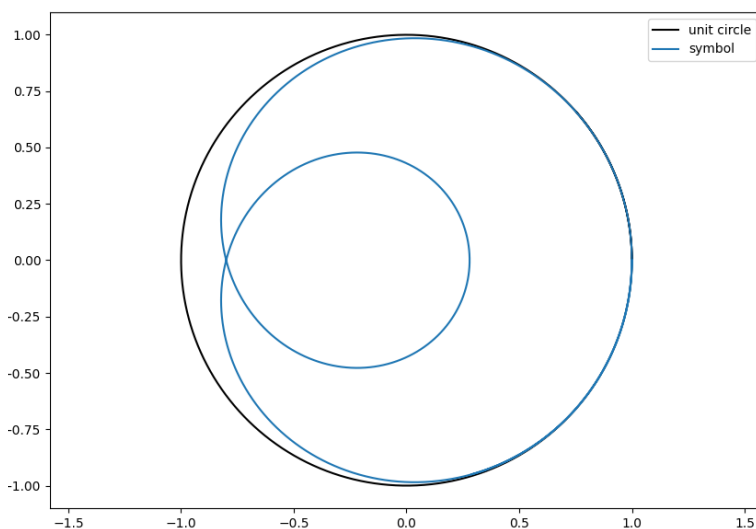


FIGURE 2. Symbol of Beam-Warming scheme for $\lambda = 1.8$.

Proposition 26. *The Beam-Warming scheme is Cauchy-stable if and only if $0 < \lambda \leq 2$.*

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

$$\begin{aligned} \gamma(\xi) &= \frac{(\lambda - 1)(\lambda - 2)}{2} + \lambda(2 - \lambda)e^{-i\xi} + \frac{\lambda(\lambda - 1)}{2}e^{-2i\xi} \\ &= e^{-i\xi} \left(\lambda(\lambda - 1) \cos \xi + \lambda(2 - \lambda) - (\lambda - 1)e^{i\xi} \right). \end{aligned}$$

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 $|\gamma(\xi)|^2 \leq 1$, so we want to have $\lambda(2 - \lambda)(\lambda - 1)^2(1 - \cos \xi)^2 \geq 0$. Because $\lambda > 0$, then the condition is $\lambda \leq 2$. \square

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.

4.4. Kreiss-Lopatinskii determinant computation for Beam-Warming scheme. First, we compute the Kreiss-Lopatinskii determinant Δ_{KL} from Definition 10 for the Beam-Warming scheme with S2ILW3 boundary condition as in (22). Assuming that the roots of (11) are distinct for a given $|z| \geq 1$, we have

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

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

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

In this rest of section, we continue the example of the Beam-Warming scheme (22) so as to illustrate practically the algebraic transformation set up in Lemma 19.

For that scheme, the corresponding \mathcal{Z} -transformed equation (10) is

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

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 $\alpha \stackrel{\text{def}}{=} \frac{-a_{-1}}{a_{-2}}$ and $\beta \stackrel{\text{def}}{=} \frac{z-a_0}{a_{-2}}$ so that the linear recurrence relation has now the form below:

$$\forall j \in \mathbb{N}, \quad \widetilde{U}_{j-2}(z) = \alpha\widetilde{U}_{j-1}(z) + \beta\widetilde{U}_j(z). \quad (23)$$

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 19, 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 $B\left(\widetilde{U}_{j-2}(z)\widetilde{U}_{j-1}(z)\widetilde{U}_j(z)\widetilde{U}_{j+1}(z)\widetilde{U}_{j+2}(z)\right)^T$ thanks to the recurrence relation (23). 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

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

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 the Corollary 14 in order to conclude on stability thanks to the Kreiss theorem 3. Actually, 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 Corollary 14, 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.

