Stability of finite difference schemes for hyperbolic initial boundary value problems

Jean-François Coulombel

STABILITY OF FINITE DIFFERENCE SCHEMES FOR HYPERBOLIC INITIAL BOUNDARY VALUE PROBLEMS

Jean-François Coulombel
CNRS, Université Lille 1 and Team Project SIMPAF of INRIA Lille - Nord Europe
Laboratoire Paul Painlevé (UMR CNRS 8524), Bâtiment M2, Cité Scientifique
59655 Villeneuve d'Ascq Cedex, France

Abstract. The aim of these notes is to present some results on the stability of finite difference approximations of hyperbolic initial boundary value problems. We first recall some basic notions of stability for the discretized Cauchy problem in one space dimension. Special attention is paid to situations where stability of the finite difference scheme is characterized by the so-called von Neumann condition. This leads us to the important class of geometrically regular operators. After discussing the discretized Cauchy problem, we turn to the case of initial boundary value problems. We introduce the notion of strongly stable schemes for zero initial data. The first main result characterizes strong stability in terms of a solvability property and an energy estimate for the resolvent equation. This first result shows that the so-called Uniform Kreiss-Lopatinskii Condition is a necessary condition for strong stability. The main result of these notes shows that the Uniform Kreiss-Lopatinskii Condition is also a sufficient condition for strong stability in the framework of geometrically regular operators. We illustrate our results on the Lax-Friedrichs and leap-frog schemes and check strong stability for various types of boundary conditions. We also extend a stability result by Goldberg and Tadmor for Dirichlet boundary conditions. In the last section of these notes, we show how to incorporate nonzero initial data and prove semigroup estimates for the discretized initial boundary value problems. We conclude with some remarks on possible improvements and open problems.

These notes have been prepared for a course taught by the author in Trieste during a trimester devoted to "Nonlinear Hyperbolic PDEs, Dispersive and Transport Equations" (SISSA, May-July 2011). The material in the notes covers three articles, one of which is a collaboration with A. Gloria (INRIA Lille, France). These notes are also the opportunity to include some simplified proofs of known results and to give some detailed examples, which may help in clarifying/demystifying the theory. The author warmly thanks the organizers as well as the participants of the trimester for inviting him to deliver these lectures and for the very kind and stimulating atmosphere in SISSA. Special thanks are addressed to Fabio Ancona, Stefano Bianchini, Gianluca Crippa and Andrea Marson for all the nice moments spent during the author's stay in Trieste.

1991 *Mathematics Subject Classification.* Primary: 65M12; Secondary: 65M06, 35L50.
Key words and phrases. Hyperbolic systems, boundary value problems, finite difference schemes, stability.
Research of the author was supported by the Agence Nationale de la Recherche, contract ANR-08-JCJC-0132-01.

1. Introduction

1.1. **What is and what is not inside these notes ?** These notes review the results derived in [4, 5, 6] on the stability of finite difference approximations for hyperbolic initial boundary value problems. In order to keep the length of the notes reasonable, the analogous results for hyperbolic partial differential equations, which have sometimes been proved quite some time ago, will be referred to without proof. This is mainly done to save space and to avoid introducing further notation. One crucial point in the analysis below is to understand why the techniques developed for partial differential equations are unfortunately not sufficient to handle finite difference schemes. Special attention is therefore paid to the main new phenomena that appear when considering discretized equations. Some examples are scattered throughout the text in order to explain how the general theory, which may look sometimes rather complicated, is often simplified when one faces a specific example. In particular, the Lax-Friedrichs and leap-frog schemes, which are some of the most simple discretizations of a hyperbolic equation, serve as a guideline throughout Sections 2, 4 and 5.

The notes are essentially self-contained. All results but one are completely proved. Of course, some familiarity with hyperbolic equations can do no harm, but the only basic requirements to follow the proofs are a good knowledge of matrices, some tools from real and complex analysis and a little bit of functional analysis.

As far as hyperbolic boundary value problems are concerned, the reader might first want to get familiar with the theory for partial differential equations before reading the discrete counterpart that is detailed here. In this case, the books [3, chapter 7] or [2, chapters 3-5] are convenient references. However, the theory for finite difference schemes can also be seen as a first step towards the theory for partial differential equations since, as detailed below, some parts of the analysis are actually simpler in the discrete case. Even though the original results were not proved historically in this way, discrete problems can also be a constructive approximation method to obtain solutions of partial differential equations. To be completely honest, the author is not convinced that this would be the most direct way to construct solutions of hyperbolic initial boundary value problems.

As far as numerical approximations are concerned, a convenient reference for our purpose is [9, chapters 5, 6, 11 and 13] where stability issues are analyzed, in particular for the discrete Cauchy problem. The techniques developed below are restricted to linear schemes for linear equations. Consequently, no knowledge of flux limiters, ENO/WENO schemes nor any other nonlinear high order approximation procedure is assumed. Extending some of the results below to such numerical schemes is definitely an open and challenging issue (which would be very interesting from the point of view of applications).

1.2. **Some notation.** Throughout these notes, the following notation is used:

$$\mathscr{U} := \{\zeta \in \mathbb{C}, |\zeta| > 1\}\,, \quad \overline{\mathscr{U}} := \{\zeta \in \mathbb{C}, |\zeta| \geq 1\}\,,$$
$$\mathbb{D} := \{\zeta \in \mathbb{C}, |\zeta| < 1\}\,, \quad \mathbb{S}^1 := \{\zeta \in \mathbb{C}, |\zeta| = 1\}\,.$$

We let $\mathscr{M}_{d,D}(\mathbb{K})$ denote the set of $d \times D$ matrices with entries in $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, and we use the notation $\mathscr{M}_D(\mathbb{K})$ when $d = D$. If $M \in \mathscr{M}_D(\mathbb{C})$, $\mathrm{sp}(M)$ denotes the spectrum of M, $\rho(M)$ denotes the spectral radius of M, while M^* denotes the conjugate transpose of M. The notation M^T is also used for the transpose of a matrix M (here M is not necessarily a square matrix). The matrix $(M + M^*)/2$ is called the real part of $M \in \mathscr{M}_D(\mathbb{C})$ and is denoted $\mathrm{Re}\,(M)$. The real vector space of Hermitian matrices of size D is denoted $\mathscr{H}_D$.

For $H_1, H_2 \in \mathscr{H}_D$, we write $H_1 \geq H_2$ if for all $x \in \mathbb{C}^D$ we have $x^*\,(H_1 - H_2)\,x \geq 0$. We let I denote the identity matrix, without mentioning the dimension. The norm of a vector $x \in \mathbb{C}^D$ is $|x| := (x^*\,x)^{1/2}$. The corresponding norm for matrices in $\mathscr{M}_D(\mathbb{C})$ is also denoted $|\cdot|$. We let ℓ^2 denote the set of square integrable sequences, and we usually do not mention the set of indeces of the sequences (sequences may be valued in $\mathbb{C}^d$ for some integer d).

The notation $\mathrm{diag}\,(M_1, \dots, M_p)$ is used to denote the diagonal matrix whose entries are (in this order) $M_1, \dots, M_p$. If the M_j's are matrices themselves, then the same notation is used to denote the corresponding block diagonal matrix.

The notation $x_1 = x_2 \leq x_3 = x_4$ means that x_1 equals x_2, x_3 equals x_4, and x_2 is not larger than x_3 (and consequently, of course, x_1 is not larger than x_4).

The letter C denotes a constant that may vary from line to line or even within the same line. The dependence of the constants on the various parameters is made precise throughout the text.

1.3. **General presentation of the stability problem.** In one space dimension, a hyperbolic initial boundary value problem reads

$$\begin{cases} \partial_t u + A\,\partial_x u = F(t,x)\,, & (t,x)\in\mathbb{R}^+\times\mathbb{R}^+\,,\\ B\,u(t,0) = g(t)\,, & t\in\mathbb{R}^+\,,\\ u(0,x) = f(x)\,, & x\in\mathbb{R}^+\,, \end{cases} \tag{1}$$

where $A\in\mathscr{M}_N(\mathbb{R})$ is diagonalizable with real eigenvalues, the unknown $u(t,x)$ is valued in $\mathbb{R}^N$, and B is a matrix - not necessarily a square matrix, see below - that encodes the boundary conditions. The functions F, g, f are given source terms, respectively, the interior source term, the boundary source term and the initial data. Of course, in one space dimension, it is rather easy to solve such a linear problem by diagonalizing A and integrating along the characteristics. More precisely, let $r_1,\dots,r_N$ denote a basis of eigenvectors of A associated with eigenvalues $\lambda_1,\dots,\lambda_N$. Let us assume for simplicity that 0 does not belong to $\mathrm{sp}(A)$, the so-called *non-characteristic* case. Up to reordering the eigenvalues, we can label them so that

$$\lambda_1,\dots,\lambda_p>0\,,\quad \lambda_{p+1},\dots,\lambda_N<0\,.$$

Then we decompose the source terms F, f and the unknown u as

$$F(t,x)=\sum_{i=1}^N F_i(t,x)\,r_i\,,\quad f(x)=\sum_{i=1}^N f_i(x)\,r_i\,,\quad u(t,x)=\sum_{i=1}^N u_i(t,x)\,r_i\,.$$

Assuming for simplicity that the solution u is smooth, at least $\mathscr{C}^1$ with respect to (t,x), (1) gives

$$\forall\, i=1,\dots,N\,,\quad \frac{\mathrm{d}}{\mathrm{d}t}\big[u_i(t,x+\lambda_i\,t)\big]=F_i(t,x+\lambda_i\,t)\,.$$

We integrate these equalities with respect to t, keeping in mind that $u_1,\dots,u_N,F_1,\dots,F_N$ are only defined on $\mathbb{R}^+\times\mathbb{R}^+$. For $i\in\{p+1,\dots,N\}$, that is when λ_i is negative, we obtain the formula

$$u_i(t,x)=f_i(x-\lambda_i\,t)+\int_0^t F_i(s,x-\lambda_i\,(t-s))\,\mathrm{d}s\,. \tag{2}$$

The latter formula makes sense for all (t,x) in the quarter-space $\mathbb{R}^+\times\mathbb{R}^+$ because in that case, all quantities $x-\lambda_i\,t$ and $x-\lambda_i\,(t-s)$ in (2) are nonnegative. In particular, the trace of u_i on the boundary $\{x=0\}$ of the space domain is entirely determined by the data:

$$u_i(t,0)=f_i(|\lambda_i|\,t)+\int_0^t F_i(s,|\lambda_i|\,(t-s))\,\mathrm{d}s\,.$$

One should be careful when performing the integration in the case $i\in\{1,\dots,p\}$. According to the sign of $x-\lambda_i\,t$, we obtain

$$u_i(t,x)=\begin{cases} f_i(x-\lambda_i\,t)+\int_0^t F_i(s,x-\lambda_i\,(t-s))\,\mathrm{d}s\,, & \text{if } x\geq\lambda_i\,t\,,\\ u_i(t-x/\lambda_i,0)+\int_{t-x/\lambda_i}^t F_i(s,x-\lambda_i\,(t-s))\,\mathrm{d}s\,, & \text{if } x\leq\lambda_i\,t\,. \end{cases} \tag{3}$$

Analyzing the formulas (2) and (3), we observe that the solution u is entirely determined provided that we can express the traces of the incoming characteristics $\{u_i(t,0)\,,\ 1\leq i\leq p\}$ in terms of the data F, g, f. Since the traces of the outgoing characteristics $\{u_i(t,0)\,,\ p+1\leq i\leq N\}$ are already determined by the formula (2), the boundary condition in (1) reads

$$\sum_{i=1}^p u_i(t,0)\,B\,r_i=g(t)-\sum_{i=p+1}^N u_i(t,0)\,B\,r_i\,,$$

where the right-hand side can be expressed in terms of F, g, f. Therefore the initial boundary value problem (1) can be well-posed in any reasonable sense (meaning at least existence and uniqueness of

a solution, even though we do not make the functional framework precise) if and only if the matrix B belongs to $\mathscr{M}_{p,N}(\mathbb{R})$, and satisfies

$$\mathbb{R}^p = \mathrm{Span}\big(B\,r_1, \dots, B\,r_p\big)\,. \tag{4}$$

In particular, (4) implies that B should have maximal rank, but this could have already been seen from (1) for otherwise there would have been an algebraic obstruction to solving the boundary condition in (1). If the matrix B satisfies (4), then we get an explicit expression for the components of the solution u along the eigenvectors r_i. Energy estimates of u in terms of F, g, f as well as qualitative properties of the solution (regularity, finite speed of propagation etc.) are readily seen from these expressions. If we try to summarize the above discussion, we obtain the following conclusion: well-posedness of (1) requires first a precise number of boundary conditions that is compatible with the hyperbolic operator, and the verification of the *algebraic condition* (4). Consequently, rather than checking energy estimates for each possible boundary conditions in (1), we are just reduced to verifying (4) which is by far easier.

A remarkable result by Kreiss [13] states that for the analogue of (1) in several space dimensions, well-posedness - that is existence, uniqueness and continuous dependence of a solution in a suitable functional framework - can still be characterized by an algebraic condition. The latter is usually referred to as the Uniform Kreiss-Lopatinskii Condition (UKLC in what follows). There is however a modification between the one-dimensional case and the multi-dimensional case. Observing that in one space dimension, the condition (4) for well-posedness equivalently reads

$$\mathrm{Ker}\ B \cap \mathrm{Span}\big(r_1, \dots, r_p\big) = \{0\}\,,$$

the UKLC in several space dimensions reads

$$\forall\, \zeta \in \Sigma\,, \quad \mathrm{Ker}\ B \cap \mathbb{E}(\zeta) = \{0\}\,,$$

where Σ is some *infinite* set of parameters and the vector spaces $\mathbb{E}(\zeta)$ all have dimension p. Verifying the UKLC in several space dimensions is therefore more complicated since it requires computing a basis of a vector space that depends on parameters, and then checking that an appropriate determinant does not vanish. This can sometimes be done with explicit computations, see for instance [2, chapter 14] for the case of gas dynamics, or it can also be done in a numerical way (this numerical strategy was used in other contexts such as the computation of Evans functions). One of the most difficult steps in the theory of [13] is to give a precise definition of the vector spaces $\mathbb{E}(\zeta)$ that enter the definition of the UKLC. Not so surprisingly, we shall also face this difficulty when dealing with numerical schemes. However, as shown on some specific examples, the general theory can be far more complicated than what one faces with one particular numerical scheme. One should therefore not be afraid to try checking the UKLC on some examples: it is the best way to manipulate the objects, to get used to them and to understand better the general theory. The reader is therefore strongly encouraged to test all the results below on his/her favourite numerical scheme.

Our main goal in these notes is to characterize - that is, find necessary and sufficient conditions - stability for the numerical schemes occurring after discretizing the initial boundary value problem (1). Existence and uniqueness for the discretized version of (1) will be completely trivial in these notes, and stability should be understood as continuous dependence of the solution with respect to the data, meaning the last requirement for "Hadamard well-posedness". In view of the existing theory for (1) and its analogue in several space dimensions, we wish to obtain a general result of the form: "*the discretization of* (1) *is stable if and only if an algebraic condition (to be determined) is satisfied*". This result will be meaningful if testing the algebraic condition is easier than checking the validity of energy estimates for the numerical schemes. As usual when one deals with problems in infinite dimensional spaces, the choice of the norm in the stability definition is crucial. Our long term goal is to develop an analogous theory for discretized multi-dimensional problems to the one detailed here in the one-dimensional case. The functional framework should therefore be compatible with such an extension, and this basically restricts us to working with L^2-type spaces (hence the use of many Hilbertian methods). As far as convergence of numerical schemes is concerned, we focus here on the stability problem since consistency is supposed to be an easier problem. In some sense, consistency of a numerical scheme follows from some Taylor expansions on an exact smooth

solution of the continuous problem (1). If we can derive a powerful stability theory, convergence should follow as a more or less direct consequence by combining stability with consistency. Instead of giving precise results in this direction, we shall refer the interested reader to [8] where this strategy is used.

Let us now detail the plan of these notes. As a warm-up, we begin in Section 2 with some considerations on the discretized Cauchy problem. This will be the opportunity to introduce some objects that are crucial in the analysis of the discretized initial boundary value problem. We also introduce and analyze some examples such as the Lax-Friedrichs and leap-frog schemes. Sections 3 and 4 are devoted to the analysis of the discretized initial boundary value problem with zero initial data. This is, technically speaking, the most difficult part of these notes. In the case of zero initial data, stability can be analyzed by applying the Laplace transform and the so-called normal modes analysis. Our main result characterizes stability by means of an algebraic condition of the same type as the UKLC. The main results in Section 3 generalize - and sometimes simplify - the fundamental contribution by Gustafsson, Kreiss and Sundström [10]. To clarify the theory, we explain in Section 4 the behaviors of all the objects (stable eigenvalues, stable subspace, UKLC...) for the Lax-Friedrichs and leap-frog schemes. Section 5 deals with the problem of incorporating nonzero initial data and adapting the notion of stability to this new framework. For one-dimensional problems, the incorporation of initial data was performed by Wu [26]. We shall explain his method and propose an alternative - though closely related - approach. The main advantage of this new approach is the fact that it can be adapted in a straightforward way to multi-dimensional problems, while Wu's method is restricted to one-dimensional problems for reasons that we shall detail. Eventually, we shall present some (of the numerous) open problems in Section 6.

2. Fully discretized hyperbolic equations

2.1. **Finite difference operators and stability for the discrete Cauchy problem.** We consider the Cauchy problem

$$\begin{cases} \partial_t u + A\,\partial_x u = 0\,, & (t,x) \in \mathbb{R}^+ \times \mathbb{R}\,, \\ u(0,x) = f(x)\,, & x \in \mathbb{R}\,, \end{cases} \tag{5}$$

on the whole real line. As in Section 1, $A \in \mathscr{M}_N(\mathbb{R})$ is diagonalizable with real eigenvalues $\lambda_1, \dots, \lambda_N$. For initial data $f \in L^2(\mathbb{R})$, there exists a unique solution $u \in \mathscr{C}(\mathbb{R}^+; L^2(\mathbb{R}))$ solution to (5). This solution can be explicitly computed by integrating along the characteristics. Decomposing along the eigenvectors r_i of A, we obtain

$$u(t,x) = \sum_{i=1}^N f_i(x - \lambda_i\, t)\, r_i\,, \quad f(x) = \sum_{i=1}^N f_i(x)\, r_i\,.$$

In particular, the following energy estimate is straightforward

$$\sup_{t\geq 0} \int_{\mathbb{R}} |u(t,x)|^2\, \mathrm{d}x \leq C \int_{\mathbb{R}} |f(x)|^2\, \mathrm{d}x\,, \tag{6}$$

with a numerical constant C that only depends on A. Another possibility for computing the solution u to (5) is to use Fourier transform with respect to the space variable x. Letting ξ denote the associated frequency variable, $\widehat{u}(t,\xi)$ satisfies the linear ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\widehat{u}(t,\xi) = -i\,\xi\,A\,\widehat{u}(t,\xi)\,, \quad \widehat{u}(0,\xi) = \widehat{f}(\xi)\,,$$

which we solve to obtain

$$\widehat{u}(t,\xi) = \exp(-i\,t\,\xi\,A)\,\widehat{f}(\xi)\,. \tag{7}$$

Let us now introduce the discretizations of (5) that we consider in these notes. Let $\Delta x, \Delta t > 0$ denote a space and a time step where the ratio $\lambda := \Delta t/\Delta x$ is a fixed positive constant. In all what follows, λ is called the CFL (for Courant-Friedrichs-Lewy) number and $\Delta t \in]0,1]$ plays the role of a small parameter, while $\Delta x = \Delta t/\lambda$ varies accordingly. Some of the assumptions in the theory are restrictions on λ. Typically, the results will hold provided that λ is chosen in a suitable interval of $\mathbb{R}^+$.

The solution to (5) is approximated by a sequence (U_j^n) defined for $n \in \mathbb{N}$ and $j \in \mathbb{Z}$. More precisely, we always identify the sequence (U_j^n) defined for $n \in \mathbb{N}$ and $j \in \mathbb{Z}$ with the step function

$$U(t,x) := U_j^n \quad \text{for } (t,x) \in [n\,\Delta t, (n+1)\,\Delta t[\times [j\,\Delta x, (j+1)\,\Delta x[\,.$$

The goal is to build a numerical scheme that produces a step function U that is close to u for the $L^\infty(\mathbb{R}^+; L^2(\mathbb{R}))$ topology. This is a natural requirement in view of (6). (The choice of the topology may look rather arbitrary, especially in one space dimension, but as detailed in the introduction, our goal is to develop some tools that may be extended to multi-dimensional problems.) Let us observe that though the solution u to (5) lies in the space $\mathscr{C}(\mathbb{R}^+; L^2(\mathbb{R}))$, the approximation U lies, in general, in the larger space $L^\infty(\mathbb{R}^+; L^2(\mathbb{R}))$. It is only in the limit process, by letting Δt tend to zero, that continuity with respect to time can be recovered.

Discretizing the initial condition of (5) is usually performed by choosing

$$\forall\, j \in \mathbb{Z}\,, \quad f_j := \frac{1}{\Delta x} \int_{j\,\Delta x}^{(j+1)\,\Delta x} f(x)\, \mathrm{d}x\,.$$

This is not the only possible choice, but it has the good property of being stable with respect to the L^2 topology, that is[1]

$$\sum_{j\in\mathbb{Z}} \Delta x\, |f_j|^2 \leq \int_{\mathbb{R}} |f(x)|^2\, \mathrm{d}x\,.$$

From now on, we assume that the initial discretization has been chosen, producing a sequence $(f_j) \in \ell^2$ such that the associated grid function is "close" - in some sense that we do not make precise - to the initial condition f of (5).

[1] This estimate is easily proved by applying Cauchy-Schwarz inequality on each interval $[j\,\Delta x, (j+1)\,\Delta x[$.

Starting from a given sequence $(f_j) \in \ell^2$, for instance the sequence defined just above, many classical finite difference approximations of (5) take the form

$$\begin{cases} U_j^{n+1} = Q\, U_j^n\,, & j \in \mathbb{Z}\,, \quad n \geq 0\,, \\ U_j^0 = f_j\,, & j \in \mathbb{Z}\,, \end{cases} \tag{8}$$

where Q is a finite difference operator whose expression is given by

$$Q := \sum_{\ell=-r}^{p} A_\ell\, \mathbf{T}^\ell\,, \quad (\mathbf{T}^\ell\, V)_k := V_{k+\ell}\,. \tag{9}$$

Let us give a few explanations on (9). The shift operator $\mathbf{T}$ is an invertible operator on $\ell^2(\mathbb{Z})$ so taking powers $\mathbf{T}^\ell$ is legitimate. The integers p, r in (9) are fixed, that is, they do not depend on the index j on the grid where the numerical scheme is applied, and neither do they depend on the small parameter Δt. In the same way, the matrices $A_{-r}, \dots, A_p \in \mathscr{M}_N(\mathbb{R})$ should not depend on Δt, nor on the initial data (f_j). In most (linear) finite difference schemes, the matrices A_ℓ are polynomial functions of the matrix $\lambda\, A$. In that case, all matrices A_ℓ can be diagonalized in the same basis. We refer to the following paragraphs for some examples. Summarizing, the numerical scheme (8) is defined by two integers p, r and by the matrices $A_{-r}, \dots, A_p$. Then the sequence U^{n+1} is computed from U^n by applying the operator Q defined in (9), which acts boundedly on ℓ^2. In particular, for all initial condition $(f_j) \in \ell^2$, there exists a unique sequence (U_j^n) that is a solution to (8), and moreover this solution satisfies $(U_j^n)_{j\in\mathbb{Z}} \in \ell^2$ for all $n \in \mathbb{N}$.