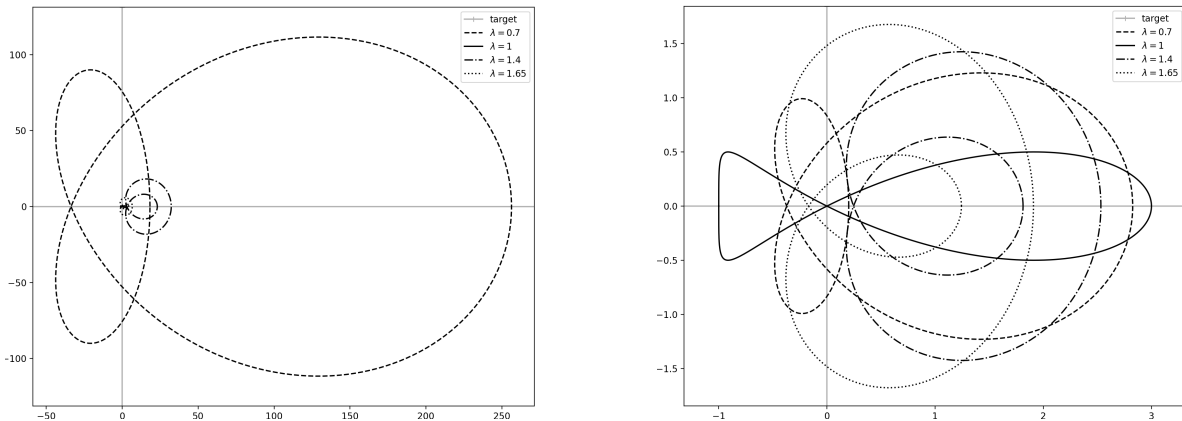


FIGURE 3. Kreiss-Lopatinskii Determinant Δ when z is on \mathbb{S} for scheme (22) for $\lambda \in \{0.7, 1, 1.4, 1.65\}$ (left) and the rescaled one $a_{-2}^2 \Delta$ (right).

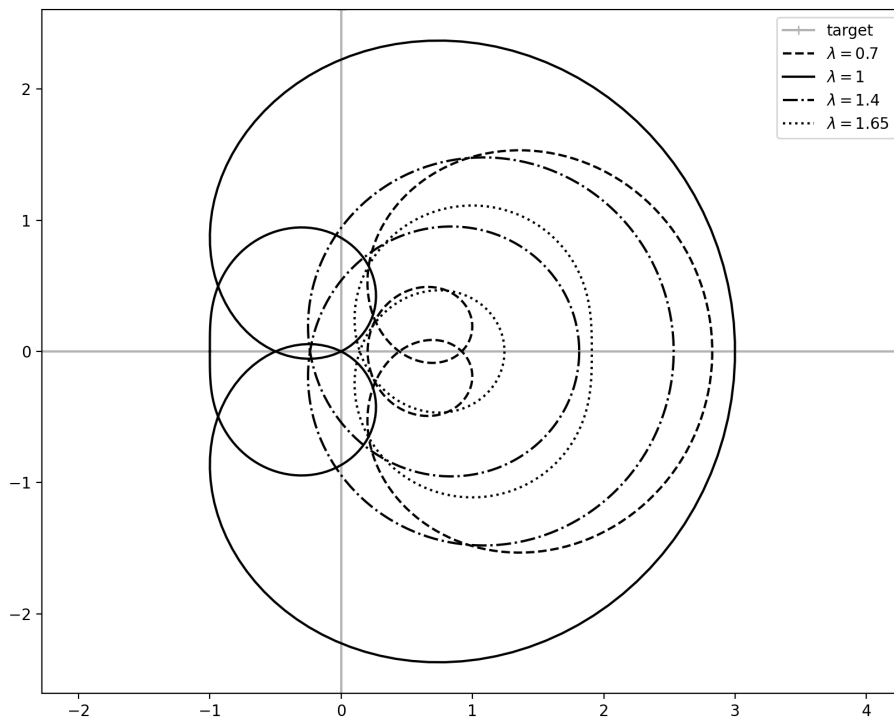


FIGURE 4. Rescaled Kreiss-Lopatinskii determinant $\frac{a_{-2}^2 \Delta}{z^2}$ for z in \mathbb{S} .