Let us briefly recall that for nonlinear schemes such as ENO or WENO schemes, the matrices A_ℓ are not fixed but depend on the solution that is computed ; for instance, to compute the sequence (U_j^1), the matrices A_ℓ at the first time step depend on (f_j), and they are updated at each time step in order to take the oscillations of the sequence (U_j^n) into account. The theory developed below relies crucially on the fact that the matrices A_ℓ are independent of the sequence (U_j^n). It therefore does not extend directly to such nonlinear schemes.

The definition of stability for the numerical scheme (8) requires that the solution to (8) satisfies the discrete analogue of (6). More precisely, we introduce

Definition 2.1 (Stability for the discrete Cauchy problem)**.** *The numerical scheme defined by (8), (9) is (ℓ^2-) stable if there exists a constant $C_0 > 0$ such that for all $\Delta t \in]0, 1]$, for all initial condition $(f_j)_{j\in\mathbb{Z}} \in \ell^2$ and for all $n \in \mathbb{N}$, there holds*

$$\sum_{j\in\mathbb{Z}} \Delta x\, |U_j^n|^2 \leq C_0 \sum_{j\in\mathbb{Z}} \Delta x\, |f_j|^2\,.$$

Of course, we could simplify the factor Δx on both sides of the stability estimate and Definition 2.1 is clearly independent of the small parameter Δt, but we prefer to keep the Δx factor in order to highlight the fact that discrete ℓ^2 norms are nothing but L^2 norms for step functions defined on the grid with uniform space step Δx. The factor Δx corresponds to the measure of the cell $[j\,\Delta x, (j+1)\,\Delta x[$. This observation is useful in order to understand the similarities between stability estimates for numerical schemes and energy estimates for partial differential equations.

Stability for the numerical scheme (8) is characterized by the following result.

Proposition 2.1 (Characterization of stability for the discrete Cauchy problem)**.** *The scheme (8) is stable in the sense of Definition 2.1 if and only if the matrices A_ℓ in (9) satisfy*

$$\forall\, n \in \mathbb{N}\,, \quad \forall\, \eta \in \mathbb{R}\,, \quad \left| \left(\sum_{\ell=-r}^{p} \mathrm{e}^{i\,\ell\,\eta}\, A_\ell \right)^n \right|^2 \leq C_0\,, \tag{10}$$

with the same constant C_0 as in Definition 2.1.

For future use, it is convenient to introduce the notation

$$\forall\, \kappa \in \mathbb{C} \setminus \{0\}\,, \quad \mathscr{A}(\kappa) := \sum_{j=-r}^{p} \kappa^\ell\, A_\ell\,, \tag{11}$$

so that (10) reads

$$\forall\, n \in \mathbb{N}\,, \quad \forall\, \eta \in \mathbb{R}\,, \quad \left|\mathscr{A}(\mathrm{e}^{i\,\eta})^n\right|^2 \leq C_0\,.$$

The matrix $\mathscr{A}(\mathrm{e}^{i\,\eta})$ is called the *amplification matrix* (or *symbol*) of the scheme (8).

Proof of Proposition 2.1. • Let us assume that the bound (10) holds, or in other words that the family $\{\mathscr{A}(\mathrm{e}^{i\,\eta}), \eta \in \mathbb{R}\}$ is uniformly power bounded with the bound $\sqrt{C_0}$. Let us consider the scheme (8). Then for all $n \in \mathbb{N}$, the step function U^n defined by

$$U^n(x) := U_j^n\,, \quad \text{for } x \in [j\,\Delta x, (j+1)\,\Delta x[\,,$$

satisfies

$$\forall\, x \in \mathbb{R}\,, \quad U^{n+1}(x) = \sum_{\ell=-r}^{p} A_\ell\, U^n(x + \ell\,\Delta x)\,.$$

We already know that U^n belongs to $L^2(\mathbb{R})$ for all n, so we can apply Fourier transform on both sides of the latter equality[2]. This operation yields the relation

$$\forall\, \xi \in \mathbb{R}\,, \quad \widehat{U^{n+1}}(\xi) = \mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi})\,\widehat{U^n}(\xi)\,,$$

from which we deduce

$$\forall\, \xi \in \mathbb{R}\,, \quad \widehat{U^n}(\xi) = \mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi})^n\,\widehat{U^0}(\xi)\,.$$

Using Plancherel Theorem and the bound (10), we obtain

$$\int_{\mathbb{R}} |U^n(x)|^2\,\mathrm{d}x = \frac{1}{2\,\pi}\int_{\mathbb{R}} \left|\widehat{U^n}(\xi)\right|^2 \mathrm{d}\xi \leq \frac{C_0}{2\,\pi}\int_{\mathbb{R}} \left|\widehat{U^0}(\xi)\right|^2 \mathrm{d}\xi = C_0\int_{\mathbb{R}} |U^0(x)|^2\,\mathrm{d}x\,.$$

Consequently, the scheme (8) is stable with the same constant C_0 as in (10).

• We now assume that the scheme (8) is stable with the constant C_0, and we fix an integer n as well as a real number η. Let also $X \in \mathbb{C}^N$ have norm 1. Then for an integer $k \geq n\,\max(p,r)$, we consider the initial condition

$$f_j := \begin{cases} \mathrm{e}^{i\,j\,\eta}\,X\,, & \text{if } |j| \leq k\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

The following computation is elementary (just recall the notation (11))

$$U_j^1 = \mathscr{A}(\mathrm{e}^{i\,\eta})\,f_j\,, \quad \text{if } |j| \leq k - \max(p,r)\,.$$

By a straightforward induction, we obtain

$$U_j^n = \mathscr{A}(\mathrm{e}^{i\,\eta})^n\,f_j\,, \quad \text{if } |j| \leq k - n\,\max(p,r)\,. \tag{12}$$

Then we have

$$\sum_{|j|\leq k-n\,\max(p,r)} \Delta x\,|U_j^n|^2 \leq \sum_{j\in\mathbb{Z}} \Delta x\,|U_j^n|^2 \leq C_0\sum_{j\in\mathbb{Z}} \Delta x\,|f_j|^2 = C_0\,\Delta x\,(2\,k+1)\,.$$

The left hand side of the latter inequality is computed by using (12) and by using the definition of the vector f_j. We obtain

$$(2\,k+1-2\,n\,\max(p,r))\,\Delta x\,\left|\mathscr{A}(\mathrm{e}^{i\,\eta})^n\,X\right|^2 \leq C_0\,\Delta x\,(2\,k+1)\,.$$

Dividing by $\Delta x\,(2\,k+1)$, letting k tend to infinity and taking the supremum with respect to X, we obtain the result of Proposition 2.1. □

Remark 2.1. *The easiest case of stability is when the matrices A_ℓ satisfy*

$$\forall\, \eta \in \mathbb{R}\,, \quad \left|\mathscr{A}(\mathrm{e}^{i\,\eta})\right| \leq 1\,.$$

Then the solution to (8) is such that the sequence of norms $(\|U^n\|_{\ell^2})$ is non-increasing. This more restrictive notion is called strong ℓ^2-stability in [24].

[2]This is the precise point where it is crucial to deal with constant matrices A_ℓ.

The main idea in the proof of Proposition 2.1 is to test the stability estimate on oscillations $\mathrm{e}^{i\,j\,\eta}$. Of course, the sequence $(\mathrm{e}^{i\,j\,\eta})_{j\in\mathbb{Z}}$ does not belong to ℓ^2 so we need to make a truncation. Fourier's inversion Theorem shows that functions can be decomposed as a superposition of oscillations so stability of the numerical scheme is encoded in a stability estimate for pure oscillations that should be uniform with respect to the frequency.

Let us now make an important remark. The grid function U^n is supposed to be an approximation of the solution u at time $n\,\Delta t$. Hence $\widehat{U^n}$ should approximate $\widehat{u}(n\,\Delta t,\cdot)$. Recalling the relation (7), the matrix $\mathscr{A}$ should satisfy

$$\mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi})^n \approx \exp(-i\,n\,\Delta t\,\xi\,A) = \exp(-i\,\Delta t\,\xi\,A)^n\,.$$

We do not wish to make the meaning of the symbol $\approx$ precise. However, a natural requirement should be to impose that in the limit $\Delta t\to 0$ with $n=1$, both expressions coincide. This yields the restriction

$$\mathscr{A}(1) = \sum_{\ell=-r}^{p} A_\ell = I\,. \tag{13}$$

A numerical scheme of the form (8), (9) that satisfies (13) is said to be *consistent*. Higher order accuracy of the numerical scheme is encoded in the Taylor expansion of $\mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi})$ as Δt tends to 0. However, this notion will not be much used in what follows, except when discussing some examples.

We shall go back to the result of Proposition 2.1 in the following paragraph. Before doing so, let us discuss a possible extension of the theory. The reader who is familiar with numerical discretizations of ordinary differential equations will probably wonder why we have restricted to numerical schemes with only one time step. As a matter of fact, there is no reason for doing so and in some situations one could prefer using a two steps (or more) numerical procedure. A well-known example is the leap-frog scheme. Another example is discussed in one of the following paragraphs. Numerical schemes with several time steps take the following form: let us consider three integers p, r, s. Starting from some sequences $(f_j^0),\dots,(f_j^s)$ in ℓ^2, the sequence (U_j^n) is defined by

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} Q_\sigma\,U_j^{n-\sigma}\,, & j\in\mathbb{Z}\,, \quad n\geq s\,,\\ U_j^n = f_j^n\,, & j\in\mathbb{Z}\,, \quad n=0,\dots,s\,, \end{cases} \tag{14}$$

where the shift operators Q_σ are given by

$$Q_\sigma := \sum_{\ell=-r}^{p} A_{\ell,\sigma}\,\mathbf{T}^\ell\,. \tag{15}$$

Again, the matrices $A_{\ell,\sigma}$ in (15) should not depend on the sequence to be computed so that the same scheme applies to all initial data and at each time step. The notion of stability for (14) is entirely analogous to Definition 2.1.

Definition 2.2 (Stability for the discrete Cauchy problem)**.** *The numerical scheme defined by* (14), (15) *is* (ℓ^2-) *stable if there exists a constant $C_0>0$ such that for all $\Delta t\in]0,1]$, for all initial condition $(f_j^0)_{j\in\mathbb{Z}},\dots,(f_j^s)_{j\in\mathbb{Z}}\in\ell^2$ and for all $n\in\mathbb{N}$, there holds*

$$\sum_{j\in\mathbb{Z}}\Delta x\,|U_j^n|^2 \leq C_0\left(\sum_{j\in\mathbb{Z}}\Delta x\,|f_j^0|^2+\dots+\sum_{j\in\mathbb{Z}}\Delta x\,|f_j^s|^2\right).$$

Similarly to Proposition 2.1, Proposition 2.2 below characterizes stability of the scheme (14) in terms of the uniform power boundedness of the corresponding amplification matrix. For future use, we therefore introduce the notation

$$\forall\,\kappa\in\mathbb{C}\setminus\{0\}\,,\quad \mathscr{A}(\kappa) := \begin{pmatrix} \widehat{Q_0}(\kappa) & \dots & \dots & \widehat{Q_s}(\kappa)\\ I & 0 & \dots & 0\\ & \ddots & \ddots & \vdots\\ 0 & 0 & I & 0 \end{pmatrix} \in \mathscr{M}_{N(s+1)}(\mathbb{C})\,,\quad \widehat{Q_\sigma}(\kappa) := \sum_{\ell=-r}^{p}\kappa^\ell\,A_{\ell,\sigma}\,, \tag{16}$$

which coincides with our former notation (11) in the case $s = 0$ (one step scheme). To avoid any possible confusion, we emphasize that in (16), the matrix $\mathscr{A}(\kappa)$ is decomposed into blocks, each of which is a square $N \times N$ matrix with complex coefficients. Such block decompositions of matrices will occur at numerous places in these notes.

Proposition 2.2 (Characterization of stability for the discrete Cauchy problem)**.** *The scheme (14) is stable in the sense of Definition 2.2 if and only if there exists a constant $C_1 > 0$ such that the amplification matrix $\mathscr{A}$ in (16) satisfies*

$$\forall\, n \in \mathbb{N}\,, \quad \forall\, \eta \in \mathbb{R}\,, \quad \left|\mathscr{A}(\mathrm{e}^{i\,\eta})^n\right|^2 \leq C_1\,. \tag{17}$$

Moreover, if the scheme is stable with a constant C_0, then one can take $C_1 = (s+1)\,C_0$ in (17), and conversely if (17) holds with a constant C_1, then one can take $C_0 = C_1$ for the stability estimate of Definition 2.2.

Proof of Proposition 2.2. • Let us assume that the bound (17) holds with the constant C_1, and let us consider the scheme (14) with initial data in ℓ^2. Then for all $n \in \mathbb{N}$, the step function U^n defined by

$$U^n(x) := U_j^n\,, \quad \text{for } x \in [j\,\Delta x, (j+1)\,\Delta x[\,,$$

satisfies

$$\forall\, n \geq s\,, \quad \forall\, x \in \mathbb{R}\,, \quad U^{n+1}(x) = \sum_{\sigma=0}^{s} \sum_{\ell=-r}^{p} A_{\ell,\sigma}\, U^{n-\sigma}(x + \ell\,\Delta x)\,.$$

It is clear that U^n belongs to $L^2(\mathbb{R})$ for all n (the operators Q_σ act boundedly on ℓ^2), so we can again apply Fourier transform and obtain

$$\forall\, n \geq s\,, \quad \forall\, \xi \in \mathbb{R}\,, \quad \widehat{U^{n+1}}(\xi) = \sum_{\sigma=0}^{s} \widehat{Q_\sigma}(\mathrm{e}^{i\,\Delta x\,\xi})\,\widehat{U^{n-\sigma}}(\xi)\,,$$

from which we deduce

$$\forall\, n \in \mathbb{N}\,, \quad \forall\, \xi \in \mathbb{R}\,, \quad \begin{pmatrix} \widehat{U^{n+s}}(\xi) \\ \vdots \\ \widehat{U^n}(\xi) \end{pmatrix} = \mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi})^n \begin{pmatrix} \widehat{U^s}(\xi) \\ \vdots \\ \widehat{U^0}(\xi) \end{pmatrix}.$$

Stability follows from Plancherel Theorem as in the proof of Proposition 2.1, and we get

$$\forall\, n \in \mathbb{N}\,, \quad \sum_{j\in\mathbb{Z}} \Delta x\, |U_j^n|^2 \leq C_1 \left(\sum_{j\in\mathbb{Z}} \Delta x\, |f_j^0|^2 + \cdots + \sum_{j\in\mathbb{Z}} \Delta x\, |f_j^s|^2 \right).$$

• Let us now assume that the scheme (14) is stable in the sense of Definition 2.2 with a constant C_0. Let $n \in \mathbb{N}$, $\eta \in \mathbb{R}$, and let $k \geq n\,\max(p,r)$. We also consider some vectors $X^0, \dots, X^s \in \mathbb{C}^N$ satisfying

$$|X^0|^2 + \cdots + |X^s|^2 = 1\,.$$

We consider the initial data

$$f_j^0 := \begin{cases} \mathrm{e}^{i\,j\,\eta}\,X^0\,, & \text{if } |j| \leq k\,, \\ 0\,, & \text{otherwise,} \end{cases} \quad \dots\,, \quad f_j^s := \begin{cases} \mathrm{e}^{i\,j\,\eta}\,X^s\,, & \text{if } |j| \leq k\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

For $|j| \leq k - \max(p,r)$, the relation (14) gives

$$U_j^{s+1} = \sum_{\sigma=0}^{s} \widehat{Q_\sigma}(\mathrm{e}^{i\,\eta})\,U_j^{s-\sigma}\,.$$

In particular, there holds $U_{j+\ell}^{s+1} = \mathrm{e}^{i\,\ell\,\eta}\,U_j^{s+1}$ for $|j| \leq k - 2\,\max(p,r)$ and $|\ell| \leq \max(p,r)$. Proceeding by induction, we get

$$U_j^{s+m+1} = \sum_{\sigma=0}^{s} \widehat{Q_\sigma}(\mathrm{e}^{i\,\eta})\,U_j^{s+m-\sigma}\,,$$

for all $m = 0, \dots, n-1$ and for all j satisfying $|j| \leq k - (m+1)\,\max(p,r)$. It is now not difficult to obtain the relation

$$\begin{pmatrix} U_j^{n+s} \\ \vdots \\ U_j^n \end{pmatrix} = \mathscr{A}(\mathrm{e}^{i\,\eta})^n \begin{pmatrix} f_j^s \\ \vdots \\ f_j^0 \end{pmatrix}, \quad \text{if } |j| \leq k - n\,\max(p,r)\,.$$

Then we have

$$\begin{aligned} \sum_{|j| \leq k - n\,\max(p,r)} \Delta x\,\left(|U_j^n|^2 + \cdots + |U_j^{n+s}|^2\right) &\leq \sum_{j \in \mathbb{Z}} \Delta x\,\left(|U_j^n|^2 + \cdots + |U_j^{n+s}|^2\right) \\ &\leq (s+1)\,C_0\,\left(|f_j^0|^2 + \cdots + |f_j^s|^2\right) \\ &= (s+1)\,C_0\,\Delta x\,(2\,k+1)\,. \end{aligned}$$

Eventually, we obtain

$$(2\,k + 1 - 2\,n\,\max(p,r))\,\Delta x\,\left|\mathscr{A}(\mathrm{e}^{i\,\eta})^n\,X\right|^2 \leq (s+1)\,C_0\,\Delta x\,(2\,k+1)\,, \quad X := \begin{pmatrix} X^s \\ \vdots \\ X^0 \end{pmatrix}.$$

Dividing by $\Delta x\,(2\,k+1)$, letting k tend to infinity and taking the supremum with respect to X, we obtain the result of Proposition 2.2. □

The following paragraph discusses how the results of Propositions 2.1 and 2.2 are useful in practice.

Remark 2.2. *When one tries to verify that the amplification matrix of a numerical scheme satisfies (10), resp. (17), the choice of the norm on $\mathscr{M}_N(\mathbb{C})$, resp. $\mathscr{M}_{N(s+1)}(\mathbb{C})$, is arbitrary because all norms are equivalent. It may sometimes be easier to work with the norm $\max_{i,j=1,\dots,N} |M_{i,j}|$, as we shall sometimes do below.*

2.2. **Possible behaviors for the eigenvalues of the amplification matrix.** In this paragraph, we recall some facts about families of matrices with uniformly bounded powers. We also analyze how the characterization of Propositions 2.1 and 2.2 can be simplified for a special class of numerical schemes.

The following result is elementary.

Lemma 2.1. *Let $M \in \mathscr{M}_d(\mathbb{C})$ be power bounded. Then $\rho(M) \leq 1$.*

Proof of Lemma 2.1. Let $\mu \in \mathrm{sp}(M)$, and let us choose an eigenvector $X \in \mathbb{C}^d$ with norm 1 associated with the eigenvalue μ. For all integer n, we have

$$|\mu|^n = |\mu^n\,X| = |M^n\,X| \leq C\,,$$

where the constant C is an upper bound for the norms of all powers M^n. The latter inequality gives $|\mu| \leq 1$ and the result follows. □

Lemma 2.1 immediately implies the following well-known necessary condition for stability.

Corollary 2.1 (von Neumann condition)**.** *Let us assume that the scheme (8), resp. (14), is stable in the sense of Definition 2.1, resp. 2.2. Then the amplification matrix $\mathscr{A}$ defined by (11), resp. (16), satisfies the so-called von Neumann condition*

$$\forall\,\eta \in \mathbb{R}\,, \quad \rho(\mathscr{A}(\mathrm{e}^{i\,\eta})) \leq 1\,. \tag{18}$$

Let us observe that for one step schemes satisfying the consistency condition (13), $\mathscr{A}(1)$ is the identity matrix so the upper bound 1 for the spectral radius allowed by the von Neumann condition is attained. In particular, when η is small, the eigenvalues of $\mathscr{A}(\mathrm{e}^{i\,\eta})$ should be close to 1 but remain within the closed unit disk. Usually, when one performs an expansion of the eigenvalues for small η, the requirement that the eigenvalues satisfy the von Neumann condition indicates some restrictions on the possible values of the CFL number λ.

The von Neumann condition in Corollary 2.1 is only a necessary condition for stability. However there is one case, that is always met in examples, where it is also a sufficient condition.

Lemma 2.2. *Let us assume that the matrices $A_{-r}, \dots, A_p$ in (9) can be simultaneously diagonalized (for instance when they are all polynomial functions of $\lambda\, A$). Then the scheme (8) is stable if and only if the von Neumann condition (18) holds.*

Proof of Lemma 2.2. The proof is elementary. Choosing an invertible matrix T that diagonalizes $A_{-r}, \dots, A_p$, the definition (11) shows that T also diagonalizes the amplification matrix $\mathscr{A}$, that is

$$\forall\, \kappa \in \mathbb{C} \setminus \{0\}\,, \quad T^{-1}\, \mathscr{A}(\kappa)\, T = \mathrm{diag}\, (z_1(\kappa), \dots, z_N(\kappa))\,.$$

If the von Neumann condition holds, the eigenvalues satisfy $|z_j(\mathrm{e}^{i\,\eta})| \leq 1$ for all $\eta \in \mathbb{R}$. This property implies

$$|\mathscr{A}(\mathrm{e}^{i\,\eta})^n| = |T\, \mathrm{diag}\, (z_1(\mathrm{e}^{i\,\eta})^n, \dots, z_N(\mathrm{e}^{i\,\eta})^n)\, T^{-1}| \leq |T|\, |T^{-1}|\,.$$

Proposition 2.1 shows that the scheme (8) is stable. □

The stability criterion of Lemma 2.2 will apply to all one step numerical schemes that appear in these notes. However, this criterion does not apply to multi-step schemes since the companion matrix $\mathscr{A}(\mathrm{e}^{i\,\eta})$ in (16) can not be diagonalized in a fixed basis that does not depend on η. We therefore need to work a little more. The following Lemma gives a more precise description of the properties of power bounded matrices.

Lemma 2.3. *A matrix $M \in \mathscr{M}_d(\mathbb{C})$ is power bounded if and only if $\rho(M) \leq 1$ and furthermore the eigenvalues of M whose modulus equals 1 are semi-simple (that is, their geometric multiplicity equals their algebraic multiplicity).*

Proof of Lemma 2.3. The proof is classical and appears in many textbooks on numerical analysis. Let $M \in \mathscr{M}_d(\mathbb{C})$ and let us consider an invertible matrix T that reduces M to its Jordan form

$$T^{-1}\, M\, T = \mathrm{diag}\, (M_1, \dots, M_p)\,,$$

where each block M_j is either of the form $\alpha_j\, I$ or a Jordan block

$$\begin{pmatrix} \alpha_j & 1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & \alpha_j \end{pmatrix},$$

whose size equals at least 2. (Let us observe that in this decomposition, the eigenvalues of the blocks M_j are not necessarily pairwise disctinct.) It is straightforward to check that M is power bounded if and only if each block is power bounded. We can now prove Lemma 2.3.

• Let us assume that M is power bounded. From Lemma 2.1, we already have $\rho(M) \leq 1$. If M is diagonalizable, then the proof is finished, so let us consider a Jordan block M_j that appears in the reduction of M and whose size is denoted d. Writing $M_j = \alpha_j\, I + N_j$, we have

$$M_j^n = \sum_{k=0}^{n} C_n^k\, \alpha_j^{n-k}\, N_j^k\,,$$

so the $(1,2)$-coefficient of M_j^n equals $n\, \alpha_j^{n-1}$ for all $n \geq 1$. Since all norms on the space $\mathscr{M}_d(\mathbb{C})$ are equivalent, there exists a constant C such that

$$\forall\, n \geq 1\,, \quad n\, |\alpha_j|^{n-1} \leq C\,,$$

and this implies $|\alpha_j| < 1$. In other words, eigenvalues of M that belong to the unit circle $\mathbb{S}^1$ must be semi-simple.

• Let us now assume that M satisfies $\rho(M) \leq 1$ and all eigenvalues of M that belong to $\mathbb{S}^1$ are semi-simple. In the Jordan reduction of M, the diagonal blocks are power bounded, so to prove Lemma 2.3, it only remains to prove that a Jordan block associated with an eigenvalue in $\mathbb{D}$ is power bounded. We keep the same notation $M_j = \alpha_j\, I + N_j$ as above. If $\alpha_j = 0$, then M_j is clearly power bounded, so we now assume $0 < |\alpha_j| < 1$. We have

$$M_j^n = \sum_{k=0}^{d-1} C_n^k\, \alpha_j^{n-k}\, N_j^k\,,$$

for all $n \geq d-1$ (here we have used $N_j^d = 0$). It is therefore sufficient to prove that for all fixed $k = 0, \ldots, d-1$, the sequence $(C_n^k \, \alpha_j^n)_{n \in \mathbb{N}}$ is bounded. This sequence tends geometrically to zero (use d'Alembert's criterion) so it is bounded and the sequence $(M_j^n)_{n \in \mathbb{N}}$ is also bounded. The proof of Lemma 2.3 is complete. □

For numerical schemes, Lemma 2.3 shows that in addition to the von Neumann condition, a necessary condition for stability is that if $\underline{\eta} \in \mathbb{R}$ is such that the matrix $\mathscr{A}(\mathrm{e}^{i\,\underline{\eta}})$ has an eigenvalue $\underline{z} \in \mathbb{S}^1$, then $\underline{z}$ should be semi-simple.

Lemma 2.3 is unfortunately not sufficient to characterize uniform power boundedness for an infinite family of matrices[3]. Indeed, let us consider the following matrices

$$M_1(x) := \begin{pmatrix} 1-x & x \\ 0 & 1-x \end{pmatrix}, \quad M_2(x) := \begin{pmatrix} 1-x^2 & x \\ 0 & 1-x^2 \end{pmatrix},$$

which both depend on a parameter $x \in [0,1]$. For all fixed $x \in [0,1]$, Lemma 2.3 shows that the matrices $M_1(x)$ and $M_2(x)$ are power bounded. However, it is not a difficult exercise to show that the family $\{M_1(x)\,,\, x \in [0,1]\}$ is uniformly power bounded while the family $\{M_2(x)\,,\, x \in [0,1]\}$ is not uniformly power bounded. As a matter of fact, there exists only one result that fully characterizes families of uniformly power bounded matrices. This famous Theorem is due to Kreiss and can be stated as follows.

Theorem 2.1 (Kreiss matrix Theorem)**.** *Let $d \in \mathbb{N}$ and let $\mathscr{F} \subset \mathscr{M}_d(\mathbb{C})$. The following conditions are equivalent.*

(i) *There exists a constant C_1 such that for all $M \in \mathscr{F}$ and for all $n \in \mathbb{N}$, $|M^n| \leq C_1$.*

(ii) *There exists a constant C_2 such that for all $M \in \mathscr{F}$, $\rho(M) \leq 1$ and for all $z \in \mathscr{U}$, there holds*

$$|(M - z\,I)^{-1}| \leq \frac{C_2}{|z|-1}\,.$$

(iii) *There exists a constant C_3 such that for all $M \in \mathscr{F}$, there exists an invertible matrix T such that $T^{-1}\,M\,T$ is upper triangular and*

$$|T| + |T^{-1}| \leq C_3\,,$$
$$\forall\, 1 \leq i < j \leq d\,, \quad |(T^{-1}\,M\,T)_{i,j}| \leq C_3 \, \min(1 - |(T^{-1}\,M\,T)_{i,i}|, 1 - |(T^{-1}\,M\,T)_{j,j}|)\,.$$

Rather than giving a complete proof of Theorem 2.1, which would take much space, we shall refer the interested reader to the nice review [22] where additional characterizations and historical references can be found. Showing that (i) implies (ii) is easy and follows from a series expansion. An elegant proof that (ii) implies (i) can be found in [23]. Optimal improvements of [23] are reported in [22].

The problem for showing uniform power boundedness for a parametrized family of matrices is to handle how a Jordan block may approach a diagonal block associated with an eigenvalue in $\mathbb{S}^1$ as the parameter varies. For numerical schemes in one space dimension, the pathology of the matrix $M_2(x)$ above is usually ruled out by the fact that as $\mathrm{e}^{i\,\eta}$ approaches a point $\mathrm{e}^{i\,\underline{\eta}}$ for which the amplification matrix has a semi-simple eigenvalue $\underline{z} \in \mathbb{S}^1$, the eigenvalues of $\mathscr{A}(\mathrm{e}^{i\,\eta})$ close to $\underline{z}$ are also semi-simple. Furthermore, eigenvalues and eigenvectors can usually be determined as smooth functions of η. A model situation for such behavior would be

$$\begin{pmatrix} 1-x^{2\,m_1} & 0 \\ 0 & 1-x^{2\,m_2} \end{pmatrix}, \quad m_1, m_2 \in \mathbb{N}\,, \quad x \in [0,1]\,.$$

To make this framework precise, we introduce the following terminology.

Definition 2.3 (Geometrically regular operator)**.** *The finite difference operator Q in (9), resp. the operators Q_σ in (15), is said to be geometrically regular if the amplification matrix $\mathscr{A}$ defined by (11), resp. (16), satisfies the following property: if $\underline{\kappa} \in \mathbb{S}^1$ and $\underline{z} \in \mathbb{S}^1 \cap \mathrm{sp}(\mathscr{A}(\underline{\kappa}))$ has algebraic multiplicity*

[3]The main reason is that the bound provided by Lemma 2.3 depends on the matrix T that reduces M to its Jordan form, and the construction of T is a highly ill-conditionned problem so that $|T|\,|T^{-1}|$ can be very large when M varies.

$\underline{\alpha}$, then there exist some functions $\beta_1(\kappa),\dots,\beta_{\underline{\alpha}}(\kappa)$ that are holomorphic in a neighborhood $\mathscr{W}$ of $\underline{\kappa}$ in $\mathbb{C}$ and that satisfy

$$\beta_1(\underline{\kappa}) = \cdots = \beta_{\underline{\alpha}}(\underline{\kappa}) = \underline{z}\,, \quad \det\big(z\,I - \mathscr{A}(\kappa)\big) = \vartheta(\kappa,z)\prod_{j=1}^{\underline{\alpha}}\big(z-\beta_j(\kappa)\big)\,,$$

with ϑ a holomorphic function of (κ,z) in some neighborhood of $(\underline{\kappa},\underline{z})$ such that $\vartheta(\underline{\kappa},\underline{z})\neq 0$, and if furthermore, there exist some vectors $e_1(\kappa),\dots,e_{\underline{\alpha}}(\kappa)\in\mathbb{C}^N$, resp. $\mathbb{C}^{N(s+1)}$, that depend holomorphically on $\kappa\in\mathscr{W}$, that are linearly independent for all $\kappa\in\mathscr{W}$, and that satisfy

$$\forall\,\kappa\in\mathscr{W}\,,\quad \forall\,j=1,\dots,\underline{\alpha}\,,\quad \mathscr{A}(\kappa)\,e_j(\kappa)=\beta_j(\kappa)\,e_j(\kappa)\,.$$

For instance, if the matrices $A_{-r},\dots,A_p$ satisfy the assumption of Lemma 2.2, it is clear that the finite difference operator Q in (9) is geometrically regular. Even better, in that case the eigenvalues and corresponding eigenvectors can be parametrized globally for $\kappa\neq 0$. The eigenvectors do not even depend on κ ! The framework of Definition 2.3 is therefore meaningful mostly for multi-step schemes, e.g. the leap-frog scheme. We hope that it will also be useful for the study of finite difference schemes in several space dimensions. We end this paragraph with the following characterization of stability by the von Neumann condition for geometrically regular operators.

Proposition 2.3 (Characterization of stability for geometrically regular operators)**.** *Let the finite difference operator Q in (9), resp. the operators Q_σ in (15), be geometrically regular. Then the scheme (8), resp. (14), is stable if and only if the von Neumann condition (18) holds.*

The precise expression, either (11) or (16), of the amplification matrix $\mathscr{A}$ is not relevant for the proof of Proposition 2.3. To unify both cases, we shall thus consider that the size of $\mathscr{A}$ is $N\,(s+1)$, which amounts to setting $s=0$ for one-step schemes.

Proof of Proposition 2.3. Using Corollary 2.1, it is sufficient to prove that the von Neumann condition implies stability. The proof of Proposition 2.3 consists in splitting the set of parameters $\eta\in\mathbb{R}$ into a first part for which the amplification matrix has eigenvalues close to $\mathbb{S}^1$ and a second part for which the eigenvalues of the amplification matrix are in $\mathbb{D}$, uniformly away from $\mathbb{S}^1$. In the first part, we shall use the geometric regularity assumption to control the powers of the amplification matrix. The second part will be easier to control. We begin with an easy consequence of Theorem 2.1 which will be useful later on.

Lemma 2.4. *Let $d\in\mathbb{N}$ and let $\mathscr{F}\subset\mathscr{M}_d(\mathbb{C})$ be a family of matrices such that there exists $\delta\in\,]0,1]$ for which*

$$\forall\,M\in\mathscr{F}\,,\quad \rho(M)\leq 1-\delta\,. \tag{19}$$

Then $\mathscr{F}$ is uniformly power bounded if and only if $\mathscr{F}$ is bounded in $\mathscr{M}_d(\mathbb{C})$.

Proof of Lemma 2.4. It is obvious that boundedness is a necessary condition for uniform power boundedness. Let now a family $\mathscr{F}\subset\mathscr{M}_d(\mathbb{C})$ be bounded and satisfy (19) for some positive δ, and let $M\in\mathscr{F}$. By Schur's Lemma, there exists a unitary matrix T such that $T^{-1}\,M\,T$ is upper triangular. Since $\mathscr{F}$ is bounded, while T and T^{-1} belong to the unitary group (a bounded subset of $\mathscr{M}_d(\mathbb{C})$), there exists a constant C that is independent of M and such that

$$\forall\,1\leq i<j\leq d\,,\quad |(T^{-1}\,M\,T)_{i,j}|\leq C\,.$$

From the assumption of Lemma 2.4, we also have

$$\min_{i=1,\dots,d}(1-|(T^{-1}\,M\,T)_{i,i}|)\geq\delta>0\,,$$

so it is easily seen that $\mathscr{F}$ satisfies condition (iii) of Theorem 2.1. The conclusion of Lemma 2.4 follows. □

The following observation is trivial and is stated without proof.

Lemma 2.5. *Let $\mathscr{K}:=\{\kappa\in\mathbb{S}^1\,,\mathrm{sp}(\mathscr{A}(\kappa))\cap\mathbb{S}^1\neq\emptyset\}$. Then $\mathscr{K}$ is a closed (hence compact) subset of $\mathbb{S}^1$.*

If $\mathscr{K}$ is empty (which never happens in practice, but let's pretend), then the von Neumann condition implies that for all $\kappa \in \mathbb{S}^1$, the spectrum of $\mathscr{A}(\kappa)$ is included in the open unit disk $\mathbb{D}$. Moreover, $\mathscr{A}(\kappa)$ depends holomorphically on $\kappa \in \mathbb{S}^1$ and $\mathbb{S}^1$ is a compact set, so there exists a constant $\delta > 0$ such that $\rho(\mathscr{A}(\kappa)) \leq 1 - \delta$ for all $\kappa \in \mathbb{S}^1$. (Here we use the continuity of the spectral radius.) Since $\{\mathscr{A}(\kappa)\,,\ \kappa \in \mathbb{S}^1\}$ is a bounded family, Lemma 2.4 shows that $\mathscr{A}(\kappa)$ is uniformly power bounded for $\kappa \in \mathbb{S}^1$ which completes the proof of Proposition 2.3.

Let us now assume that $\mathscr{K}$ is not empty. The following Lemma gives a description of $\mathscr{A}(\kappa)$ in the neighborhood of any point of $\mathscr{K}$.

Lemma 2.6. *For all $\underline{\kappa} \in \mathscr{K}$, there exist an integer q, two positive constants C and δ, an open neighborhood $\mathscr{W}$ of $\underline{\kappa}$ in $\mathbb{C}$ and an invertible matrix $T(\kappa)$ that depends holomorphically on $\kappa \in \mathscr{W}$ and that satisfies*

- *for all $\kappa \in \mathscr{W}$, $|T(\kappa)| + |T(\kappa)^{-1}| \leq C$,*
- *for all $\kappa \in \mathscr{W}$, $T(\kappa)^{-1}\,\mathscr{A}(\kappa)\,T(\kappa) = \mathrm{diag}\,(\beta_1(\kappa), \ldots, \beta_q(\kappa), \mathscr{A}_\sharp(\kappa))$, with $\beta_1(\kappa), \ldots, \beta_q(\kappa) \in \mathbb{C}$, $\mathscr{A}_\sharp(\kappa) \in \mathscr{M}_{N(s+1)-q}(\mathbb{C})$, $|\mathscr{A}_\sharp(\kappa)| \leq C$ and $\rho(\mathscr{A}_\sharp(\kappa)) \leq 1 - \delta$.*

Let us complete the proof of Proposition 2.3 using Lemma 2.6. We use a finite covering of the compact set $\mathscr{K}$ by open sets $\mathscr{W}_1, \ldots, \mathscr{W}_K \subset \mathbb{C}$ given in Lemma 2.6 (on each $\mathscr{W}_k$, we have a change of basis $T_k(\kappa)$ that satisfies suitable properties). Let now $\kappa = \mathrm{e}^{i\,\eta} \in \mathbb{S}^1 \cap \mathscr{W}_k$ with $1 \leq k \leq K$. The von Neumann condition shows that the eigenvalues of $\mathscr{A}(\kappa)$ belong to $\mathbb{D} \cup \mathbb{S}^1$. Moreover, there exist some positive constants C_k and δ_k that do not depend on $\kappa \in \mathbb{S}^1 \cap \mathscr{W}_k$ such that the diagonal block $\mathscr{A}_\sharp(\kappa)$ satisfies $|\mathscr{A}_\sharp(\kappa)| \leq C_k$ and $\rho(\mathscr{A}_\sharp(\kappa)) \leq 1 - \delta_k$. Applying Lemma 2.4, we find that the family $\{\mathscr{A}_\sharp(\kappa)\,,\ \kappa \in \mathbb{S}^1 \cap \mathscr{W}_k\}$ is uniformly power bounded. Using the block diagonal decomposition of $\mathscr{A}(\kappa)$, it follows that the family of matrices $\{\mathscr{A}(\kappa)\,,\ \kappa \in \mathbb{S}^1 \cap \mathscr{W}_k\}$ is also uniformly power bounded. (Here we use the property $|\beta_j(\kappa)| \leq 1$ for $\kappa \in \mathbb{S}^1 \cap \mathscr{W}_k$ which follows from the von Neumann condition.) In other words, there exists a constant $C_1 > 0$ such that

$$\forall\, \kappa \in \mathbb{S}^1 \cap (\mathscr{W}_1 \cup \cdots \cup \mathscr{W}_K)\,, \quad \forall\, n \in \mathbb{N}\,, \quad |\mathscr{A}(\kappa)^n| \leq C_1\,.$$

For κ in the closed (hence compact) subset $\mathbb{S}^1 \setminus (\mathscr{W}_1 \cup \cdots \cup \mathscr{W}_K)$ of $\mathbb{S}^1$, we know that the spectrum of $\mathscr{A}(\kappa)$ lies inside $\mathbb{D}$. Consequently, there exists a constant $\delta' > 0$ such that $\rho(\mathscr{A}(\kappa)) \leq 1 - \delta'$ for $\kappa \in \mathbb{S}^1 \setminus (\mathscr{W}_1 \cup \cdots \cup \mathscr{W}_K)$. Applying Lemma 2.4 again, there exists a constant $C_2 > 0$ such that

$$\forall\, \kappa \in \mathbb{S}^1 \setminus (\mathscr{W}_1 \cup \cdots \cup \mathscr{W}_K)\,, \quad \forall\, n \in \mathbb{N}\,, \quad |\mathscr{A}(\kappa)^n| \leq C_2\,.$$

Consequently the matrix $\mathscr{A}(\kappa)$ is uniformly power bounded for $\kappa \in \mathbb{S}^1$, and the proof of Proposition 2.3 is complete. It only remains to prove Lemma 2.6... Since it is the first occurence in these notes of arguments that will appear in several other places, we give a detailed proof of Lemma 2.6. Similar arguments will be sometimes used as a "black box" later on.

Proof of Lemma 2.6. Let $\underline{\kappa} \in \mathscr{K}$. From the von Neumann condition, the eigenvalues of the amplification matrix split into two groups: eigenvalues on $\mathbb{S}^1$ and eigenvalues in $\mathbb{D}$. We let $\underline{z}_1, \ldots, \underline{z}_m$ denote the pairwise distinct eigenvalues of $\mathscr{A}(\underline{\kappa})$ on $\mathbb{S}^1$. The corresponding algebraic multiplicities are denoted $\underline{\alpha}_1, \ldots, \underline{\alpha}_m$. We also introduce the notation $q := \underline{\alpha}_1 + \cdots + \underline{\alpha}_m$.

Let us consider an open neighborhood $\mathscr{W}$ of $\underline{\kappa}$ and a positive constant δ such that for all $\kappa \in \mathscr{W}$, the eigenvalues of $\mathscr{A}(\kappa)$ belong to one of the following sets:

$$\{\zeta \in \mathbb{C}\,,\ |\zeta - \underline{z}_1| \leq \delta\}\,, \ldots,\ \{\zeta \in \mathbb{C}\,,\ |\zeta - \underline{z}_m| \leq \delta\}\,,\ \{\zeta \in \mathbb{C}\,,\ |\zeta| \leq 1 - 3\,\delta\}\,.$$

Up to shrinking δ and $\mathscr{W}$, we can always assume that these disks do not intersect. Hence the disk with center $\underline{z}_1$ contains $\underline{\alpha}_1$ eigenvalues of $\mathscr{A}(\kappa)$, the disk with center $\underline{z}_m$ contains $\underline{\alpha}_m$ eigenvalues, and the disk centered at the origin containes $N\,(s+1) - q$ eigenvalues.

For $\kappa \in \mathscr{W}$, the spectral projector $\Pi(\kappa)$ of $\mathscr{A}(\kappa)$ on the generalized eigenspace $E(\kappa)$ associated with eigenvalues in $\{\zeta \in \mathbb{C}\,,\ |\zeta| \leq 1 - 3\,\delta\}$ is given by the Dunford-Taylor formula

$$\Pi(\kappa) = \frac{1}{2\,i\,\pi} \int_{\{|w| = 1 - 2\,\delta\}} (w\,I - \mathscr{A}(\kappa))^{-1}\,\mathrm{d}w\,.$$

The projector $\Pi(\kappa)$ depends holomorphically on $\kappa \in \mathscr{W}$, and its image has rank $N(s+1)-q$. Choosing a basis $\underline{e}_{q+1}, \dots, \underline{e}_{N(s+1)}$ of the generalized eigenspace $E(\underline{\kappa})$, the vectors

$$\Pi(\kappa)\,\underline{e}_{q+1}\,, \dots,\, \Pi(\kappa)\,\underline{e}_{N(s+1)}\,,$$

are linearly independent for κ sufficiently close to $\underline{\kappa}$, and moreover they belong to $E(\kappa)$. We have thus constructed a basis $(e_{q+1}(\kappa), \dots, e_{N(s+1)}(\kappa))$ of $E(\kappa)$ that depends holomorphically on κ for κ sufficiently close to $\underline{\kappa}$ (that is, for all $\kappa \in \mathscr{W}$ up to shrinking $\mathscr{W}$).

The geometric regularity condition shows that the $\underline{\alpha}_1$ eigenvalues of $\mathscr{A}(\kappa)$ which belong to the disk $\{\zeta \in \mathbb{C}\,,\, |\zeta - \underline{z}_1| \leq \delta\}$ behave holomorphically on κ. Collecting the eigenvalues of $\mathscr{A}(\kappa)$ which do not contribute to $E(\kappa)$, there exist some holomorphic functions $\beta_1, \dots, \beta_q$ on the neighborhood $\mathscr{W}$ of $\underline{\kappa}$ such that

$$\forall\, \kappa \in \mathscr{W}\,, \quad \mathrm{sp}(\mathscr{A}(\kappa)) \cap \{\zeta \in \mathbb{C}\,,\, |\zeta| > 1 - 3\,\delta\} = \{\beta_1(\kappa)\,, \dots,\, \beta_q(\kappa)\}\,.$$

The geometric regularity condition also shows that the eigenvalues $\beta_i(\kappa)$ admit some eigenvectors $e_i(\kappa)$ that are defined holomorphically on the neighborhood $\mathscr{W}$. To complete the proof of Lemma 2.6, it remains to observe that the vectors $e_1(\underline{\kappa}), \dots, e_{N(s+1)}(\underline{\kappa})$ are linearly independent, so this property remains true for all $\kappa \in \mathscr{W}$, up to shrinking $\mathscr{W}$ once again. We have therefore constructed the column vectors of the invertible matrix $T(\kappa)$. Since T and T^{-1} are holomorphic, they are also bounded up to shrinking $\mathscr{W}$. □

□

To conclude this paragraph, we show that geometric regularity can also arise as a necessary condition for stability of a finite difference scheme. In one space dimension, this notion seems to be central in the analysis of the discrete Cauchy problem and we shall see in the next sections that it also plays a central role in the analysis of discrete initial boundary value problems. Our result is the following.

Lemma 2.7. *Let us consider the numerical scheme (8), resp. (14), in the scalar case $N = 1$. If (8), resp. (14), is stable in the sense of Definition 2.1, resp. Definition 2.2, then the finite difference operator Q, resp. the operators Q_σ, is geometrically regular.*

Proof of Lemma 2.7. In the case of the one step scheme (8), the amplification matrix $\mathscr{A}(\kappa)$ in (11) is a complex number so geometric regularity is trivial. The only coefficient of $\mathscr{A}$ depends holomorphically on $\kappa \in \mathbb{C} \setminus \{0\}$ and the eigenvector is independent of κ. We thus turn to the case of multistep schemes. The proof of Lemma 2.7 relies on a very simple observation which we state separately since it will be useful later on.

Lemma 2.8. *Let $M \in \mathscr{M}_m(\mathbb{C})$ be a companion matrix, that is*

$$M = \begin{pmatrix} \mu_1 & \dots & \dots & \mu_m \\ 1 & 0 & \dots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then for all eigenvalue λ of M, the dimension of $\mathrm{Ker}\,(M - \lambda\, I)$ equals 1 and the eigensapce is spanned by the vector $(\lambda^{m-1}, \dots, \lambda, 1)^T$.