A particular situation occurs for $\lambda = 1$, due to the fact that $a_{-2} = 0$ 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 have to multiply by $\frac{1}{\beta} = \frac{1}{z^2}$ to find the Kreiss-Lopatinskii determinant (because $m = 3$ and $r = 1$).

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

4.5. **Numerical illustration.** Figure 4 may help to conclude on the stability of the scheme (22) as we said in Section 2.4, indeed, as we said in Section 4.1, one can compute the winding number while using a numerical procedure [31] and draw the winding number with respect to λ , as seen in Figure 5 for the case S2ILW3. It makes simpler the observation of the number of zeros of the Kreiss-Lopatinskii determinant. Hence, the numerical experiments indicate that the scheme (22) is strongly stable for $\lambda \in]0, 1[$ but also for $\lambda \in]1.52, 1, 78[$ approximately, but is unstable outside these domains.

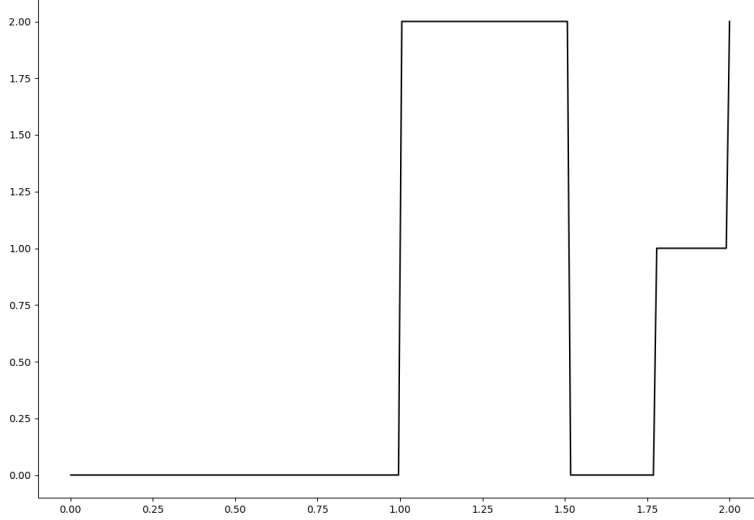


FIGURE 5. Number of zeros of Kreiss-Lopatinskii determinant with respect to λ for Beam-Warming scheme (22) with S2ILW3 boundary condition.

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 SILW boundary condition (defined by the equation (21)) as done in Figure 6. Surprisingly, in any of these experiments a small interval with a large CFL number λ provide the strong stability property.

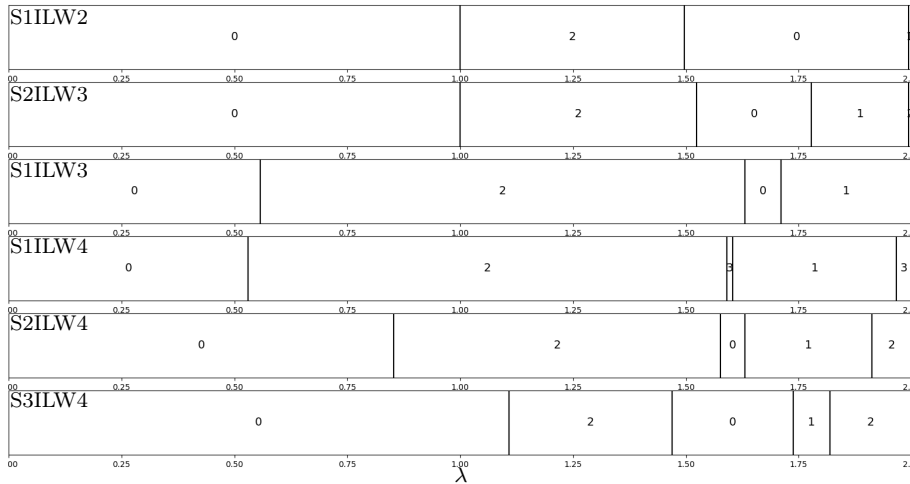


FIGURE 6. Number of zeros of Kreiss-Lopatinskii determinant for Beam-Warming scheme with different SILW boundary with respect to λ .