The proof of Lemma 2.8 follows from a simple calculation and is omitted. Let us complete the proof of Lemma 2.7. If (14) is stable, we know that the amplification matrix $\mathscr{A}(\kappa)$ in (16) is uniformly power bounded for $\kappa \in \mathbb{S}^1$. Let us now consider a point $\underline{\kappa} \in \mathbb{S}^1$ for which $\mathscr{A}(\underline{\kappa})$ has an eigenvalue $\underline{z} \in \mathbb{S}^1$. According to Lemma 2.3, we know that $\underline{z}$ is a semi-simple eigenvalue. Since Lemma 2.8 shows that the geometric multiplicity of $\underline{z}$ equals 1, we can conclude that $\underline{z}$ is a simple eigenvalue of $\mathscr{A}(\underline{\kappa})$. In particular, the Weierstrass preparation Theorem shows that for κ in a neighborhood of $\underline{\kappa}$, $\mathscr{A}(\kappa)$ has a unique simple eigenvalue $\beta(\kappa)$ that depends holomorphically on κ such that $\beta(\underline{\kappa}) = \underline{z}$. Lemma 2.8 shows that the eigenspace $\mathrm{Ker}\,(\mathscr{A}(\kappa) - \beta(\kappa)\, I)$ is spanned by the vector $(\beta(\kappa)^{m-1}, \dots, \beta(\kappa), 1)^T$ which also depends holomorphically on κ. We have thus proved that the operators Q_σ in (14) are geometrically regular. □

We now show on a series of examples that either Lemma 2.2 or Proposition 2.3 can be used to prove stability for various well-known numerical schemes. In these examples, we shall also be interested in giving a precise description of the eigenvalues of $\mathscr{A}(\kappa)$ near a point $\underline{\kappa}$ where the spectrum of $\mathscr{A}(\underline{\kappa})$ meets $\mathbb{S}^1$.

2.3. **The Lax-Friedrichs and leap-frog schemes.** The Lax-Friedrichs approximation of (5) corresponds to the scheme (8) where

$$p = r = 1\,, \quad Q_{LF} := \frac{I + \lambda\, A}{2}\, \mathbf{T}^{-1} + \frac{I - \lambda\, A}{2}\, \mathbf{T}\,.$$

In other words, the corresponding numerical scheme reads

$$\begin{cases} U_j^{n+1} = \dfrac{U_{j-1}^n + U_{j+1}^n}{2} - \dfrac{\lambda\, A}{2}\,(U_{j+1}^n - U_{j-1}^n)\,, & j \in \mathbb{Z}\,, \quad n \geq 0\,, \\ U_j^0 = f_j\,, & j \in \mathbb{Z}\,. \end{cases} \tag{20}$$

We recall that the CFL number λ is a constant that is fixed from the beginning and that stands for the ratio $\Delta t/\Delta x$. Since A is diagonalizable with real eigenvalues, the result of Lemma 2.2 applies and stability for (20) is encoded in the von Neumann condition. Letting $\lambda_1, \dots, \lambda_N$ denote the eigenvalues of A and letting T denote an invertible matrix that diagonalizes A, an easy computation gives[4]

$$T^{-1}\, \mathscr{A}_{LF}(\mathrm{e}^{i\,\eta})\, T = \mathrm{diag}\ (z_1(\eta), \dots, z_N(\eta))\,, \quad z_j(\eta) := \cos\eta - i\,\lambda\,\lambda_j\,\sin\eta\,.$$

In particular, we have

$$|z_j(\eta)|^2 = \cos^2\eta + (\lambda\,\lambda_j)^2\,\sin^2\eta = 1 + [(\lambda\,\lambda_j)^2 - 1]\,\sin^2\eta\,. \tag{21}$$

It is easy to deduce from (21) that if λ satisfies $\lambda\,\rho(A) \leq 1$, then the von Neumann condition (18) is satisfied and the scheme (20) is stable. Conversely, let us consider the case where λ satisfies $\lambda\,\rho(A) > 1$, with for instance $\lambda\,|\lambda_1| > 1$. For small η, we compute

$$|z_1(\eta)|^2 = 1 + [(\lambda\,\lambda_1)^2 - 1]\,\eta^2 + O(\eta^4)\,.$$

In particular, there holds $|z_1(\eta)| > 1$ for all $\eta \neq 0$ sufficiently small. Corollary 2.1 then shows that (20) is not stable. Summing up our computations, we have proved that the Lax-Friedrichs scheme (20) is stable if and only if $\lambda\,\rho(A) \leq 1$.

Let us now fix a constant $\lambda > 0$ such that $\lambda\,\rho(A) \leq 1$. We wish to study the behavior of the eigenvalues $z_j(\eta)$ near points where these eigenvalues touch the unit circle. A first possible case is when λ satisfies $\lambda\,|\lambda_j| = 1$ for some index j. Then $z_j(\eta) \in \mathbb{S}^1$ for all $\eta \in \mathbb{R}$. Moreover, it is easy to verify the property $z_j'(\eta) \neq 0$ in this case. Consequently, the parametrized curve $\{z_j(\eta)\,,\, \eta \in \mathbb{R}\}$ coincides with $\mathbb{S}^1$ and contains only regular points. The second possible case is when λ satisfies $\lambda\,|\lambda_j| < 1$. Then (21) shows that $z_j(\eta) \in \mathbb{S}^1$ if and only if $\eta \in \mathbb{Z}\,\pi$. Furthermore, there holds $z_j'(0) = -z_j'(\pi) = -i\,\lambda\,\lambda_j$. Assuming for simplicity that A is invertible, so that all the eigenvalues λ_j are nonzero, the parametrized curve $\{z_j(\eta)\,,\, \eta \in \mathbb{R}\}$ is an ellipse that is included in the unit disk, and that meets the unit circle at ± 1 which correspond to regular points. (When 0 is an eigenvalue of A, the corresponding eigenvalue of the amplification matrix yields a parametrized curve that describes the segment $[-1, 1]$, whose contact points ± 1 with $\mathbb{S}^1$ are singular points.)

The leap-frog scheme is more or less the most simple approximation of (5) with a two-steps scheme. It corresponds to the scheme (14) where

$$s = p = r = 1\,, \quad Q_{lf,0} := -\lambda\, A\,(\mathbf{T} - \mathbf{T}^{-1})\,, \quad Q_{lf,1} := I\,.$$

In other words, the corresponding numerical scheme reads

$$\begin{cases} U_j^{n+1} = U_j^{n-1} - \lambda\, A\,(U_{j+1}^n - U_{j-1}^n)\,, & j \in \mathbb{Z}\,, \quad n \geq 0\,, \\ U_j^0 = f_j^0\,, & j \in \mathbb{Z}\,, \\ U_j^1 = f_j^1\,, & j \in \mathbb{Z}\,. \end{cases} \tag{22}$$

[4]The reader should be cautious with the notation λ which stands for the CFL number and the notation λ_j which stands for the eigenvalues of A.

In this case, the amplification matrix is the block companion matrix

$$\mathscr{A}_{lf}(\kappa) := \begin{pmatrix} -\lambda\,(\kappa-\kappa^{-1})\,A & I \\ I & 0 \end{pmatrix}.$$

Diagonalizing A and permuting rows and columns, there exists a fixed invertible matrix T such that

$$T^{-1}\,\mathscr{A}_{lf}(\kappa)\,T := \operatorname{diag}\left(\begin{pmatrix} -\lambda\,\lambda_1\,(\kappa-\kappa^{-1}) & 1 \\ 1 & 0 \end{pmatrix},\dots,\begin{pmatrix} -\lambda\,\lambda_N\,(\kappa-\kappa^{-1}) & 1 \\ 1 & 0 \end{pmatrix}\right).$$

Our goal is first to determine the CFL parameters λ for which the von Neumann condition holds. For a fixed index j and a real number η, we need to determine the eigenvalues of the matrix

$$\begin{pmatrix} -2\,i\,\lambda\,\lambda_j\,\sin\eta & 1 \\ 1 & 0 \end{pmatrix}.$$

The eigenvalues are the roots to the polynomial equation

$$(\omega + i\,\lambda\,\lambda_j\,\sin\eta)^2 + (\lambda\,\lambda_j)^2\,\sin^2\eta - 1 = 0\,. \tag{23}$$

Let us first consider the case $\lambda\,|\lambda_j| > 1$. Then choosing $\eta = \pi/2$, there exists one purely imaginary root of (23) whose modulus equals $\lambda\,|\lambda_j| + \sqrt{(\lambda\,\lambda_j)^2 - 1}$ and the von Neumann condition is not satisfied. Let us now consider the case $\lambda\,|\lambda_j| = 1$. Choosing $\underline{\eta} = \pi/2\,\mathrm{sgn}(\lambda\,\lambda_j)$, $-i$ is a double eigenvalue and the corresponding 2×2 matrix is not diagonalizable. This shows that when $\lambda\,\rho(A)$ equals 1, there exists a non-semi-simple eigenvalue $\underline{z}\in\mathbb{S}^1$ of $\mathscr{A}_{lf}(\mathrm{e}^{i\,\underline{\eta}})$. Using Lemma 2.3, the scheme can not be stable.

Eventually, let us show that in the case $\lambda\,\rho(A) < 1$ the leap-frog scheme (22) is stable. We are going to apply Proposition 2.3. Since $\lambda\,|\lambda_j| < 1$, the roots to the polynomial equation (23) are

$$\omega_{1,j}(\eta) := \sqrt{1-(\lambda\,\lambda_j)^2\,\sin^2\eta} - i\,\lambda\,\lambda_j\,\sin\eta\,,\quad \omega_{2,j}(\eta) := -\sqrt{1-(\lambda\,\lambda_j)^2\,\sin^2\eta} - i\,\lambda\,\lambda_j\,\sin\eta\,.$$

Both roots $\omega_{1,j},\omega_{2,j}$ are real analytic functions, and $\omega_{1,j}-\omega_{2,j}$ does not vanish. Furthermore, it is straightforward to check that both eigenvalues $\omega_{1,j}(\eta),\omega_{2,j}(\eta)$ belong to $\mathbb{S}^1$ for all $\eta\in\mathbb{R}$. Let $\underline{\kappa}\in\mathbb{S}^1$. We have already seen that the spectrum of the amplification matrix $\mathscr{A}(\underline{\kappa})$ is included in $\mathbb{S}^1$. The eigenvalues of each matrix

$$\begin{pmatrix} -\lambda\,\lambda_j\,(\underline{\kappa}-\underline{\kappa}^{-1}) & 1 \\ 1 & 0 \end{pmatrix},$$

are simple. For $\kappa\in\mathbb{C}$ in a sufficiently small neighborhood of $\underline{\kappa}$, the two eigenvalues and corresponding eigenvectors of

$$\begin{pmatrix} -\lambda\,\lambda_j\,(\kappa-\kappa^{-1}) & 1 \\ 1 & 0 \end{pmatrix},$$

depend holomorphically on κ. Collecting the eigenvalues and eigenvectors of each diagonal block in $\mathscr{A}(\kappa)$, we have proved that the operators in the leap-frog scheme are geometrically regular when $\lambda\,\rho(A) < 1$. Applying Proposition (2.3), we conclude that the leap-frog scheme is stable (and in this case it is also geometrically regular) if and only if $\lambda\,\rho(A) < 1$. In that case, the parametrized curve $\{\omega_{1,j}(\eta)\,,\,\eta\in\mathbb{R}\}$ describes part of the unit circle $\mathbb{S}^1$, and it has exactly two singular points of order 2 corresponding to the values $\eta-\pi/2\in\mathbb{Z}\,\pi$. (The curve parametrized by $\omega_{2,j}$ has exactly the same behavior.)

Let us develop here an elementary calculation which shows stability for the leap-frog scheme (22) when $\lambda\,\rho(A) < 1$. We make the additional assumption that the matrix A is symmetric, and therefore $|A| = \rho(A)$. We start from the relation (22), take the scalar product with the vector $U_j^{n+1}+U_j^{n-1}$ and sum with respect to $j\in\mathbb{Z}$. This yields

$$\sum_{j\in\mathbb{Z}}|U_j^{n+1}|^2 - \sum_{j\in\mathbb{Z}}|U_j^{n-1}|^2 = -\sum_{j\in\mathbb{Z}}(U_j^{n+1})^*\,\left[\lambda\,A\,(U_{j+1}^n - U_{j-1}^n)\right] - \sum_{j\in\mathbb{Z}}(U_j^{n-1})^*\,\left[\lambda\,A\,(U_{j+1}^n - U_{j-1}^n)\right].$$

Using the symmetry of A and performing a "discrete integration by parts", we obtain

$$\sum_{j\in\mathbb{Z}}|U_j^{n+1}|^2 - \sum_{j\in\mathbb{Z}}|U_j^{n-1}|^2 = \sum_{j\in\mathbb{Z}}\left[\lambda\,A\,(U_{j+1}^{n+1} - U_{j-1}^{n+1})\right]^*\,U_j^n - \sum_{j\in\mathbb{Z}}(U_j^{n-1})^*\,\left[\lambda\,A\,(U_{j+1}^n - U_{j-1}^n)\right].$$

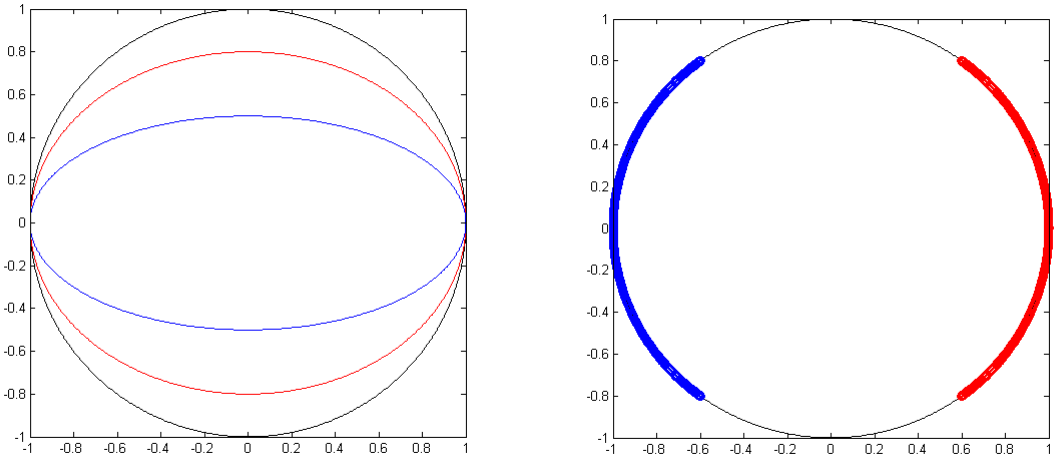

FIGURE 1. Left : parametrized curves of eigenvalues for the Lax-Friedrichs scheme (20) (the unit circle in black, the eigenvalue curve for $\lambda\,|\lambda_j| = 0.8$ in red, and the eigenvalue curve for $\lambda\,|\lambda_j| = 0.5$ in blue). Right : parametrized curves of eigenvalues for the leap-frog scheme (22) (the unit circle in black, the eigenvalues curves for $\lambda\,|\lambda_j| = 0.8$ in red and blue).

We use the latter relation for both cases n odd and n even, and sum the corresponding two equalities. Using new indeces and summing with respect to n, we obtain

$$\sum_{j\in\mathbb{Z}} |U_j^{2\,n}|^2 + \sum_{j\in\mathbb{Z}} |U_j^{2\,n+1}|^2 - \sum_{j\in\mathbb{Z}} |f_j^0|^2 - \sum_{j\in\mathbb{Z}} |f_j^1|^2 \\ = \sum_{j\in\mathbb{Z}} \left[\lambda\, A\,(U_{j+1}^{2\,n+1} - U_{j-1}^{2\,n+1})\right]^*\, U_j^{2\,n} - \sum_{j\in\mathbb{Z}} \left[\lambda\, A\,(f_{j+1}^1 - f_{j-1}^1)\right]^*\, f_j^0\,.$$

Applying Cauchy-Schwarz inequality and collecting terms, we obtain

$$(1-\lambda\,|A|)\sum_{j\in\mathbb{Z}} |U_j^{2\,n}|^2 + |U_j^{2\,n+1}|^2 \le (1+\lambda\,|A|)\sum_{j\in\mathbb{Z}} |f_j^0|^2 + |f_j^1|^2\,.$$

Multiplying by Δx, we have thus proved stability for (22) under the assumption that A is symmetric and satisfies $\lambda\,\rho(A) < 1$. Of course, this "energy method" based on integration by parts does not predict instability in the case $\lambda\,\rho(A) \ge 1$, neither does it give information about the behavior of the eigenvalues of the amplification matrix.

For the Lax-Friedrichs and leap-frog schemes, we have focused on the description of the parametrized curves $\{z_j(\eta)\}$, where $z_j(\eta)$ is an eigenvalue of the amplification matrix $\mathscr{A}(\mathrm{e}^{i\,\eta})$. In these two examples, the eigenvalues can be parametrized globally by smooth periodic functions of $\eta \in \mathbb{R}$. Such curves are represented in Figure 1. The following paragraph will give examples of numerical schemes for which the eigenvalues can still be parametrized globally but the associated parametrized curve can have a more complex behavior than above. It is important to understand which are the possible behaviors for these curves since these observations will play an important role in Section 3.

2.4. **The Runge-Kutta schemes or how to produce singular points of even order.** In this paragraph we follow [9, chapter 6] and introduce a class of high order numerical schemes based on the Runge-Kutta approximation for ordinary differential equations. The general method is the following: we start from (5) and first introduce a discretization of the space variable (this is usually called *semi-discretization*). This amounts to introducing a space step $\Delta x > 0$ and approximating the solution $u(t,x)$ to (5) by a sequence of function $(v_j(t))_{j\in\mathbb{Z}}$ where for all $j \in \mathbb{Z}$, $v_j(t)$ represents an

approximation of $u(t, j\,\Delta x)$. One example is obtained by observing that for all sufficiently smooth function f, there holds

$$\frac{2}{3\,\varepsilon}(f(\varepsilon) - f(-\varepsilon)) - \frac{1}{12\,\varepsilon}(f(2\,\varepsilon) - f(-2\,\varepsilon)) = f'(0) + O(\varepsilon^4)\,.$$

Then the Cauchy problem (5) can be approximated by the semi-discrete problem[5]

$$\begin{cases} \dot v_j = -\dfrac{2}{3\,\Delta x}\,A\,(v_{j+1} - v_{j-1}) + \dfrac{1}{12\,\Delta x}\,A\,(v_{j+2} - v_{j-2})\,, & j \in \mathbb{Z}\,,\ t \geq 0\,, \\ v_j(0) = f(j\,\Delta x)\,, & j \in \mathbb{Z}\,. \end{cases}$$

The latter problem is a linear (infinite) system of ordinary differential equations for which we can apply the fourth order Runge-Kutta integration rule[6] with time step $\Delta t = \lambda\,\Delta x$ (recall that the CFL number λ is a fixed constant). The following observation follows from a straightforward computation: for a linear ordinary differential equation

$$\dot X = L\,X\,, \quad X(0) = X_0\,,$$

the fourth order Runge-Kutta method reads

$$X_{n+1} = \sum_{k=0}^{4} \frac{(\Delta t\,L)^k}{k!}\,X_n\,.$$

Applying this rule to the above linear system for the v_j's, we obtain the following fully discrete approximation for (5):

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=0}^{4} \frac{(\lambda\,A\,\widetilde Q)^k}{k!}\,U_j^n\,, & j \in \mathbb{Z}\,,\ n \in \mathbb{N}\,, \\ U_j^0 = f_j\,, & j \in \mathbb{Z}\,, \end{cases} \qquad \widetilde Q := -\frac{2}{3}\,(\mathbf{T} - \mathbf{T}^{-1}) + \frac{1}{12}\,(\mathbf{T}^2 - \mathbf{T}^{-2})\,. \tag{24}$$

The scheme (24) can be written under the form (8), (9) with $p = r = 8$. Our goal is now to determine the values of the CFL number λ for which the scheme (24) is stable. Applying Lemma 2.2, we already know that it is sufficient to verify the von Neumann condition. Once again, we let $\lambda_1, \dots, \lambda_N$ denote the (real) eigenvalues of A, and we compute the eigenvalues of the corresponding amplification matrix $\mathscr{A}$ by diagonalizing A. The eigenvalues $z_1(\eta), \dots, z_N(\eta)$ of $\mathscr{A}(\mathrm{e}^{i\,\eta})$ are given by

$$\forall\, j = 1, \dots, N\,, \quad z_j(\eta) = \sum_{\ell=0}^{4} \frac{(\lambda\,\lambda_j\,q(\eta))^\ell}{\ell!}\,, \quad q(\eta) := -i\,\frac{\sin\eta}{3}\,(4 - \cos\eta)\,.$$

The modulus of $z_j(\eta)$ is computed by using the fact that $q(\eta)$ is purely imaginary, and we obtain

$$|z_j(\eta)|^2 = 1 - \frac{(\lambda\,\lambda_j)^6}{52488}\,h(\eta)^6\,\left(1 - \frac{(\lambda\,\lambda_j)^2}{72}\,h(\eta)^2\right)\,, \quad h(\eta) := \sin\eta\,(4 - \cos\eta)\,. \tag{25}$$

It follows from (25) that the scheme (24) satisfies the von Neumann condition if and only if $\lambda\,\rho(A)\,\max_{\mathbb{R}}|h| \leq 6\sqrt{2}$. The maximum of $|h|$ on $\mathbb{R}$ can be explicitely computed (!) by studying the variations of h and we obtain

$$\max_{\mathbb{R}}|h| = \left(3 + \frac{\sqrt{6}}{2}\right)\,\sqrt{\sqrt{6} - \frac{3}{2}}\,.$$

The maximum value for $\lambda\,\rho(A)$ that ensures stability equals approximately 2.06. The reader can check that $|h|$ attains its maximum for $\eta \pm \eta_0 \in \mathbb{Z}\,2\,\pi$ where η_0 is uniquely determined by $\eta_0 \in\,]\pi/2, \pi[$ and $\cos\eta_0 = 1 - \sqrt{3/2}$.

We now wish to analyze the behavior of the parametrized curve $\{z_j(\eta)\,,\ \eta \in \mathbb{R}\}$ according to the possible values of $\lambda\,\lambda_j$. For simplicity again, we assume that 0 does not belong to $\mathrm{sp}(A)$. Let us first consider the case where $\lambda\,|\lambda_j|\,\max_{\mathbb{R}}|h| < 6\sqrt{2}$. Then it follows from (25) that $z_j(\eta)$ belongs to $\mathbb{S}^1$ if and only if $\eta \in \mathbb{Z}\,\pi$. Moreover, there holds $z_j(0) = z_j(\pi) = 1$, $z_j'(0) = -i\,\lambda\,\lambda_j \neq 0$ and

[5] Here we use the rather standard "dot" notation for the time derivative in an ordinary differential equation.

[6] We refer to [20] for an introduction to the discretization of ordinary differential equations.

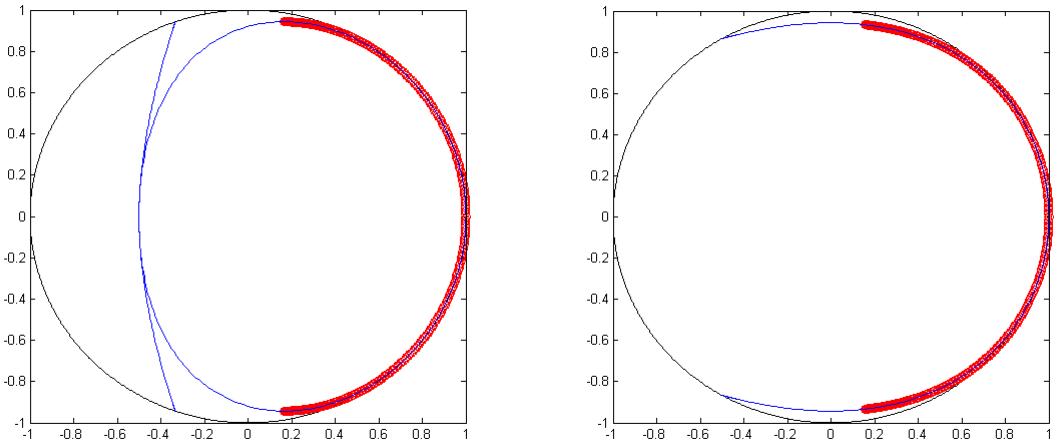

FIGURE 2. Left : parametrized curves of eigenvalues for the Runge-Kutta scheme (24) (the unit circle in black, the eigenvalue curve for $\lambda\,|\lambda_j|\,\max_{\mathbb{R}}|h| = 6\sqrt{2}\times 0.8$ in red, and the eigenvalue curve for $\lambda\,|\lambda_j|\,\max_{\mathbb{R}}|h| = 6\sqrt{2}$ in blue). Right : parametrized curves of eigenvalues for the Runge-Kutta scheme (27) (the unit circle in black, the eigenvalue curve for $\lambda\,|\lambda_j|\,M_J = 3\sqrt{3}/4$ in red and the eigenvalue curve for $\lambda\,|\lambda_j|\,M_J = \sqrt{3}$ in blue).

$z_j'(\pi) = 5\,i\,\lambda\,\lambda_j/3 \neq 0$. Consequently the curve $\{z_j(\eta)\,,\,\eta\in\mathbb{R}\}$ has one regular contact point with the unit circle (this point is attained in two different ways but each time it corresponds to a regular point). An example of such a curve is depicted in red in the left picture of Figure 2. The unit circle is depicted in black.

Let us now consider the more interesting case where $\lambda\,|\lambda_j|\,\max_{\mathbb{R}}|h| = 6\sqrt{2}$, and let us even further assume $\lambda_j > 0$, the case $\lambda_j < 0$ being entirely similar. The formula (25) shows that $z_j(\eta)\in\mathbb{S}^1$ if and only if $\eta\in\mathbb{Z}\,\pi$ or $\eta\pm\eta_0\in\mathbb{Z}\,2\,\pi$. As above, we compute $z_j(0) = z_j(\pi) = 1$, $z_j'(0) = -i\,\lambda\,\lambda_j \neq 0$ and $z_j'(\pi) = 5\,i\,\lambda\,\lambda_j/3 \neq 0$. We also compute $z_j'(\pm\eta_0) = 0$ since $h'(\pm\eta_0) = 0$. An elementary calculation yields the relations

$$z_j(\eta_0) = \overline{z_j(-\eta_0)} = -\frac{1}{3} + i\,\frac{2\sqrt{2}}{3}\,,\quad z_j''(\eta_0) = \overline{z_j''(-\eta_0)} = \lambda\,\lambda_j\,h''(\eta_0)\left(\frac{2\sqrt{2}}{9} + i\right),\quad h''(\eta_0) < 0\,.$$

The points $z_j(\pm\eta_0)$ are singular points of order 2 on the curve $\{z_j(\eta)\,,\,\eta\in\mathbb{R}\}$. Moreover, there exists a constant $c > 0$ such that for all η close to η_0, there holds

$$|z_j(\eta)| = 1 - c\,(\eta-\eta_0)^2 + o((\eta-\eta_0)^2)\,,$$

and there is a similar behavior in the neighborhood of $-\eta_0$. The curve parametrized by z_j is depicted in blue in the left picture of Figure 2.

The scheme (24) gives an example for an eigenvalue z_j of the amplification matrix such that the curve $\{z_j(\eta)\,,\,\eta\in\mathbb{R}\}$ has a singular contact point of order 2 with $\mathbb{S}^1$ and this curve is not included in $\mathbb{S}^1$ (as was the case with the leap-frog scheme). As a matter of fact, it is now not so difficult to generalize the example (24) in order to give an example of a stable scheme which produces some eigenvalues whose corresponding parametrized curves have a singular contact point with $\mathbb{S}^1$ of arbitrarily large even order. Moreover these parametrized curves will not be included in $\mathbb{S}^1$. Let us detail how this generalization can be performed.

Let us consider an integer $J\in\mathbb{N}$ that is fixed once and for all. Then we define the numbers

$$\forall\, j = 0,\dots,J\,,\quad q_j := \frac{C_{2\,J+1}^{J-j}}{2^{2\,J+1}\,(2\,j+1)}\,, \tag{26}$$

where C_n^k denotes the binomial coefficient. Using these numbers, we define the following finite difference operator (we feel free to use similar notation as above)

$$\widetilde{Q} := \sum_{j=0}^{J} q_j \left(\mathbf{T}^{1+2\,j} - \mathbf{T}^{-1-2\,j} \right).$$

This operator is constructed as an approximation of the space derivative ∂_x. Indeed, the properties of the binomial coefficients show that for all sufficiently smooth function f, there holds

$$\sum_{j=0}^{J} q_j \left(f((1+2\,j)\,\varepsilon) - f(-(1+2\,j)\,\varepsilon) \right) = \varepsilon\, f'(0) + O(\varepsilon^3).$$

We now consider the Runge-Kutta integration rule of order 3 for the linear system of ordinary differential equations obtained after semi-discretizing the space derivative ∂_x by means of the operator $\widetilde{Q}/\Delta x$ (we recall that λ still denotes the CFL number $\Delta t/\Delta x$)[7]. This procedure gives the fully discretized scheme

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=0}^{3} \frac{\left(-\lambda\, A\, \widetilde{Q}\right)^k}{k!}\, U_j^n\,, & j \in \mathbb{Z}\,,\ n \in \mathbb{N}\,, \\ U_j^0 = f_j\,, & j \in \mathbb{Z}\,. \end{cases} \tag{27}$$

For the scheme (27), we have $p = r = 3\,(1+2\,J)$, and applying Lemma 2.2 again, stability is equivalent to the von Neumann condition. The latter condition is verified by diagonalizing the matrix A. The eigenvalues $z_j(\eta)$ of the amplification matrix $\mathscr{A}(\mathrm{e}^{i\,\eta})$ are given by

$$z_j(\eta) = 1 - \frac{(\lambda\,\lambda_j)^2}{2}\, h(\eta)^2 - i\,\lambda\,\lambda_j\, h(\eta) \left(1 - \frac{(\lambda\,\lambda_j)^2}{6}\, h(\eta)^2 \right),$$

$$h(\eta) := \sum_{j=0}^{J} 2\, q_j\, \sin((2\,j+1)\,\eta)\,. \tag{28}$$

We compute

$$|z_j(\eta)|^2 = 1 - \frac{(\lambda\,\lambda_j)^4}{12}\, h(\eta)^4 \left(1 - \frac{(\lambda\,\lambda_j)^2}{3}\, h(\eta)^2 \right),$$

so stability of (27) is equivalent to the condition $\lambda\,\rho(A)\,\max_{\mathbb{R}} |h| \leq \sqrt{3}$. The main properties of the function h are summarized in Lemma 2.9 below.

Lemma 2.9. *Let the numbers q_j be defined by (26) and let h be defined in (28). Then h is odd and satisfies*

$$\forall\, \eta \in \mathbb{R}\,, \quad h'(\eta) = \cos^{2\,J+1}\eta\,.$$

The function h vanishes exactly for $\eta \in \mathbb{Z}\,\pi$. The maximum of h on $\mathbb{R}$, that we denote M_J, is positive and is attained when $\eta - \pi/2 \in \mathbb{Z}\,2\,\pi$. The minimum of h on $\mathbb{R}$ equals $-M_J$, and is attained when $\eta + \pi/2 \in \mathbb{Z}\,2\,\pi$.

Proof of Lemma 2.9. It is clear that h is odd, and we now differentiate h using the expression (26) of the q_j's, obtaining

$$h'(\eta) = \frac{1}{2^{2\,J}} \sum_{j=0}^{J} C_{2\,J+1}^{J-j} \cos((2\,j+1)\,\eta) = \frac{1}{2^{2\,J}} \sum_{j=0}^{J} C_{2\,J+1}^{j} \cos((2\,J+1-2\,j)\,\eta)$$

$$= \frac{1}{2^{2\,J+1}} \sum_{j=0}^{2\,J+1} C_{2\,J+1}^{j} \cos((2\,J+1-2\,j)\,\eta) = \mathrm{Re} \left(\frac{\mathrm{e}^{i\,\eta} + \mathrm{e}^{-i\,\eta}}{2} \right)^{2\,J+1} = \cos^{2\,J+1}\eta\,,$$

where we have first changed j for $J-j$, and then used the symmetry of the binomial coefficients.

It follows that h behaves exactly as the sine function: h vanishes at 0, is increasing on $[0, \pi/2]$, attains its maximum at $\pi/2$, is decreasing on $[\pi/2, 3\,\pi/2]$ and vanishes at π, attains its minimum at $3\,\pi/2$, and so on. □

[7] We could have used again the Runge-Kutta integration rule of order 4 as in the preceeding example, but we propose this new example to convince the reader that there is a very wide choice of approximation procedures.

Remark 2.3. *The value of M_J in Lemma 2.9 coincides with the Wallis integral $\int_0^{\pi/2} \cos^{2J+1}\eta\,\mathrm{d}\eta$, that is $2^{2J}\,(J!)^2/(2J+1)!$. Since M_J tends to 0 as J tends to $+\infty$, we see that the range of stability $\lambda\,\rho(A) \in [0;\sqrt{3}/M_J]$ for the scheme (27) is getting larger and larger with J going to $+\infty$ (meaning that for large J, the CFL number λ can be chosen large).*

We now analyze the behavior of the curve $\{z_j(\eta)\ \eta \in \mathbb{R}\}$, dealing first with the easier case $\lambda\,|\lambda_j|\,M_J < \sqrt{3}$. We also assume that 0 does not belong to $\mathrm{sp}(A)$ for simplicity. Then $z_j(\eta) \in \mathbb{S}^1$ if and only if $\eta \in \mathbb{Z}\,\pi$, and we compute $z_j(0) = z_j(\pi) = 1$, $z_j'(0) = \overline{z_j'(\pi)} = -i\,\lambda\,\lambda_j \neq 0$. The contact point with the unit circle is a regular point, as can be seen in the right picture of Figure 2 (red curve).

Let us now assume that the CFL number is chosen such that $\lambda\,\lambda_j\,M_J = \sqrt{3}$ (we consider the case $\lambda_j > 0$). Then Lemma 2.9 shows that $z_j(\eta) \in \mathbb{S}^1$ if and only if $\eta \in \mathbb{Z}\,\pi/2$. We still have $z_j(0) = z_j(\pi) = 1$, $z_j'(0) = \overline{z_j'(\pi)} \neq 0$, and we focus from now on on the behavior of z_j near $\eta = \pi/2$. We first compute $z_j(\pi/2) = -1/2 - i\,\sqrt{3}/2$. Using Lemma 2.9, we also have

$$h'(\pi/2) = \cdots = h^{(2J+1)}(\pi/2) = 0\,, \quad h^{(2J+2)}(\pi/2) = -(2J+1)!\,.$$

Performing a Taylor expansion in (28), we obtain

$$z_j(\eta) = -\frac{1}{2} - i\,\frac{\sqrt{3}}{2} + \frac{\lambda\,\lambda_j}{2J+2}\left(\sqrt{3} - \frac{i}{2}\right)(\eta - \pi/2)^{2J+2} + O((\eta-\pi/2)^{2J+3})\,.$$

In particular, $z_j(\pi/2)$ is a singular point of order $2J+2$ and we have

$$|z_j(\eta)| = 1 - \frac{3}{8\,M_J\,(J+1)}\,(\eta - \pi/2)^{2J+2} + o((\eta-\pi/2)^{2J+2})\,.$$

The behavior of the curve parametrized by z_j near $\eta = -\pi/2$ is similar (it is just obtained by a complex conjugation). We refer to the right picture in Figure 2 for a representation of this curve with two singular points of high order[8].

2.5. **Multisteps schemes or how to produce singular points of odd order.** In this paragraph, we are going to construct an example of a scheme of the form (14) with $s = 1$, $r = 3$, $p = 4$, that is stable as long as $\lambda\,\rho(A) \leq 1$, that is geometrically regular and such that in the case $\lambda\,\rho(A) = 1$, one of the parametrized curves associated with eigenvalues of the amplification matrix has a singular contact point of order 3 with $\mathbb{S}^1$. This example is mainly constructed in order to convince the reader that singular contact points of odd order do exist ! However the reader should keep in mind that the scheme defined below is probably useless for practical applications, as was the case for the scheme (27). Its interest is purely theoretical. As it will appear below, it is not so straightforward to construct such an example, or at least we have not found - despite repeated efforts - an easier construction.

We start from (5), semi-discretize the space variable by means of a finite difference operator, leading to the system of ordinary differential equations

$$\dot{v}_j = \frac{1}{\Delta x}\,A\,Q_\sharp\,v_j\,, \quad j \in \mathbb{Z}\,.$$

Then we apply the Adams-Bashforth quadrature rule of order 2. The numerical scheme thus reads

$$\begin{cases} U_j^{n+1} = U_j^n + \lambda\left(\dfrac{3}{2}\,A\,Q_\sharp\,U_j^n - \dfrac{1}{2}\,A\,Q_\sharp\,U_j^{n-1}\right)\,, & j \in \mathbb{Z}\,,\ n \geq 1\,, \\ U_j^0 = f_j^0\,, \quad U_j^1 = f_j^1\,, & j \in \mathbb{Z}\,. \end{cases} \tag{29}$$

We choose the finite difference operator $Q_\sharp$ of the form

$$Q_\sharp := \sum_{\ell=-3}^{4} q_\ell\,\mathbf{T}^\ell\,,$$

[8]Of course, when one only considers the curve and not its parametrization, it is impossible to distinguish between a singular point of order 2 and a singular point of order $2J+2$. The two pictures in Figure 2 look similar even though the right picture represents a more degenerate situation.

where the real numbers $q_{-3},\dots,q_4$ are defined as the unique solution to the linear system

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 \\ 9 & 4 & 1 & 0 & 1 & 4 & 9 & 16 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 3 & -2 & 1 & 0 & -1 & 2 & -3 & 4 \\ -9 & 4 & -1 & 0 & -1 & 4 & -9 & 16 \\ 27 & -8 & 1 & 0 & -1 & 8 & -27 & 64 \\ -81 & 16 & -1 & 0 & -1 & 16 & -81 & 256 \end{pmatrix} \begin{pmatrix} q_{-3} \\ q_{-2} \\ q_{-1} \\ q_0 \\ q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ -1 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix} . \tag{30}$$

The first two rows of the linear system (30) ensure that for all smooth function f, there holds

$$\sum_{\ell=-3}^{4} q_\ell \, f(\ell\,\varepsilon) = -f'(0)\,\varepsilon + o(\varepsilon)\,,$$

so (29) is really an approximation of (5). The reader can easily check either by hand made calculations or on a computer that the matrix of the above linear system is invertible so the scheme (29) is well-defined. The amplification matrix of (29) is given by the formula (16). Diagonalizing A and permuting rows and columns, there exists an invertible matrix T such that for all $\eta\in\mathbb{R}$, there holds

$$\mathscr{A}(\mathrm{e}^{i\,\eta}) = \mathrm{diag}\left(\begin{pmatrix} 1+\dfrac{3\,\lambda\,\lambda_1}{2}\,q(\eta) & -\dfrac{\lambda\,\lambda_1}{2}\,q(\eta) \\ 1 & 0 \end{pmatrix}, \dots, \begin{pmatrix} 1+\dfrac{3\,\lambda\,\lambda_N}{2}\,q(\eta) & -\dfrac{\lambda\,\lambda_N}{2}\,q(\eta) \\ 1 & 0 \end{pmatrix} \right),$$

$$q(\eta) := \sum_{\ell=-3}^{4} q_\ell\,\mathrm{e}^{i\,\ell\,\eta}\,.$$

The function q satisfies

$$\begin{aligned} &q(0)=0\,, \quad q'(0)=-i\,, \quad q''(0)=-1\,, \\ &q(\pi)=-1\,, \quad q'(\pi)=q''(\pi)=0\,, \quad q^{(3)}(\pi)=i\,, \quad q^{(4)}(\pi)=1\,, \end{aligned} \tag{31}$$

as can be checked by using (30).

We now wish to determine the CFL numbers λ for which the scheme (29) is stable. More precisely, we are going to show that if all eigenvalues of A are nonnegative and if $\lambda\,\rho(A)\le 1$, then the operators in (29) are geometrically regular and the amplification matrix of (29) satisfies the von Neumann condition. This will enable us to apply Proposition 2.3 and deduce stability for (29). We shall need the following preliminary result.

Lemma 2.10. *The mapping*

$$\kappa\in\mathbb{S}^1 \longmapsto \frac{2\,\kappa\,(\kappa-1)}{3\,\kappa-1}\,,$$

is injective and thus defines a closed simple curve $\mathscr{C}\subset\mathbb{C}\simeq\mathbb{R}^2$. The interior $\mathscr{I}$ of $\mathscr{C}$ is a strictly convex region that contains the segment $]-1,0[$. Moreover, 1 belongs to the exterior of $\mathscr{C}$.

Proof of Lemma 2.10. We consider the mapping

$$\theta\in[-\pi,\pi] \longmapsto \frac{2\,\mathrm{e}^{i\,\theta}\,(\mathrm{e}^{i\,\theta}-1)}{3\,\mathrm{e}^{i\,\theta}-1} = x(\theta)+i\,y(\theta)\,.$$

Direct computations yield $y(0)=y(\pm\pi)=0$, and $\pm y(\theta)>0$ if $\pm\theta\in\,]0,\pi[$. Furthermore, x is increasing on $[-\pi,0]$ and decreasing on $[0,\pi]$. These properties imply that $\mathscr{C}$ is a simple closed curve (see Figure 3 for a representation of $\mathscr{C}$). The reader can also check that $(x')^2+(y')^2$ does not vanish so every point of $\mathscr{C}$ is regular.

The interior of $\mathscr{C}$ is well-defined thanks to Jordan's Theorem. It is strictly convex provided that the curvature of $\mathscr{C}$ is nonnegative and vanishes at finitely many points. This amounts to proving that $x'\,y''-x''\,y'$ is nonnegative and vanishes at finitely many points. We compute

$$x'(\theta)\,y''(\theta)-x''(\theta)\,y'(\theta) = \frac{6\,(1-X)\,(3\,X^2-3\,X+4)}{(5-3\,X)^3}\Big|_{X=\cos\theta} \ge 0\,,$$

so $\mathscr{I}$ is strictly convex. □

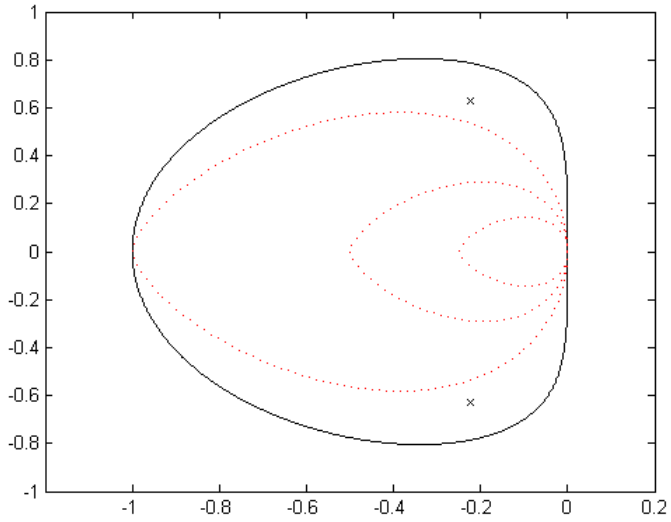

FIGURE 3. The curve $\mathscr{C}$ (black line), and the curve $\{\lambda\,\lambda_j\,q(\eta)\,,\eta\in\mathbb{R}\}$ for $\lambda\,\lambda_j = 1/4$, $\lambda\,\lambda_j = 1/2$ and $\lambda\,\lambda_j = 1$ (red dots). The crosses represent the points $-2/9 \pm i\,4\,\sqrt{2}/9$.

The following Lemma explains the link between the curve $\mathscr{C}$ and stability of the scheme (29).

Lemma 2.11. *Let us assume that for all $\eta \notin \mathbb{Z}\,\pi$, $q(\eta) \in \mathscr{I}$, where the region $\mathscr{I}$ is defined in Lemma 2.10. If all eigenvalues of A are nonnegative and if furthermore $\lambda\,\rho(A) \leq 1$, then the scheme (29) is stable.*

Proof of Lemma 2.11. • Let us start with the following simple observations. The matrix

$$M(\alpha) := \begin{pmatrix} 1+3\,\alpha/2 & -\alpha/2 \\ 1 & 0 \end{pmatrix}$$

has at least one eigenvalue in $\mathbb{S}^1$ if and only if $\alpha \in \mathscr{C}$. By a connectedness argument, this means that for $\alpha \in \mathscr{I}$, $M(\alpha)$ has two eigenvalues in $\mathbb{D}$ (just look at the case $\alpha = -1/2$). If α belongs to the exterior of $\mathscr{C}$, then $M(\alpha)$ has one eigenvalue in $\mathbb{D}$ and one eigenvalue in $\mathscr{U}$ (look at the case $\alpha = 1$). Moreover, $M(\alpha)$ has a double eigenvalue if and only if $\alpha = -2/9 \pm i\,4\,\sqrt{2}/9$, and in that case the double root belongs to $\mathbb{D}$. If $\alpha \in \mathscr{C}$, then $M(\alpha)$ can not have two distinct eigenvalues on $\mathbb{S}^1$ (use Lemma 2.10) so $M(\alpha)$ has exactly one eigenvalue in $\mathbb{D}$ and one eigenvalue on $\mathbb{S}^1$. If we summarize, the eigenvalues of $M(\alpha)$ belong to the closed unit disk provided that α belongs to $\mathscr{I} \cup \mathscr{C}$.

• According to the reduction of the amplification matrix, the von Neumann condition will be satisfied if for all eigenvalue λ_j of A and for all $\eta \in \mathbb{R}$, the eigenvalues of $M(\lambda\,\lambda_j\,q(\eta))$ belong to the closed unit disk. We compute $q(0) = 0 \in \mathscr{C}$ and $q(\pi) = -1 \in \mathscr{C}$, so for all $\eta \in \mathbb{R}$, there holds $q(\eta) \in \mathscr{I} \cup \mathscr{C}$ thanks to the assumption of Lemma 2.11. The convexity of $\mathscr{I}$ shows that under the CFL condition $\lambda\,\rho(A) \leq 1$, there holds $\lambda\,\lambda_j\,q(\eta) \in \mathscr{I} \cup \mathscr{C}$. (Here we have used the fact that eigenvalues of A are nonnegative.) Using the above observations, we conclude that the eigenvalues of the matrix $M(\lambda\,\lambda_j\,q(\eta))$ belong to the closed unit disk. Consequently the von Neumann condition is satisfied.

• It remains to show that the amplification matrix satisfies the geometric regularity condition stated in Definition 2.3 and we shall be able to apply Proposition 2.3 to conclude. Using the diagonalization of $\mathscr{A}(\mathrm{e}^{i\,\eta})$ in blocks of the form $M(\lambda\,\lambda_j\,q(\eta))$, we already see that it is sufficient to prove a geometric regularity condition on each 2×2 block. Moreover, the exponential is locally

a biholomorphic diffeomorphism so working in a complex neighborhood of some $\underline{\kappa} = \mathrm{e}^{i\,\underline{\eta}} \in \mathbb{S}^1$ is equivalent to working in a complex neighborhood of $\underline{\eta} \in \mathbb{R}$.

Let us first consider the case $\lambda\,\lambda_j < 1$. The strict convexity of $\mathscr{I}$ shows that $\lambda\,\lambda_j\,q(\eta) \in \mathscr{C}$ if and only if $q(\eta) \in \mathbb{Z}\,2\,\pi$. For $\eta = 0$, the eigenvalues of $M(0)$ are 0 and 1, so 1 is a simple hence geometrically regular eigenvalue of $M(\lambda\,\lambda_j\,q(\eta))$. If we consider the case $\lambda\,\lambda_j = 1$, we have $\lambda\,\lambda_j\,q(\eta) \in \mathscr{C}$ if and only if $q(\eta) \in \mathbb{Z}\,\pi$. For $\eta = \pi$, the eigenvalues of $M(-1)$ are -1 and $1/2$ so -1 is also a simple hence geometrically regular eigenvalue of $M(\lambda\,\lambda_j\,q(\eta))$. The proof of Lemma 2.11 is complete. □

Figure 3 gives some numerical evidence that the curve $\{q(\eta)\,,\ \eta \in \mathbb{R}\}$ remains within the interior of $\mathscr{C}$. However, we must confess that we have not been able (or not brave enough) to find a complete proof of this fact. As such, stability of (29) under the appropriate CFL condition remains an "if result".