All the figures can be computed easily and quickly (less than one minute of computation on a classical 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.

4.6. Misalignment between boundaries and grid points. Motivated for example by the solving of multidimensional problems through cartesian grids, 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 can be gathered with the inverse Lax-Wendroff procedure in order to improve the accuracy at the boundary, see [9], [29] and [20]. As an archetype of 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 $[x_\sigma, 1]$:

$$\begin{cases} \partial_t u + a \partial_x u = 0, & t \geq 0, x \in [x_\sigma, 1], \\ u(t, x_\sigma) = g(t), & t \geq 0, \\ u(0, x) = f(x), & x \in [x_\sigma, 1]. \end{cases} \quad (24)$$

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 nonzero) $\sigma \in [-\frac{1}{2}, \frac{1}{2}[$. The scheme (4)-(5)-(6) is then implemented for $j \geq j_\sigma$ only, but with r ghost points at $j_\sigma - 1, \dots, j_\sigma - r$. 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 (21) presented above has to be slightly adapted (see [29]). The numerical boundary condition reads

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

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 (22) 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.

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

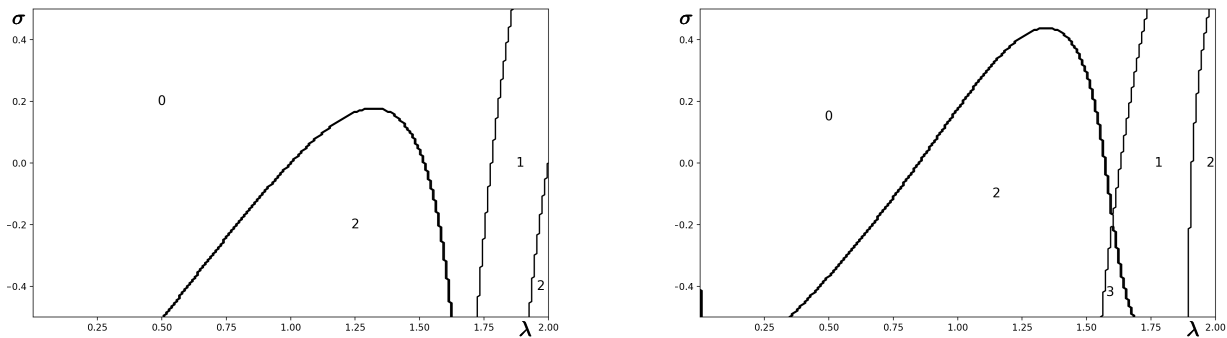


FIGURE 8. Stability of the Beam-Warming (22) with S2ILW3 boundary condition (left) and with S2ILW4 boundary condition (right).

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

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

In the simulations, the velocity is $a = 1$, the boundary condition is $g(t, y) = e^{-200(t-0.25)^2}$ and the initial data is $f \equiv 0$. The numerical solution is computed at time $T = 0.3$ thanks to 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.

5. CONCLUSION

The present theoretical and numerical results are restricted to the class of totally upwind schemes. This particular assumption enables a specific simple analysis of the Kreiss-Lopatinskii determinant, thanks to the explicit formula (20), and a numerical strategy to conclude to the existence of eigenvalue or generalized eigenvalue and thus to the stability issue. Actually, the present paper is only an initial effort in order to use and design some efficient numerical tools based on the Kreiss-Lopatinskii determinant and will lead hopefully to a more general study.