Let us focus on the behavior of the eigenvalues of the block $M(q(\eta))$, assuming that $\lambda\,\lambda_j = 1$. As we have seen in the proof of Lemma 2.11, $M(q(\eta))$ has an eigenvalue on $\mathbb{S}^1$ if and only if $\eta \in \mathbb{Z}\,\pi$. If $\eta = 0$, 1 is a simple eigenvalue whose Taylor expansion near $\eta = 0$ reads (use the relations (31))

$$z(\eta) = 1 - i\,\eta - \eta^2 + o(\eta^2)\,, \quad |z(\eta)| = 1 - \frac{1}{2}\,\eta^2 + o(\eta^2)\,.$$

If $\eta = \pi$, -1 is a simple eigenvalue whose Taylor expansion near $\eta = \pi$ reads (use the relations (31) again)

$$z(\eta) = -1 + \frac{2\,i}{9}\,(\eta-\pi)^3 + \frac{1}{18}\,(\eta-\pi)^4 + o((\eta-\pi)^4)\,, \quad |z(\eta)| = 1 - \frac{1}{18}\,(\eta-\pi)^4 + o((\eta-\pi)^4)\,.$$

In particular, the above Taylor expansions show that for all $\eta \neq 0$ sufficiently small and for all $\eta \neq \pi$ sufficiently close to π, the eigenvalues of $M(q(\eta))$ belong to $\mathbb{D}$. Furthermore, the eigenvalue curve passing through -1 has a singular contact point of order 3. We refer to Figure 4 for a representation of the spectrum of $M(q(\eta))$, that is for the case $\lambda\,\lambda_j = 1$.

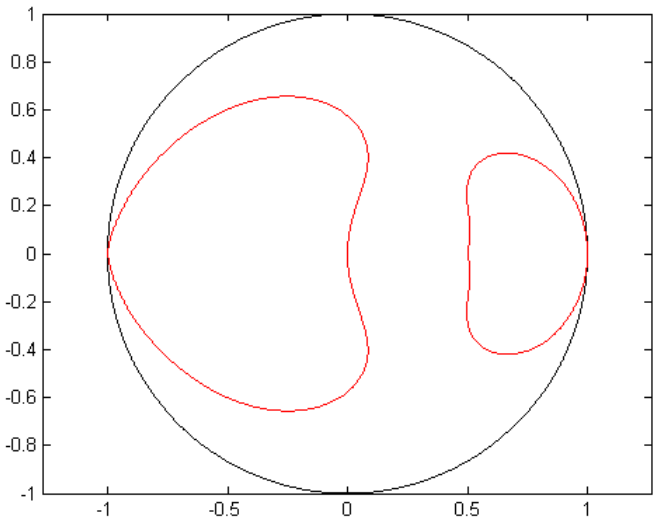

FIGURE 4. The eigenvalues of $M(q(\eta))$ in red and the unit circle in black.

2.6. **A few facts to remember in view of what follows, and a (not very interesting) conjecture.** We try to summarize here a few facts that should be kept in mind since they will play an important role in the following Section. The von Neumann condition is only a necessary sufficient condition for stability. However, in one space dimension, the geometric regularity condition for the amplification matrix is more or less always satisfied. This is rather clear for one step schemes ($s = 0$)

since usually the matrices A_ℓ can be simultaneously diagonalized. For multistep schemes as the leap-frog scheme, this is a little less obvious but it can usually be checked by rather simple arguments. The main advantage of the geometric regularity property is that it characterizes stability by means of the von Neumann condition, thus ruling out pathological Jordan blocks.

The main difference between the theory of hyperbolic partial differential equations and their discrete counterparts lies in the behavior of eigenvalues of the amplification matrix. For the continuous Cauchy problem, one passes from $\widehat{u}(0,\xi)$ to $\widehat{u}(\Delta t,\xi)$ through a multiplication by the matrix $\exp(-i\,\Delta t\,\xi\,A)$, see (7). In particular, the eigenvalues of the *exact* amplification matrix belong to $\mathbb{S}^1$ for all frequency ξ. On the Fourier side, this property shows that modes associated with any frequency are not damped so that the L^2 norm of the solution is conserved (at least up to an appropriate change of unknown that diagonalizes A). At the discrete level, the eigenvalues of the amplification matrix are not necessarily located on $\mathbb{S}^1$ since they can also belong to $\mathbb{D}$. Eigenvalues in $\mathbb{D}$ correspond to an exponential damping as what happens more or less for parabolic equations. In order to have the lowest dissipation, the eigenvalues of the amplification matrix should remain as close as possible to $\mathbb{S}^1$ (compare for instance the two pictures in Figure 2 and guess which scheme dissipates less).

What makes the situation for discrete problems far more complicated is that for high frequencies, eigenvalues of the amplification matrix may approach $\mathbb{S}^1$ again. (For consistent schemes, there is always a group of eigenvalues located at 1 for $\eta = 0$.) This may give rise to singular contact points that are analogous to glancing frequencies in the theory of partial differential equations. Here, such glancing frequencies are also associated to some kind of dissipation phenomenon. We refer to [25] for some more details on this issue. The previous examples were given in order to show that many possible singular contact points may appear. As a matter of fact, our conjecture is the following: if we consider for simplicity the scalar transport equation

$$\partial_t u + \partial_x u = 0\,, \quad u(0,x) = f(x)\,,$$

and if we consider two integers $m_1 \in \mathbb{N}$, $m_2 \in \mathbb{N}$ such that $m_1 > 0$ and $2\,m_2 \geq m_1$, then there exists a stable and geometrically regular numerical scheme such that there exists one eigenvalue curve defined in the neighborhood of some $\underline{\eta} \in \mathbb{R}$ and satisfying

$$z(\eta) = \underline{z} + \alpha\,(\eta - \underline{\eta})^{m_1} + o((\eta - \underline{\eta})^{m_1})\,, \quad |z(\eta)| = 1 - c\,(\eta - \underline{\eta})^{2\,m_2} + o((\eta - \underline{\eta})^{2\,m_2})\,,$$

where $\underline{z} \in \mathbb{S}^1$, $\alpha \in \mathbb{C} \setminus \{0\}$ and $c > 0$. Of course, the numerical scheme should also be consistent with the transport equation in order to be meaningful. The examples above show that the conjecture is true at least for $m_1 = 2\,m_2$ as well as for $m_1 = 3$ and $m_2 = 2$. We do not think however that this conjecture is really meaningful from a mathematical point of view. Our message is the following: if we wish to develop a general stability theory that covers all "reasonable" numerical schemes in one space dimension, then geometric regularity is not a strong assumption but the price to pay is the appearance of infinitely many possible singular contact points with $\mathbb{S}^1$ corresponding to the above Taylor expansions. Such glancing/dissipative frequencies do not appear in the analogous theory for partial differential equations, see for instance [2, chapter 4].

3. Fully discrete initial boundary value problems: stability with zero initial data

3.1. **Finite difference discretizations and strong stability.** From now on, we consider the continuous problem (1) which we discretize by means of a finite difference scheme. Let us assume that we have already chosen one discretization of the hyperbolic operator, as in Section 2, and that this scheme involves r points on the left and p points on the right, see (9) or (15). Here the space grid is not indexed by $\mathbb{Z}$ anylonger since we consider a problem on a half-line. Up to using a translation on the indeces, we can always assume that the space grid is indexed by $\{j \in \mathbb{Z}\,,\, j \geq 1-r\}$. This means that the solution u to (1) is approximated by a sequence (U_j^n) defined for $j \geq 1-r$ and $n \geq 0$. If the initial condition $(U_j^0)_{j\geq 1-r}$ is known, then we can not apply the discretization of the hyperbolic operator at points $j = 1-r, \dots, 0$ because this would require using some values U_ℓ^0 with $\ell \leq -r$. Consequently, a discretization of (1) must involve (i) one discretization of the hyperbolic operator to be used at the grid points $j \geq 1$, and (ii) one way to discretize the boundary conditions to be used at the grid points $j = 1-r, \dots, 0$. As we have already seen in Section 2, there are many possible choices for discretizing the hyperbolic operator and the reader will no doubt imagine that there is also a wide choice of possibilities for discretizing the boundary conditions. We do not aim here at considering the most general schemes but we shall try nevertheless to encompass a wide class of discretizations, both in terms of the hyperbolic operator and in terms of the boundary conditions. Some rather simple examples are given in the following Section. More examples may be found in [10] and [9, chapters 11, 13] as well as in the references cited therein. In the examples that we shall detail in these notes, we shall see that discretizing the boundary conditions is not especially difficult in one space dimension since one can then separate incoming from outgoing characteristics. Achieving high order approximation together with stability is however more delicate.

After this short introduction, let us now introduce the finite difference approximation of (1). We let $\Delta x, \Delta t > 0$ denote a space and a time step where $\lambda = \Delta t/\Delta x$ is a fixed positive constant, and we also let p, q, r, s denote some fixed integers. The solution to (1) is approximated by a sequence (U_j^n) defined for $n \in \mathbb{N}$, and $j \in 1-r+\mathbb{N}$. For $j = 1-r, \dots, 0$, the vector U_j^n should be understood as an approximation of the trace $u(n\,\Delta t, 0)$ on the boundary $\{x = 0\}$, and possibly the trace of normal derivatives. For instance, in the case $r = 1$, there is one grid point in the *discrete boundary*, and U_0^n is an approximate value of $u(n\,\Delta t, 0)$. In the case $r = 2$, there are two grid points in the discrete boundary: U_0^n is still an approximate value of $u(n\,\Delta t, 0)$ and the scheme can be built in such a way that $(U_0^n - U_{-1}^n)/\Delta x$ is an approximation of $\partial_x u(n\,\Delta t, 0)$. In some sense, the integer r can give a measure of the order of approximation at the boundary. (It is rather clear that with only one grid point in the discrete boundary, one will hardly reach an approximation of order 10...) The boundary meshes $[j\,\Delta x, (j+1)\,\Delta x[$, $j = 1-r, \dots, 0$, shrink to $\{0\}$ as Δx tends to 0, so the formal continuous limit problem as Δx tends to 0 is set on the half-line $\mathbb{R}^+$. In these notes, we consider finite difference approximations of (1) that read

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} Q_\sigma\, U_j^{n-\sigma} + \Delta t\, F_j^n\,, & j \geq 1\,, \quad n \geq s\,, \\ U_j^{n+1} = \displaystyle\sum_{\sigma=-1}^{s} B_{j,\sigma}\, U_1^{n-\sigma} + g_j^{n+1}\,, & j = 1-r, \dots, 0\,, \quad n \geq s\,, \\ U_j^n = f_j^n\,, & j \geq 1-r\,, \quad n = 0, \dots, s\,, \end{cases} \tag{32}$$

where the operators Q_σ and $B_{j,\sigma}$ are given by:

$$Q_\sigma := \sum_{\ell=-r}^{p} A_{\ell,\sigma}\, \mathbf{T}^\ell\,, \quad B_{j,\sigma} := \sum_{\ell=0}^{q} B_{\ell,j,\sigma}\, \mathbf{T}^\ell\,. \tag{33}$$

In (33), all matrices $A_{\ell,\sigma}, B_{\ell,j,\sigma}$ belong to $\mathscr{M}_N(\mathbb{R})$ and are independent of the small parameter Δt, while $\mathbf{T}$ still denotes the shift operator as in Section 2. Let us emphasize that we deal here with explicit schemes for simplicity. If the solution is known up to the time index $n \geq s$, then the scheme first determines U_j^{n+1} for $j \geq 1$ by applying the discretization of the hyperbolic operator. Then the scheme determines the values $U_{1-r}^{n+1}, \dots, U_0^{n+1}$ by applying the operators $B_{j,\sigma}$. We believe that most of the arguments below can be adapted to some implicit discretizations as in [10].

In Section 2, we have studied the stability of fully discrete hyperbolic equations on the whole real line. Stability for a numerical scheme had been defined in order to reproduce the energy estimate (6) that was known to hold for the continuous problem. The definition of stability for (32) follows the same approach, except that here we wish to study the sensitivity of the solution with respect to three possible source terms: the interior source term (F_j^n), the boundary source term (g_j^n) and the initial data $f^0, \dots, f^s$. We shall in some sense cut the problems into two pieces and deal first with the case of zero initial data. Nonzero initial data will be considered in Section 5. For zero initial data, an appropriate notion of stability was introduced in [10]:

Definition 3.1 (Strong stability [10]). *The finite difference approximation (32) is said to be strongly stable if there exists a constant C_0 such that for all $\gamma > 0$ and all $\Delta t \in]0,1]$, the solution (U_j^n) of (32) with $f^0 = \cdots = f^s = 0$ satisfies the estimate:*

$$\frac{\gamma}{\gamma\,\Delta t + 1} \sum_{n \geq s+1} \sum_{j \geq 1-r} \Delta t\, \Delta x\, \mathrm{e}^{-2\,\gamma\, n\, \Delta t}\, |U_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{p} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\, \Delta t}\, |U_j^n|^2$$
$$\leq C_0 \left\{ \frac{\gamma\,\Delta t + 1}{\gamma} \sum_{n \geq s} \sum_{j \geq 1} \Delta t\, \Delta x\, \mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\, |F_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\, \Delta t}\, |g_j^n|^2 \right\}. \quad (34)$$

In Definition (3.1), the stability estimate (34) should be understood as follows: if the source terms $(F_j^n), (g_j^n)$ are such that the right hand-side of the inequality is finite, then the solution (U_j^n) should satisfy the latter inequality and the constant C_0 is independent of $\gamma > 0$ and $\Delta t \in]0,1]$. If the source terms are such that the right hand-side of the inequality is infinite, then (32) still uniquely defines a sequence (U_j^n) but we do not require this solution to satisfy anything. The terminology "strong stability" is used to emphasize that the solution is estimated in the same norm as the data. Here there are an interior source term and a boundary source term so the natural requirement is to ask for a control of U in the interior domain and a control of the "trace" of U. To be completely honest, we should warn the reader that Definition 3.1 above is not exactly the notion of strong stability introduced in [10]. The difference is the following. In [10], the authors considered in the left-hand side of the inequality the term

$$\sum_{n \geq s+1} \sum_{j=1-r}^{0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\, \Delta t}\, |U_j^n|^2\,,$$

in order to estimate the trace of the solution (U_j^n) while here we make the sum run from $1-r$ to p. This modification is motivated by the results of the following paragraphs where we wish to characterize - as easily as possible - strong stability by means of an estimate for the so-called resolvent equation. Such a characterization is easily proved if we consider this slightly stronger notion of stability, while we have not been able to fill the gap in [10] with their weaker notion. But this does not so much matter since we show a better property on the solution that what appeared in [10].

There are two ways to remember the stability estimate of Definition (3.1), and to understand why the various weights involving γ and Δt are meaningful. Studying first the limit $\Delta t \to 0$, we should recover formally an energy estimate for the continuous problem (1). Indeed, if we let formally Δt tend to 0, assuming that all quantities have a limit, we obtain

$$\gamma \iint_{\mathbb{R}^+ \times \mathbb{R}^+} \mathrm{e}^{-2\,\gamma\, t} |u(t,x)|^2\, \mathrm{d}t\, \mathrm{d}x + \int_{\mathbb{R}^+} \mathrm{e}^{-2\,\gamma\, t} |u(t,0)|^2\, \mathrm{d}t$$
$$\leq C_0 \left\{ \frac{1}{\gamma} \iint_{\mathbb{R}^+ \times \mathbb{R}^+} \mathrm{e}^{-2\,\gamma\, t} |F(t,x)|^2\, \mathrm{d}t\, \mathrm{d}x + \int_{\mathbb{R}^+} \mathrm{e}^{-2\,\gamma\, t} |g(t)|^2\, \mathrm{d}t \right\}.$$

The latter energy estimate is known to hold for solutions of (1) with zero initial data as soon as the well-posedness condition (4) holds. This can be checked by using the formulae (2), (3). Of course, the above limit is completely formal since there is already some problem with the size of the source terms on the boundary: in (32), the vectors g_j^n belong to $\mathbb{R}^N$ while, for the continuous problem (1), $g(t)$ belongs to $\mathbb{R}^p$, and in general p is strictly smaller than N. However, the above formal limit

shows the link between the energy estimate for (1) and the stability estimate (34) of Definition 3.1. We also note that in the first sum on the left-hand side of (34), the factor $\Delta t\,\Delta x$ is the measure of the mesh $[n\,\Delta t,(n+1)\,\Delta t[\times[j\,\Delta x,(j+1)\,\Delta x[$ so the sum represents an L^2 norm in the variables (t,x) of a piecewise constant function. All other sums in (34) represent L^2 norms in t or in (t,x) as well.

Another interesting observation is to consider the limit $\gamma\to+\infty$ in (34). At a formal level, the term $\exp(-2\,\gamma\,m\,\Delta t)$ is negligible with respect to $\exp(-2\,\gamma\,n\,\Delta t)$ for $m>n$. Multiplying (34) by $\exp(2\,\gamma\,(s+1)\,\Delta t)$ and letting γ tend to $+\infty$ (recall that the initial data vanish), the scheme (32) should verify

$$\frac{1}{\Delta t}\sum_{j\geq 1-r}\Delta t\,\Delta x\,|U_j^{s+1}|^2+\sum_{j=1-r}^{p}\Delta t\,|U_j^{s+1}|^2\leq C_0\left\{\Delta t\sum_{j\geq 1}\Delta t\,\Delta x\,|F_j^s|^2+\sum_{j=1-r}^{0}\Delta t\,|g_j^{s+1}|^2\right\},$$

or equivalently

$$\frac{1}{\lambda}\sum_{j\geq 1-r}|U_j^{s+1}|^2+\sum_{j=1-r}^{p}|U_j^{s+1}|^2\leq C_0\left\{\frac{1}{\lambda}\,\Delta t^2\sum_{j\geq 1}|F_j^s|^2+\sum_{j=1-r}^{0}|g_j^{s+1}|^2\right\}.$$

Such an estimate can be easily deduced from (32) with $U^0=\cdots=U^s=0$.

Remark 3.1. *The estimate (34) can be made independent of Δt by simply observing that in (32), the small parameter Δt appears only in the source term $\Delta t\,F_j^n$. One easily sees that strong stability for (32) is equivalent to requiring that the solution (U_j^n) to*

$$\begin{cases} U_j^{n+1}=\displaystyle\sum_{\sigma=0}^{s}Q_\sigma\,U_j^{n-\sigma}+F_j^n\,, & j\geq 1\,,\quad n\geq s\,,\\ U_j^{n+1}=\displaystyle\sum_{\sigma=-1}^{s}B_{j,\sigma}\,U_1^{n-\sigma}+g_j^{n+1}\,, & j=1-r,\dots,0\,,\quad n\geq s\,,\\ U_j^n=0\,, & j\geq 1-r\,,\quad n=0,\dots,s\,, \end{cases} \tag{35}$$

satisfies the estimate

$$\begin{aligned}\frac{\gamma}{\gamma+1}\sum_{n\geq s+1}\sum_{j\geq 1-r}\mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2+\sum_{n\geq s+1}\sum_{j=1-r}^{p}\mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2\\ \leq C\left\{\frac{\gamma+1}{\gamma}\sum_{n\geq s}\sum_{j\geq 1}\mathrm{e}^{-2\,\gamma\,(n+1)}\,|F_j^n|^2+\sum_{n\geq s+1}\sum_{j=1-r}^{0}\mathrm{e}^{-2\,\gamma\,n}\,|g_j^n|^2\right\},\end{aligned} \tag{36}$$

for all $\gamma>0$ and a constant C that is independent of γ. In other words, one can always assume $\Delta t=1$ (and $\Delta x=1/\lambda$) when checking strong stability.

In the following paragraph, we shall explain how strong stability can be characterized by means of an estimate for the so-called *resolvent equation*. This characterization relies on the Laplace transform. The strategy is entirely analogous to the analysis for the continuous problem, see [2, chapter 4].

3.2. **The nomal modes analysis and the Godunov-Ryabenkii condition.** The resolvent equation is obtained by formally looking for solutions to (32) of the form $U_j^n=z^n\,W_j$, $z\in\mathbb{C}\setminus\{0\}$. The source terms in (32) should have similar expressions. Of course, this is a formal procedure since such sequences do not satisfy $U^0=\cdots=U^s=0$, while we are restricting here to the case of zero initial data! Solutions to (32) should be thought of as superpositions of such elementary solutions that we call *normal modes* (this is the same strategy as in Section 2 when we performed some plane wave analysis by looking for solutions to (8) under the form of pure oscillations). Plugging the expression

$U_j^n = z^n\, W_j$ in (32) yields a system of the form

$$\begin{cases} W_j - \displaystyle\sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, W_j = F_j\,, \qquad j \geq 1\,, \\ W_j - \displaystyle\sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{j,\sigma}\, W_1 = g_j\,, \quad j = 1-r, \dots, 0\,, \end{cases} \tag{37}$$

where we do not wish to make the expression of the source terms precise since it would be useless. Our goal here is to characterize strong stability for the scheme (32) in terms of an estimate for the solution to the resolvent equation (37). The main advantage for doing so is that studying (37) has reduced the dimension since time has been replaced by one complex parameter. For clarity, we shall divide some of the arguments below into several intermediate results. The main results are summarized at the end of this paragraph for future use. Our first main result is

Theorem 3.1 (Gustafsson, Kreiss, Sundström [10]). *Assume that the scheme (32) is strongly stable in the sense of Definition 3.1 with a constant $C_0 > 0$ such that (36) holds. Then for all $z \in \mathscr{U}$, for all $(F_j) \in \ell^2$ and for all vectors $g_{1-r}, \dots, g_0 \in \mathbb{C}^N$, the resolvent equation (37) has a unique solution $(W_j) \in \ell^2$ and this solution satisfies*

$$\frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 \leq 4\, C_0 \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}. \tag{38}$$

The proof of Theorem 3.1 relies on two preliminary results, which we prove now.

Lemma 3.1. *For all $x > 0$, there holds*

$$\frac{x}{1+x} \leq \frac{\mathrm{e}^x - 1}{\mathrm{e}^x} \leq 2\, \frac{x}{1+x}\,,$$

or equivalently

$$\frac{1}{2}\, \frac{1+x}{x} \leq \frac{\mathrm{e}^x}{\mathrm{e}^x - 1} \leq \frac{1+x}{x}\,.$$

Proof of Lemma 3.1. The inequality

$$\frac{x}{1+x} \leq \frac{\mathrm{e}^x - 1}{\mathrm{e}^x}\,,$$

is easily seen to be equivalent to $\mathrm{e}^x \geq 1 + x$, and this inequality follows from the power series expansion of the exponential function.