REFERENCES

- [1] A. Arnold, M. Ehrhardt, and I. Sofronov. Discrete transparent boundary conditions for the schrödinger equation: fast calculation, approximation, and stability. *Communications in Mathematical Sciences*, 1(3):501–556, 2003.
- [2] B. Barker and K. Zumbrun. Numerical proof of stability of viscous shock profiles. *Math. Models Methods Appl. Sci.*, 26(13):2451–2469, 2016.
- [3] R. M. Beam and R. F. Warming. The asymptotic spectra of banded toeplitz and quasi-toeplitz matrices. *SIAM Journal on Scientific Computing*, 14(4):971–1006, 1993.
- [4] N. Borovikh and M. N. Spijker. Resolvent conditions and bounds on the powers of matrices, with relevance to numerical stability of initial value problems. *J. Comput. Appl. Math.*, 125(1-2):41–56, 2000.
- [5] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. In *HCDTE lecture notes. Part I. Nonlinear hyperbolic PDEs, dispersive and transport equations*, volume 6 of *AIMS Ser. Appl. Math.*, page 146. Am. Inst. Math. Sci. (AIMS), Springfield, MO, 2013.
- [6] J.-F. Coulombel. Transparent numerical boundary conditions for evolution equations: derivation and stability analysis. *Ann. Fac. Sci. Toulouse, Math. (6)*, 28(2):259–327, 2019.

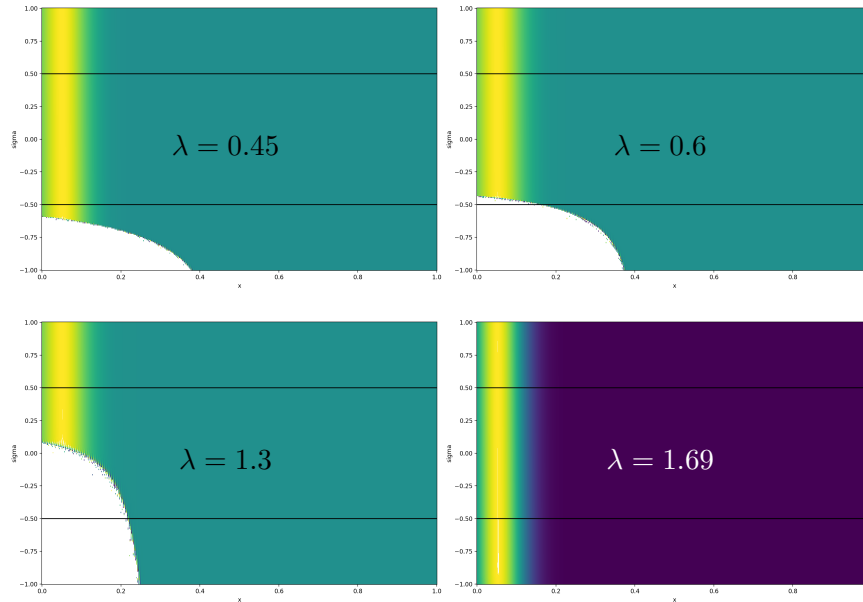


FIGURE 9. Numerical simulation of Beam-Warming scheme with S2ILW3 for CFL number $\lambda \in \{0.45, 0.6, 1.3, 1.69\}$