On the other hand, the inequality

$$\frac{\mathrm{e}^x - 1}{\mathrm{e}^x} \leq 2\, \frac{x}{1+x}\,,$$

is equivalent to $(x-1)\,\mathrm{e}^x + x + 1 \geq 0$. The latter function of x vanishes at 0 and is increasing on $\mathbb{R}^+$ so the result follows. □

Lemma 3.2. *For all $\nu \geq 1$, we define the function ρ_ν on $\mathbb{R}$ by*

$$\forall\, \theta \in \mathbb{R}\,, \quad \rho_\nu(\theta) := \frac{1}{\sqrt{\nu}} \sum_{k=0}^{\nu-1} \mathrm{e}^{-i\,k\,\theta}\,.$$

Then the sequence (ρ_ν) satisfies the following properties:

(i) *For all $\nu \geq 1$, ρ_ν is $2\,\pi$-periodic and*

$$\frac{1}{2\,\pi} \int_{-\pi}^{\pi} |\rho_\nu(\theta)|^2\, \mathrm{d}\theta = 1\,.$$

(ii) *For all $\alpha \in\,]0, \pi/2]$, there holds*

$$\lim_{\nu \to +\infty} \int_{\alpha}^{\pi} |\rho_\nu(\theta)|^2\, \mathrm{d}\theta = 0\,.$$

(iii) *For all continuous function H on $\mathbb{R}$ verifying $\sup_{\theta\in\mathbb{R}}(1+\theta^2)\,|H(\theta)|<+\infty$, there holds*

$$\lim_{\nu\to+\infty}\frac{1}{2\,\pi}\int_{\mathbb{R}}H(\theta)\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta=\sum_{k\in\mathbb{Z}}H(2\,k\,\pi)\,.$$

Proof of Lemma 3.2. • The proof of property (i) follows from a straightforward computation:

$$\int_{-\pi}^{\pi}|\rho_\nu(\theta)|^2\,\mathrm{d}\theta=\frac{1}{\nu}\sum_{k_1,k_2=0}^{\nu-1}\int_{-\pi}^{\pi}\mathrm{e}^{i\,(k_1-k_2)\,\theta}\,\mathrm{d}\theta=2\,\pi\,.$$

• For $\alpha\in]0,\pi/2]$ and $\theta\in[\alpha,\pi]$, we have

$$|\rho_\nu(\theta)|=\frac{1}{\sqrt{\nu}}\left|\frac{1-\mathrm{e}^{-i\,\nu\,\theta}}{1-\mathrm{e}^{-i\,\theta}}\right|\le\frac{2}{\sqrt{\nu}\,\sqrt{(1-\cos\theta)^2+\sin^2\theta}}\le\begin{cases}2/\sqrt{\nu} & \text{if }\theta\in[\pi/2,\pi],\\ 2/(\sqrt{\nu}\,\sin\alpha) & \text{if }\theta\in[\alpha,\pi/2].\end{cases}$$

Property (ii) follows by integrating the latter inequalities.

• Let us first observe that both the integral on $\mathbb{R}$ and the sum over $\mathbb{Z}$ in property (iii) are well-defined thanks to the assumption on H. Moreover, property (i) gives

$$A_\nu:=\left|\frac{1}{2\,\pi}\int_{\mathbb{R}}H(\theta)\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta-\sum_{k\in\mathbb{Z}}H(2\,k\,\pi)\right|\le\frac{1}{2\,\pi}\sum_{k\in\mathbb{Z}}\int_{(2\,k-1)\,\pi}^{(2\,k+1)\,\pi}|H(\theta)-H(2\,k\,\pi)|\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\,.$$

Our goal is to show that the sequence $(A_\nu)_{\nu\ge1}$ converges towards 0. Let therefore $\varepsilon>0$. We first note that there exists an integer K_ε, that is independent of ν, such that

$$\forall\,\nu\ge1\,,\quad\frac{1}{2\,\pi}\sum_{|k|>K_\varepsilon}\int_{(2\,k-1)\,\pi}^{(2\,k+1)\,\pi}|H(\theta)-H(2\,k\,\pi)|\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\le\varepsilon\,.$$

Indeed the assumption on H yields, for a suitable constant M that depends on H but not on ν,

$$\begin{aligned}\frac{1}{2\,\pi}\sum_{|k|>K}\int_{(2\,k-1)\,\pi}^{(2\,k+1)\,\pi}|H(\theta)-H(2\,k\,\pi)|\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta&\le\frac{1}{2\,\pi}\sum_{|k|>K}\int_{(2\,k-1)\,\pi}^{(2\,k+1)\,\pi}\frac{M}{1+k^2}\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\\&=M\sum_{|k|>K}\frac{1}{1+k^2}\,.\end{aligned}$$

The right-hand side of the latter inequality is small provided that K is large, independently of ν. We thus have

$$A_\nu\le\varepsilon+\frac{1}{2\,\pi}\sum_{|k|\le K_\varepsilon}\int_{(2\,k-1)\,\pi}^{(2\,k+1)\,\pi}|H(\theta)-H(2\,k\,\pi)|\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\,.$$

The continuity of H at the points $2\,k\,\pi$, $|k|\le K_\varepsilon$, gives the existence of some $\alpha\in]0,\pi/2]$, that is independent of ν, such that

$$\forall\,k=-K_\varepsilon,\dots,K_\varepsilon\,,\quad\forall\,\theta\in[2\,k\,\pi-\alpha,2\,k\,\pi+\alpha]\,,\quad|H(\theta)-H(2\,k\,\pi)|\le\frac{\varepsilon}{2\,K_\varepsilon+1}\,.$$

For all $\nu\ge1$, we thus have

$$\begin{aligned}A_\nu&\le2\,\varepsilon+\frac{1}{2\,\pi}\sum_{|k|\le K_\varepsilon}\int_{(2\,k-1)\,\pi}^{2\,k\,\pi-\alpha}|H(\theta)-H(2\,k\,\pi)|\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\\&\quad+\frac{1}{2\,\pi}\sum_{|k|\le K_\varepsilon}\int_{2\,k\,\pi+\alpha}^{(2\,k+1)\,\pi}|H(\theta)-H(2\,k\,\pi)|\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\\&\le2\,\varepsilon+\frac{4\,\|H\|_{L^\infty(\mathbb{R})}}{2\,\pi}\,(2\,K_\varepsilon+1)\int_{\alpha}^{\pi}|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\,,\end{aligned}$$

where we have used property (i) for the integrals on $[2\,k\,\pi-\alpha,2\,k\,\pi+\alpha]$ and the fact that $|\rho_\nu|$ is even. Using property (ii), we can complete the proof of property (iii) because we have obtained $A_\nu\le3\,\varepsilon$ for ν sufficiently large. □

Let us now prove Theorem 3.1.

Proof of Theorem 3.1. Before proving that the resolvent equation (37) has a unique solution for all data in ℓ^2, we shall prove an a priori estimate for any solution to (37). In other words, we shall consider that we already have a solution to the resolvent equation and we wish to prove that this solution satisfies the estimate (38). We introduce the notation

$$\forall\, z \in \mathscr{U}\,, \quad L(z)\,:\, W \in \ell^2 \longmapsto L(z)\,W \in \ell^2$$
$$\text{with } (L(z)\,W)_j := \begin{cases} W_j - \sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, W_j\,, & \text{if } j \geq 1, \\ W_j - \sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{j,\sigma}\, W_1\,, & \text{if } 1-r \leq j \leq 0. \end{cases} \tag{39}$$

Let now $(W_j)_{j\geq 1-r} \in \ell^2$, and let $z_0 \in \mathscr{U}$. For all integer $\nu \geq 1$, we define the sequences

$$\forall\, j \geq 1-r\,, \quad \forall\, n \geq 0\,, \quad U_j^n(\nu) := \begin{cases} z_0^n\, W_j/\sqrt{\nu}\,, & \text{if } s+1 \leq n \leq s+\nu, \\ 0\,, & \text{otherwise.} \end{cases}$$
$$\forall\, j \geq 1\,, \quad \forall\, n \geq s\,, \quad F_j^n(\nu) := U_j^{n+1}(\nu) - \sum_{\sigma=0}^{s} Q_\sigma\, U_j^{n-\sigma}(\nu)\,,$$
$$\forall\, j = 1-r, \dots, 0\,, \quad \forall\, n \geq s+1\,, \quad g_j^n(\nu) := U_j^n(\nu) - \sum_{\sigma=-1}^{s} B_{j,\sigma}\, U_1^{n-1-\sigma}(\nu)\,.$$

In other words, the sequence $(U_j^n(\nu))$ satisfies

$$\begin{cases} U_j^{n+1}(\nu) = \sum_{\sigma=0}^{s} Q_\sigma\, U_j^{n-\sigma}(\nu) + F_j^n(\nu)\,, & j \geq 1\,, \quad n \geq s\,, \\ U_j^{n+1}(\nu) = \sum_{\sigma=-1}^{s} B_{j,\sigma}\, U_1^{n-\sigma}(\nu) + g_j^{n+1}(\nu)\,, & j = 1-r, \dots, 0\,, \quad n \geq s\,, \\ U_j^n(\nu) = 0\,, & j \geq 1-r\,, \quad n = 0, \dots, s\,. \end{cases} \tag{40}$$

It is not difficult to check that for all fixed n, $(U_j^n(\nu))$ and $(F_j^n(\nu))$ belong to ℓ^2. Moreover, $F_j^n(\nu) = 0$ for $n \geq 2\,s + \nu + 1$, and $g_j^n(\nu) = 0$ for $n \geq 2\,s + \nu + 2$. We can apply the strong stability estimate (36) for $\gamma = \gamma_0 := \ln|z_0| > 0$:

$$\frac{\gamma_0}{\gamma_0+1} \sum_{n\geq s+1} \sum_{j\geq 1-r} \mathrm{e}^{-2\,\gamma_0\, n}\, |U_j^n(\nu)|^2 + \sum_{n\geq s+1} \sum_{j=1-r}^{p} \mathrm{e}^{-2\,\gamma_0\, n}\, |U_j^n(\nu)|^2$$
$$\leq C_0 \left\{ \frac{\gamma_0+1}{\gamma_0} \sum_{n\geq s} \sum_{j\geq 1} \mathrm{e}^{-2\,\gamma_0\,(n+1)}\, |F_j^n(\nu)|^2 + \sum_{n\geq s+1} \sum_{j=1-r}^{0} \mathrm{e}^{-2\,\gamma_0\, n}\, |g_j^n(\nu)|^2 \right\}. \tag{41}$$

The right-hand side of (41) is finite because the sum over n involves finitely many terms. The above definition of $U_j^n(\nu)$ gives

$$\forall\, j \geq 1-r\,, \quad \sum_{n\geq s+1} \mathrm{e}^{-2\,\gamma_0\, n}\, |U_j^n(\nu)|^2 = \sum_{n=s+1}^{s+\nu} \mathrm{e}^{-2\,\gamma_0\, n}\, |z_0|^{2\,n}\, \frac{|W_j|^2}{\nu} = |W_j|^2\,,$$

so (41) reduces to

$$\frac{\gamma_0}{\gamma_0+1} \sum_{j\geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2$$
$$\leq C_0 \left\{ \frac{\gamma_0+1}{\gamma_0} \sum_{n\geq s} \sum_{j\geq 1} \mathrm{e}^{-2\,\gamma_0\,(n+1)}\, |F_j^n(\nu)|^2 + \sum_{n\geq s+1} \sum_{j=1-r}^{0} \mathrm{e}^{-2\,\gamma_0\, n}\, |g_j^n(\nu)|^2 \right\}.$$

Using Lemma 3.1, we have

$$\frac{|z_0|-1}{|z_0|}\sum_{j\geq 1-r}|W_j|^2+\sum_{j=1-r}^{p}|W_j|^2 \\ \leq 4\,C_0\left\{\frac{|z_0|}{|z_0|-1}\sum_{n\geq s}\sum_{j\geq 1}\mathrm{e}^{-2\,\gamma_0\,n}\,|z_0|^{-2}\,|F_j^n(\nu)|^2+\sum_{n\geq s+1}\sum_{j=1-r}^{0}\mathrm{e}^{-2\,\gamma_0\,n}\,|g_j^n(\nu)|^2\right\}. \quad (42)$$

The left-hand side of the inequality (42) does not depend on ν, and we are now going to compute the limit of the right-hand side in (42) as ν tends to $+\infty$.

Let us define the following functions on $\mathbb{R}^+$:

$$U_j(\nu,t):=\begin{cases}0\,, & \text{if } t\in[0,s+1[,\\ U_j^n(\nu)\,, & \text{if } t\in[n,n+1[,\ n\geq s+1,\end{cases}\qquad F_j(\nu,t):=\begin{cases}0\,, & \text{if } t\in[0,s[,\\ F_j^n(\nu)\,, & \text{if } t\in[n,n+1[,\ n\geq s,\end{cases}$$

$$g_j(\nu,t):=\begin{cases}0\,, & \text{if } t\in[0,s+1[,\\ g_j^n(\nu)\,, & \text{if } t\in[n,n+1[,\ n\geq s+1.\end{cases}$$

It is not difficult to check that the Laplace transform of each function $U_j(\nu,\cdot)$, $F_j(\nu,\cdot)$, $g_j(\nu,\cdot)$ is well-defined and holomorphic on $\mathbb{C}$, because each one of these functions is bounded with compact support in $\mathbb{R}^+$. To avoid any possible confusion, we recall that the Laplace transform of a function f defined on $\mathbb{R}^+$ is

$$\widehat{f}(\tau):=\int_{\mathbb{R}^+}\mathrm{e}^{-\tau\,t}\,f(t)\,\mathrm{d}t\,,$$

for all complex number τ such that the above integral makes sense. The system (40) equivalently reads

$$\begin{cases}U_j(\nu,t+1)=\displaystyle\sum_{\sigma=0}^{s}Q_\sigma\,U_j(\nu,t-\sigma)+F_j(\nu,t)\,, & j\geq 1\,,\quad t\geq s\,,\\ U_j(\nu,t)=\displaystyle\sum_{\sigma=-1}^{s}B_{j,\sigma}\,U_1(\nu,t-1-\sigma)+g_j(\nu,t)\,, & j=1-r,\dots,0\,,\quad t\geq s+1\,.\end{cases}$$

Multiplying the above equation by $\mathrm{e}^{-\tau\,t}$ and integrating over $[s,+\infty[$ or $[s+1,+\infty[$, we obtain

$$\begin{cases}\widehat{U_j}(\nu,\tau)-\displaystyle\sum_{\sigma=0}^{s}z^{-\sigma-1}\,Q_\sigma\,\widehat{U_j}(\nu,\tau)=z^{-1}\,\widehat{F_j}(\nu,\tau)\,, & j\geq 1\,,\\ \widehat{U_j}(\nu,\tau)-\displaystyle\sum_{\sigma=-1}^{s}z^{-\sigma-1}\,B_{j,\sigma}\,\widehat{U_j}(\nu,\tau)=\widehat{g_j}(\nu,\tau)\,, & j=1-r,\dots,0\,,\end{cases} \quad (43)$$

where we use the short notation $z:=\mathrm{e}^{\tau}$.

The Laplace transform $\widehat{U_j}(\nu,\tau)$ can be explicitely computed from the definition of $U_j^n(\nu)$. If we consider one $\tau_0\in\mathbb{C}$ such that $z_0:=\mathrm{e}^{\tau_0}$, then we get

$$\forall\,\theta\in\mathbb{R}\,,\quad \widehat{U_j}(\nu,\tau_0+i\,\theta)=\frac{1-z_0^{-1}\,\mathrm{e}^{-i\,\theta}}{\tau_0+i\,\theta}\,\mathrm{e}^{-i\,(s+1)\theta}\,\rho_\nu(\theta)\,W_j\,, \quad (44)$$

where ρ_ν stands for the function defined in Lemma 3.2. Using the first relation in (43), we obtain

$$z_0^{-1}\,\mathrm{e}^{-i\,\theta}\,\widehat{F_j}(\nu,\tau_0+i\,\theta)=\frac{1-z_0^{-1}\,\mathrm{e}^{-i\,\theta}}{\tau_0+i\,\theta}\,\mathrm{e}^{-i\,(s+1)\theta}\,\rho_\nu(\theta)\left(W_j-\sum_{\sigma=0}^{s}z_0^{-\sigma-1}\,\mathrm{e}^{-i\,(\sigma+1)\theta}\,Q_\sigma\,W_j\right).$$

Applying Plancherel's Theorem and Fubini's Theorem, we have

$$\begin{aligned}\sum_{n\geq s}\sum_{j\geq 1}\mathrm{e}^{-2\,\gamma_0\,n}\,|z_0|^{-2}\,|F_j^n(\nu)|^2 &= \frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\sum_{j\geq 1}|z_0|^{-2}\int_{\mathbb{R}^+}\mathrm{e}^{-2\,\gamma_0\,t}\,|F_j(\nu,t)|^2\,\mathrm{d}t\\ &= \frac{2\,\gamma_0}{2\,\pi\,(1-\mathrm{e}^{-2\,\gamma_0})}\sum_{j\geq 1}|z_0|^{-2}\int_{\mathbb{R}}\left|\widehat{F_j}(\nu,\tau_0+i\,\theta)\right|^2\mathrm{d}\theta\\ &= \frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\,\frac{1}{2\,\pi}\int_{\mathbb{R}}H(\theta)\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\,,\end{aligned}$$

where we have used the notation

$$\forall\,\theta\in\mathbb{R}\,,\quad H(\theta):=\left|\frac{1-z_0^{-1}\,\mathrm{e}^{-i\,\theta}}{\tau_0+i\,\theta}\right|^2\sum_{j\geq 1}\left|W_j-\sum_{\sigma=0}^{s}z_0^{-\sigma-1}\,\mathrm{e}^{-i\,(\sigma+1)\theta}\,Q_\sigma\,W_j\right|^2.$$

It is not so difficult to check that the function H satisfies the assumptions of property (iii) of Lemma 3.2. We thus have (recall the notation (39)):

$$\lim_{\nu\to+\infty}\sum_{n\geq s}\sum_{j\geq 1}\mathrm{e}^{-2\,\gamma_0\,n}\,|z_0|^{-2}\,|F_j^n(\nu)|^2=\frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\sum_{k\in\mathbb{Z}}\frac{|1-z_0^{-1}|^2}{|\tau_0+2\,i\,k\,\pi|^2}\sum_{j\geq 1}|(L(z_0)\,W)_j|^2\,. \tag{45}$$

With completely similar arguments, we can also obtain

$$\lim_{\nu\to+\infty}\sum_{n\geq s+1}\sum_{j=1-r}^{0}\mathrm{e}^{-2\,\gamma_0\,n}\,|g_j^n(\nu)|^2=\frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\sum_{k\in\mathbb{Z}}\frac{|1-z_0^{-1}|^2}{|\tau_0+2\,i\,k\,\pi|^2}\sum_{j=1-r}^{0}|(L(z_0)\,W)_j|^2\,. \tag{46}$$

Passing to the limit in (42) and using (45), (46), we get

$$\begin{aligned}&\frac{|z_0|-1}{|z_0|}\sum_{j\geq 1-r}|W_j|^2+\sum_{j=1-r}^{p}|W_j|^2\\ &\leq 4\,C_0\left\{\frac{|z_0|}{|z_0|-1}\sum_{j\geq 1}|(L(z_0)\,W)_j|^2+\sum_{j=1-r}^{0}|(L(z_0)\,W)_j|^2\right\}\frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\sum_{k\in\mathbb{Z}}\frac{|1-z_0^{-1}|^2}{|\tau_0+2\,i\,k\,\pi|^2}\,.\end{aligned} \tag{47}$$

Recalling the expression (44) for the Laplace transform of $U_j(\nu,\cdot)$, we have

$$\begin{aligned}|W_j|^2=\sum_{n\geq s+1}\mathrm{e}^{-2\,\gamma_0\,n}\,|U_j^n(\nu)|^2 &= \frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\int_{\mathbb{R}^+}\mathrm{e}^{-2\,\gamma_0\,t}\,|U_j(\nu,t)|^2\,\mathrm{d}t\\ &= \frac{2\,\gamma_0}{2\,\pi\,(1-\mathrm{e}^{-2\,\gamma_0})}\int_{\mathbb{R}}\left|\widehat{U_j}(\nu,\tau_0+i\,\theta)\right|^2\mathrm{d}\theta\\ &= \frac{2\,\gamma_0}{2\,\pi\,(1-\mathrm{e}^{-2\,\gamma_0})}\int_{\mathbb{R}}\left|\frac{1-z_0^{-1}\,\mathrm{e}^{-i\,\theta}}{\tau_0+i\,\theta}\right|^2|W_j|^2\,|\rho_\nu(\theta)|^2\,\mathrm{d}\theta\\ &\longrightarrow \frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\sum_{k\in\mathbb{Z}}\frac{|1-z_0^{-1}|^2}{|\tau_0+2\,i\,k\,\pi|^2}\,|W_j|^2\,.\end{aligned}$$

We have thus derived the formula

$$\frac{2\,\gamma_0}{1-\mathrm{e}^{-2\,\gamma_0}}\sum_{k\in\mathbb{Z}}\frac{|1-z_0^{-1}|^2}{|\tau_0+2\,i\,k\,\pi|^2}=1\,,$$

so we can simplify in (47) and obtain

$$\frac{|z_0|-1}{|z_0|}\sum_{j\geq 1-r}|W_j|^2+\sum_{j=1-r}^{p}|W_j|^2\leq 4\,C_0\left\{\frac{|z_0|}{|z_0|-1}\sum_{j\geq 1}|(L(z_0)\,W)_j|^2+\sum_{j=1-r}^{0}|(L(z_0)\,W)_j|^2\right\}. \tag{48}$$

The inequality (48) is only an a priori estimate for the operators $L(z)$, $z\in\mathscr{U}$. We emphasize that the constant $4\,C_0$ is independent of $z_0\in\mathscr{U}$ and $W\in\ell^2$. To complete the proof of Theorem

3.1, we only need to prove that each (bounded) operator $L(z)$ is invertible. This property is shown by combining two arguments.

Lemma 3.3. *There exists $R_0 \geq 1$ such that for all $z \in \mathbb{C}$ with $|z| > R_0$, the operator $L(z)$ defined by (39) is an isomorphism on ℓ^2.*

Proof of Lemma 3.3. Let L_∞ be defined by

$$L_\infty \,:\, W \in \ell^2 \longmapsto L_\infty W \in \ell^2 \quad \text{with } (L_\infty W)_j := \begin{cases} W_j\,, & \text{if } j \geq 1, \\ W_j - B_{j,-1}\, W_1\,, & \text{if } 1-r \leq j \leq 0. \end{cases}$$

It is easy to check that L_∞ is a bounded invertible operator on ℓ^2. Moreover, for $z \in \mathscr{U}$ and $W \in \ell^2$, we have

$$z\,\big((L_\infty - L(z))\,W\big)_j = \begin{cases} \sum_{\sigma=0}^{s} z^{-\sigma}\, Q_\sigma\, W_j\,, & \text{if } j \geq 1, \\ \sum_{\sigma=0}^{s} z^{-\sigma}\, B_{j,\sigma}\, W_1\,, & \text{if } 1-r \leq j \leq 0. \end{cases}$$

Consequently there exists a constant C such that

$$\forall\, z \in \mathscr{U}\,, \quad \|L_\infty - L(z)\|_{\mathscr{B}(\ell^2)} \leq \frac{C}{|z|}\,,$$

with $\mathscr{B}(\ell^2)$ the set of bounded operators on ℓ^2. This property implies that for $|z| > C\,\|L_\infty^{-1}\|_{\mathscr{B}(\ell^2)}$, $L(z)$ is an isomorphism. □

Lemma 3.4. *Let E be a Banach space, and let $\mathscr{T}$ denote a nonempty connected set. Let $\mathscr{L}$ be a continuous function on $\mathscr{T}$ with values in the space $\mathscr{B}(E)$ of bounded operators on E. Assume moreover that the two following conditions are satisfied:*

- *there exists a constant $M > 0$ such that for all $t \in \mathscr{T}$ and for all $x \in E$, we have $\|x\|_E \leq M\,\|\mathscr{L}(t)\,x\|_E$,*
- *there exists some $t_0 \in \mathscr{T}$ such that $\mathscr{L}(t_0)$ is an isomorphism.*

Then $\mathscr{L}(t)$ is an isomorphism for all $t \in \mathscr{T}$.