- [7] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen differenzgleichungen der mathematischen physik. *Mathematische Annalen*, 100(1):32–74, 1928.
- [8] J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. 43(1):50–67, 1947.
- [9] G. Dakin, B. Després, and S. Jaouen. Inverse Lax–Wendroff Boundary Treatment for Compressible Lagrange-Remap Hydrodynamics on Cartesian Grids. *Journal of Computational Physics*, 353:228–257, 2018.
- [10] M. Ehrhardt. Absorbing boundary conditions for hyperbolic systems. *Numer. Math., Theory Methods Appl.*, 3(3):295–337, 2010.
- [11] B. Engquist and A. Majda. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comput.*, 31:629–651, 1977.
- [12] C. Gasquet and P. Witomski. *Fourier analysis and applications: filtering, numerical computation, wavelets*, volume 30. Springer Science & Business Media, 2013.
- [13] S. K. Godunov and V. S. Ryabenkii. Spectral stability criteria for boundary-value problems for non-self-adjoint difference equations. *Russian Mathematical Surveys*, 18(3):1–12, 1963.
- [14] M. Goldberg. On a boundary extrapolation theorem by kreiss. *Mathematics of Computation*, 31(138):469–477, 1977.
- [15] B. Gustafsson. *High Order Difference Methods for Time Dependent PDE*. Number 38 in Springer series in computational mathematics. Springer, Berlin, 2008.
- [16] B. Gustafsson, H. Kreiss, and J. Olinger. *Time-Dependent Problems and Difference Methods*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.
- [17] B. Gustafsson, H.-O. Kreiss, and A. Sundström. Stability theory of difference approximations for mixed initial boundary value problems. II. *Mathematics of Computation*, 26(119):649–649, 1972.
- [18] R. Hersh. Mixed problems in several variables. *Journal of Mathematics and Mechanics*, 12(3):317–334, 1963.
- [19] P. D. Lax and R. D. Richtmyer. Survey of the stability of linear finite difference equations. *Communications on pure and applied mathematics*, 9(2):267–293, 1956.
- [20] T. Li, C.-W. Shu, and M. Zhang. Stability Analysis of the Inverse Lax–Wendroff Boundary Treatment for High Order Upwind-Biased Finite Difference Schemes. *Journal of Computational and Applied Mathematics*, 299:140–158, 2016.
- [21] T. Li, C.-W. Shu, and M. Zhang. Stability Analysis of the Inverse Lax–Wendroff Boundary Treatment for High Order Central Difference Schemes for Diffusion Equations. *Journal of Scientific Computing*, 70(2):576–607, 2017.
- [22] G. Métivier and K. Zumbrun. Symmetrizers and continuity of stable subspaces for parabolic-hyperbolic boundary value problems. *Discrete & Continuous Dynamical Systems - A*, 11(1):205–220, 2004.

- [23] M. N. Spijker, S. Tracogna, and B. D. Welfert. About the sharpness of the stability estimates in the kreiss matrix theorem. *Mathematics of Computation*, 72(242):697–714, 2002.
- [24] S. Tan and C.-W. Shu. Inverse lax-wendroff procedure for numerical boundary conditions of conservation laws. *Journal of Computational Physics*, 229(21):8144 – 8166, 2010.
- [25] M. Thuné. Automatic GKS stability analysis. *SIAM J. Sci. Statist. Comput.*, 7(3):959–977, 1986.
- [26] L. N. Trefethen. Group velocity interpretation of the stability theory of Gustafsson, Kreiss, and Sundström. *J. Comput. Phys.*, 49(2):199–217, 1983.
- [27] L. N. Trefethen. Instability of difference models for hyperbolic initial boundary value problems. *Communications on Pure and Applied Mathematics*, 37(3):329–367, 1984.
- [28] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, N.J, 2005.
- [29] F. Vilar and C.-W. Shu. Development and stability analysis of the inverse lax wendroff boundary treatment for central compact schemes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(1):39–67, 2015.
- [30] L. Wu. The semigroup stability of the difference approximations for initial-boundary value problems. *Mathematics of Computation*, 64(209):71–71, 1995.
- [31] J.-L. G. Zapata and J. C. D. Martín. A geometric algorithm for winding number computation with complexity analysis. *Journal of Complexity*, 28(3):320–345, 2012.
- [32] J. L. G. Zapata and J. C. D. Martín. A geometrical root finding method for polynomials, with complexity analysis, 2013.

UNIV RENNES, CNRS, IRMAR - UMR 6625,, F-35000 RENNES, FRANCE.

Email address: benjamin.boutin@univ-rennes1.fr

Email address: pierre.lebarbenchon@univ-rennes1.fr

Email address: nicolas.seguin@univ-rennes1.fr