Proof of Lemma 3.4. We already know that $\mathscr{B}(E)$ is a Banach space and that the set of isomorphisms $Gl(E)$ is an open subset of $\mathscr{B}(E)$. This first property shows that the set $\{t \in \mathscr{T} / \mathscr{L}(t) \in Gl(E)\}$ is open because $\mathscr{L}$ is continuous. It only remains to show that this set is closed and the claim will follow (this set is nonempty thanks to the assumption of Lemma 3.4). We thus consider a sequence (t_n) in $\mathscr{T}$ that converges towards $\underline{t} \in \mathscr{T}$ and such that for all n, $\mathscr{L}(t_n)$ belongs to $Gl(E)$. We are going to show that $\mathscr{L}(\underline{t})$ also belongs to $Gl(E)$. Using the Banach isomorphism Theorem, it is enough to prove that $\mathscr{L}(\underline{t})$ is a bijection.

Due to the uniform bound $\|x\|_E \leq M\,\|\mathscr{L}(t)\,x\|_E$, it is clear that $\mathscr{L}(\underline{t})$ is injective and that for all n we have $\|\mathscr{L}(t_n)^{-1}\|_{\mathscr{B}(E)} \leq M$. It remains to show that $\mathscr{L}(\underline{t})$ is surjective. Let $y \in E$. For all integers n and p, we have:

$$\begin{aligned} &\big\|\mathscr{L}(t_{n+p})^{-1}\,y - \mathscr{L}(t_n)^{-1}\,y\big\|_E \leq \big\|\mathscr{L}(t_{n+p})^{-1} - \mathscr{L}(t_n)^{-1}\big\|_{\mathscr{B}(E)}\,\|y\|_E \\ &\quad \leq \big\|\mathscr{L}(t_{n+p})^{-1}\,(\mathscr{L}(t_n) - \mathscr{L}(t_{n+p}))\,\mathscr{L}(t_n)^{-1}\big\|_{\mathscr{B}(E)}\,\|y\|_E \leq M^2\,\|\mathscr{L}(t_{n+p}) - \mathscr{L}(t_n)\|_{\mathscr{B}(E)}\,\|y\|_E\,. \end{aligned}$$

These inequalities show that $(\mathscr{L}(t_n)^{-1}\,y)$ is a Cauchy sequence in E and therefore converges towards $x \in E$. Moreover we have $\mathscr{L}(t_n)\,\mathscr{L}(t_n)^{-1}\,y = y$ for all n so, passing to the limit, we get $\mathscr{L}(\underline{t})\,x = y$. Here we use again the continuity of $\mathscr{L}$. This shows that $\mathscr{L}(\underline{t})$ is surjective, which completes the proof. □

Lemma 3.4 shows that the resolvent equation can be uniquely solved for all $z \in \mathscr{U}$. Indeed, for all integer ν sufficiently large, the mapping L restricted to the annulus $\{z \in \mathbb{C}\,,\, 1 + 2^{-\nu} \leq |z| \leq 2^\nu\}$ satisfies the assumptions of Lemma 3.4 (use Lemma 3.3 for the existence of one point where L is an isomorphism and (48) for the uniform bound). We leave to the reader the verification that $L(z) \in \mathscr{B}(\ell^2)$ depends continuously on $z \in \mathscr{U}$. Eventually we can conclude that $L(z)$ is an isomorphism for all $z \in \mathscr{U}$ and $L(z)^{-1}$ satisfies the estimate (48) which is nothing else but (38). The proof of Theorem 3.1 is now complete. □

Theorem 3.1 has an important consequence, which is the following well-known necessary condition for strong stability.

Corollary 3.1 (Godunov-Ryabenkii condition)**.** *Assume that the scheme (32) is strongly stable in the sense of Definition 3.1. Then for all $z \in \mathscr{U}$, any $W \in \ell^2$ satisfying*

$$\begin{cases} W_j - \displaystyle\sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, W_j = 0\,, & j \geq 1\,, \\ W_j - \displaystyle\sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{j,\sigma}\, W_1 = 0\,, & j = 1-r, \dots, 0\,, \end{cases}$$

must be zero.

The Godunov-Ryabenkii condition is a preliminary test in view of showing strong stability. It is analogous to the Lopatinskii condition for hyperbolic initial boundary value problems. As we shall see later on, it is unfortunately not a sufficient condition for strong stability (see the following paragraphs for more comments).

In the remaining part of this paragraph, we are going to show the converse result of Theorem 3.1.

Theorem 3.2 (Gustafsson, Kreiss, Sundström [10])**.** *Assume that there exists a constant $C_1 > 0$ such that for all $z \in \mathscr{U}$, for all $(F_j) \in \ell^2$ and for all vectors $g_{1-r}, \dots, g_0 \in \mathbb{C}^N$, the resolvent equation (37) has a unique solution $(W_j) \in \ell^2$ and this solution satisfies*

$$\frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 \leq C_1 \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

Then the scheme (32) is strongly stable and satisfies (34) with the constant $C_1 \max(1,\lambda)/\min(1,\lambda)$.

The proof of Theorem 3.2 splits in several steps. In what follows, we shall say that a sequence (U_j^n) has compact support if the terms U_j^n vanish except for a finite number of indeces (j,n).

Proof of Theorem 3.2. • We consider some source terms (F_j^n), (g_j^n) for (35) with compact support. We also let (U_j^n) denote the solution to (35). It is easy to show by induction on n that for all n, the sequence $(U_j^n)_{j \geq 1-r}$ belongs to ℓ^2. For $\gamma > 0$, we introduce the quantities

$$\mathscr{I}_N(\gamma) := \sum_{n=0}^{N} \sum_{j \geq 1} \mathrm{e}^{-2\gamma n} |U_j^n|^2\,, \qquad \mathscr{B}_N(\gamma) := \sum_{n=0}^{N} \sum_{j=1-r}^{0} \mathrm{e}^{-2\gamma n} |U_j^n|^2\,,$$

$$S_F(\gamma) := \sum_{n \geq s} \sum_{j \geq 1} \mathrm{e}^{-2\gamma(n+1)} |F_j^n|^2\,, \qquad S_g(\gamma) := \sum_{n \geq s+1} \sum_{j=1-r}^{0} \mathrm{e}^{-2\gamma n} |g_j^n|^2\,.$$

Performing very crude estimates in (35), we immediately see that there exists a constant $C > 0$ that is independent of F, g, U such that

$$\forall\, j \geq 1\,, \quad \forall\, n \geq s\,, \quad |U_j^{n+1}|^2 \leq C \left(|F_j^n|^2 + \sum_{\sigma=0}^{s} \sum_{\ell=-r}^{p} |U_{j+\ell}^{n-\sigma}|^2 \right),$$

$$\forall\, j = 1-r, \dots, 0\,, \quad \forall\, n \geq s\,, \quad |U_j^{n+1}|^2 \leq C \left(|g_j^{n+1}|^2 + \sum_{\sigma=-1}^{s} \sum_{\ell=0}^{q} |U_{1+\ell}^{n-\sigma}|^2 \right).$$

Multiplying each inequality by $\exp(-2\gamma(n+1))$ and taking the sum, we obtain

$$\forall\, N \geq s+1\,, \quad \mathscr{I}_N(\gamma) \leq C\, S_F(\gamma) + C\, \mathrm{e}^{-2\gamma} \sum_{\sigma=0}^{s} \mathscr{I}_{N-1-\sigma}(\gamma) + \mathscr{B}_{N-1-\sigma}(\gamma)\,,$$

$$\mathscr{B}_N(\gamma) \leq C\, \mathscr{I}_N(\gamma) + C\, S_g(\gamma) + C\, \mathrm{e}^{-2\gamma} \sum_{\sigma=0}^{s} \mathscr{I}_{N-1-\sigma}(\gamma)\,,$$

with a possibly larger constant C. Using the obvious inequalities

$$\mathscr{I}_{N-1-\sigma}(\gamma) \leq \mathscr{I}_N(\gamma)\,, \quad \mathscr{B}_{N-1-\sigma}(\gamma) \leq \mathscr{B}_N(\gamma)\,,$$

and combining the above estimates for $\mathscr{I}_N(\gamma), \mathscr{B}_N(\gamma)$, we obtain that for some large enough $\underline{\gamma} > 0$, that is independent of F, g, U and N, there holds

$$\forall\, \gamma \geq \underline{\gamma}\,, \quad \forall\, N \geq s+1\,, \quad \mathscr{I}_N(\gamma) + \mathscr{B}_N(\gamma) \leq C\left(S_F(\gamma) + S_g(\gamma)\right).$$

In other words, for $\gamma \geq \underline{\gamma}$, we have

$$\sum_{n \geq s+1} \sum_{j \geq 1-r} \mathrm{e}^{-2\,\gamma\, n}\, |U_j^n|^2 < +\infty\,. \tag{49}$$

• As in the proof of Theorem 3.1, it is convenient to introduce the following functions defined on $\mathbb{R}^+$:

$$U_j(t) := \begin{cases} 0\,, & \text{if } t \in [0, s+1[, \\ U_j^n\,, & \text{if } t \in [n, n+1[,\ n \geq s+1, \end{cases} \qquad F_j(t) := \begin{cases} 0\,, & \text{if } t \in [0, s[, \\ F_j^n\,, & \text{if } t \in [n, n+1[,\ n \geq s, \end{cases}$$

$$g_j(t) := \begin{cases} 0\,, & \text{if } t \in [0, s+1[, \\ g_j^n\,, & \text{if } t \in [n, n+1[,\ n \geq s+1. \end{cases}$$

Then (49) reads

$$\forall\, \gamma \geq \underline{\gamma}\,, \quad \sum_{j \geq 1-r} \int_{\mathbb{R}^+} \mathrm{e}^{-2\,\gamma\, t}\, |U_j(t)|^2\, \mathrm{d}t < +\infty\,. \tag{50}$$

The Laplace transforms $\widehat{F_j}, \widehat{g_j}$ are well-defined and holomorphic on $\mathbb{C}$, and $\widehat{F_j}$ is identically zero for j large enough. Moreover, (50) shows that the Laplace transforms $\widehat{U_j}$, $j \geq 1-r$, are well-defined and holomorphic on $\{\mathrm{Re}\ \tau > \underline{\gamma}\}$, with $\underline{\gamma}$ independent of j. Applying Plancherel's Theorem in (50), we find that for all $\gamma > \underline{\gamma}$ and for almost every $\theta \in \mathbb{R}$, the sequence $\left(\widehat{U_j}(\gamma + i\,\theta)\right)_{j \geq 1-r}$ belongs to ℓ^2.

Applying the Laplace transform on (35) with $\mathrm{Re}\ \tau > \underline{\gamma}$, we get

$$\begin{cases} \widehat{U_j}(\tau) - \displaystyle\sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, \widehat{U_j}(\tau) = z^{-1}\, \widehat{F_j}(\tau)\,, & j \geq 1\,, \\ \widehat{U_j}(\tau) - \displaystyle\sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{j,\sigma}\, \widehat{U_j}(\tau) = \widehat{g_j}(\tau)\,, & j = 1-r, \dots, 0\,, \end{cases} \tag{51}$$

where we still use the short notation $z := \mathrm{e}^\tau$ as in the proof of Theorem 3.1.

Since $\widehat{F_j}$ vanishes for j large, we have

$$\forall\, \tau \in \mathbb{C}\,, \quad \sum_{j \geq 1} |z|^{-2} \left|\widehat{F_j}(\tau)\right|^2 < +\infty\,.$$

For all $\tau \in \mathbb{C}$ with $\mathrm{Re}\ \tau > 0$, we can thus define $(W_j(\tau))_{j \geq 1-r}$ as the unique solution in ℓ^2 to the resolvent equation

$$\begin{cases} W_j(\tau) - \displaystyle\sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, W_j(\tau) = z^{-1}\, \widehat{F_j}(\tau)\,, & j \geq 1\,, \\ W_j(\tau) - \displaystyle\sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{j,\sigma}\, W_j(\tau) = \widehat{g_j}(\tau)\,, & j = 1-r, \dots, 0\,. \end{cases} \tag{52}$$

Moreover, $(W_j(\tau))_{j \geq 1-r}$ satisfies

$$\forall\, \tau \in \mathbb{C}\,, \quad \mathrm{Re}\ \tau > 0\,, \quad \frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j(\tau)|^2 + \sum_{j=1-r}^{p} |W_j(\tau)|^2$$

$$\leq C_1 \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |z|^{-2} \left|\widehat{F_j}(\tau)\right|^2 + \sum_{j=1-r}^{0} \left|\widehat{g_j}(\tau)\right|^2 \right\}. \tag{53}$$

The difference between (51) and (52) is that (51) holds only for $\mathrm{Re}\,\tau > \underline{\gamma}$ while (52) holds for $\mathrm{Re}\,\tau > 0$. Our goal is to identify (W_j) and $(\widehat{U_j})$ and to show that (51) holds for $\mathrm{Re}\,\tau > 0$. This is based on the following result, the proof of which is left to the reader.

Lemma 3.5. *The operator $L(z) \in \mathscr{B}(\ell^2)$ in (39) depends holomorphically on $z \in \mathbb{C} \setminus \{0\}$. Consequently, under the assumptions of Theorem 3.2, $L(\mathrm{e}^\tau)^{-1}$ depends holomorphically on τ for $\mathrm{Re}\,\tau > 0$.*

Lemma 3.5 implies that for all $j \geq 1-r$, W_j is holomorphic on $\{\mathrm{Re}\,\tau > 0\}$ because the source terms in (52) depend holomorphically on τ. Furthermore, we know that $\widehat{U_j}$ is holomorphic on $\{\mathrm{Re}\,\tau > \underline{\gamma}\}$, and that for all $\gamma > \underline{\gamma}$ and almost every $\theta \in \mathbb{R}$, $(\widehat{U_j}(\gamma + i\,\theta))$ belongs to ℓ^2 and is a solution to (51). This implies $\widehat{U_j}(\gamma + i\,\theta) = W_j(\gamma + i\,\theta)$ for $\gamma > \underline{\gamma}$ and for almost every $\theta \in \mathbb{R}$. Since both functions are holomorphic, we have obtained

$$\forall\, j \geq 1-r\,, \quad \forall\, \gamma > \underline{\gamma}\,, \quad \forall\, \theta \in \mathbb{R}\,, \quad \widehat{U_j}(\gamma + i\,\theta) = W_j(\gamma + i\,\theta)\,.$$

Let now $\gamma_0 > 0$. We integrate (53) with respect to $\theta \in \mathbb{R}$ for $\tau = \gamma + i\,\theta$ and $\gamma > \gamma_0$, and use Plancherel's Theorem to compute the right-hand side of the inequality. We thus obtain

$$\sup_{\gamma > \gamma_0} \sum_{j \geq 1-r} \int_{\mathbb{R}} |W_j(\gamma + i\,\theta)|^2 \,\mathrm{d}\theta < +\infty\,.$$

Applying the Paley-Wiener Theorem for which we refer to [19], this means that for all $j \geq 1-r$, there exists a measurable function V_j on $\mathbb{R}^+$ such that

$$\int_{\mathbb{R}^+} \mathrm{e}^{-2\,\gamma_0\,t}\,|V_j(t)|^2 \,\mathrm{d}t < +\infty\,,$$

and $W_j = \widehat{V_j}$ on $\{\mathrm{Re}\,\tau > \gamma_0\}$. By injectivity of the Laplace transform, V_j must equal U_j. In other words, we have just proved that for all $\gamma_0 > 0$, $\exp(-\gamma_0\,t)\,U_j$ belongs to $L^2(\mathbb{R}^+)$, so $\widehat{U_j}$ is well-defined on $\{\mathrm{Re}\,\tau > 0\}$ and coincides with W_j. Hence (53) holds with $\widehat{U_j}$ instead of W_j. We now integrate (53) with respect to $\theta = \mathrm{Im}\,\tau$ and use Plancherel's Theorem, which yields

$$\frac{\mathrm{e}^\gamma - 1}{\mathrm{e}^\gamma} \sum_{n \geq s+1} \sum_{j \geq 1-r} \mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{p} \mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2$$
$$\leq C_1 \left\{ \frac{\mathrm{e}^\gamma}{\mathrm{e}^\gamma - 1} \sum_{n \geq s} \sum_{j \geq 1} \mathrm{e}^{-2\,\gamma\,(n+1)}\,|F_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{0} \mathrm{e}^{-2\,\gamma\,n}\,|g_j^n|^2 \right\},$$

for all $\gamma > 0$. Applying Lemma 3.1, we get

$$\frac{\gamma}{\gamma+1} \sum_{n \geq s+1} \sum_{j \geq 1-r} \mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{p} \mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2$$
$$\leq C_1 \left\{ \frac{\gamma+1}{\gamma} \sum_{n \geq s} \sum_{j \geq 1} \mathrm{e}^{-2\,\gamma\,(n+1)}\,|F_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{0} \mathrm{e}^{-2\,\gamma\,n}\,|g_j^n|^2 \right\},$$

• It is useful to recall that the latter estimate was derived under the assumption that the source terms had compact support. To complete the proof of Theorem 3.2, it is sufficient to prove Lemma 3.6 below.

Lemma 3.6. *Assume that for all data (F_j^n) and (g_j^n) with compact support, the solution (U_j^n) to (35) satisfies*

$$\frac{\gamma}{\gamma+1} \sum_{n \geq s+1} \sum_{j \geq 1-r} \mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{p} \mathrm{e}^{-2\,\gamma\,n}\,|U_j^n|^2$$
$$\leq C_1 \left\{ \frac{\gamma+1}{\gamma} \sum_{n \geq s} \sum_{j \geq 1} \mathrm{e}^{-2\,\gamma\,(n+1)}\,|F_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{0} \mathrm{e}^{-2\,\gamma\,n}\,|g_j^n|^2 \right\},$$

for all $\gamma > 0$. Then the scheme (32) satisfies (34) with the constant $C_1 \max(1,\lambda)/\min(1,\lambda)$.

Proof of Lemma 3.6. Let us consider some source terms (F_j^n), (g_j^n) for (35), not necessarily with compact support. Let $\gamma > 0$ such that the right-hand side of (36) is finite. For $\nu \geq 1$, we define

$$F_j^n(\nu) := \begin{cases} F_j^n\,, & \text{if } s \leq n \leq s+\nu-1 \text{ and } 1 \leq j \leq 1+q+\nu p, \\ 0\,, & \text{otherwise,} \end{cases}$$

$$g_j^n(\nu) := \begin{cases} g_j^n\,, & \text{if } s+1 \leq n \leq s+\nu \text{ and } 1-r \leq j \leq 0, \\ 0\,, & \text{otherwise.} \end{cases}$$

A direct induction argument shows that the corresponding solution $(U_j^n(\nu))$ to (35) satisfies $U_j^n(\nu) = U_j^n$ for $0 \leq n \leq s+\nu$ and $1-r \leq j \leq \nu$. We thus have

$$\frac{\gamma}{\gamma+1} \sum_{n \leq s+\nu} \sum_{1-r \leq j \leq \nu} \mathrm{e}^{-2\gamma n} |U_j^n|^2 + \sum_{n \leq s+\nu} \sum_{j=1-r}^{p} \mathrm{e}^{-2\gamma n} |U_j^n|^2 \\ \leq C_1 \left\{ \frac{\gamma+1}{\gamma} \sum_{n \geq s} \sum_{j \geq 1} \mathrm{e}^{-2\gamma(n+1)} |F_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{0} \mathrm{e}^{-2\gamma n} |g_j^n|^2 \right\},$$

for all $\gamma > 0$ and all $\nu \geq p+1$. Passing to the limit $\nu \to +\infty$, we have proved that (36) holds with the constant C_1 and without any assumption of compact support on the data. To prove that (34) holds, it is sufficient to apply (36) with the source term $\Delta t\, F_j^n$ instead of F_j^n, and with the parameter $\gamma\,\Delta t > 0$ instead of γ. Recalling the relation $\Delta t = \lambda\,\Delta x$, we obtain the result. The details are left to the reader. □

□

We summarize the main results of this paragraph into the following result.

Theorem 3.3 (Characterization of strong stability [10])**.** *The scheme (32) is strongly stable in the sense of Definition 3.1 if and only if there exists a constant $C > 0$ such that for all $z \in \mathscr{U}$, for all $(F_j) \in \ell^2$ and for all vectors $g_{1-r}, \dots, g_0 \in \mathbb{C}^N$, the resolvent equation (37) has a unique solution $(W_j) \in \ell^2$ and this solution satisfies*

$$\frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 \leq C \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

In particular, if the scheme (32) is strongly stable, then the Godunov-Ryabenkii condition of Corollary 3.1 holds.

It is also useful to keep in mind that showing the unique solvability of the resolvent equation relies on a rather simple argument of functional analysis that reduces to the verification of an a priori estimate. Furthermore, the resolvent equation becomes trivially solvable for $|z|$ large. More precisely we state a slightly refined version of Theorem 3.3 which will be useful in the following paragraph. Theorem 3.4 below shows that it is sufficient to consider the resolvent equation for bounded parameters z.

Theorem 3.4 (Characterization of strong stability)**.** *The scheme (32) is strongly stable in the sense of Definition 3.1 if and only if for all $R \geq 2$, there exists a constant $C_R > 0$ such that for all $z \in \mathscr{U}$ with $|z| \leq R$, for all $(F_j) \in \ell^2$ and for all vectors $g_{1-r}, \dots, g_0 \in \mathbb{C}^N$, the resolvent equation (37) has a unique solution $(W_j) \in \ell^2$ and this solution satisfies*

$$\frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

Proof of Theorem 3.4. • Let us first assume that the scheme (32) is strongly stable in the sense of Definition 3.1. Then we apply Theorem 3.3: the resolvent equation can be uniquely solved in ℓ^2 for all $z \in \mathscr{U}$ with the estimate

$$\frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 \leq C \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

This shows that the conclusion of Theorem 3.4 holds with a constant $C_R := C$ that is independent of $R \geq 2$.

• Let us now assume that for all $R \geq 2$, the resolvent equation can be uniquely solved in ℓ^2 for all $z \in \mathscr{U}$, $|z| \leq R$, with the estimate

$$\frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

We first apply Lemma 3.3 and keep the notation introduced in the proof of this Lemma. There exists $R_0 \geq 2$ such that for all $z \in \mathbb{C}$ with $|z| \geq R_0$, the mapping $L(z) \in \mathscr{B}(\ell^2)$ is an isomorphism and $\|L(z) - L_\infty\|_{\mathscr{B}(\ell^2)} \leq \|L_\infty^{-1}\|_{\mathscr{B}(\ell^2)}^{-1}/2$. In particular, there exists a constant $\underline{C} > 0$ such that for all $z \in \mathbb{C}$ with $|z| \geq R_0$, the unique solution $(W_j) \in \ell^2$ to (37) satisfies

$$\sum_{j \geq 1-r} |W_j|^2 \leq \underline{C} \left\{ \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

This estimate yields

$$\begin{aligned} \frac{|z|-1}{|z|} \sum_{j \geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 &\leq 2 \sum_{j \geq 1-r} |W_j|^2 \\ &\leq 2\,\underline{C} \left\{ \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\} \\ &\leq 2\,\underline{C} \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}. \end{aligned}$$

It remains to use the assumption for the radius R_0 and consider the constant $\max(2\,\underline{C}, C_{R_0})$. Theorem 3.3 then shows that the scheme (32) is strongly stable. □

In the following paragraph, we shall write the resolvent equation into an equivalent but more convenient form. This will lead to the formulation of our main result which characterizes strong stability in terms of an algebraic condition that is analogous to the so-called uniform Kreiss-Lopatinskii condition.

3.3. **An equivalent form of the resolvent equation.** The equation

$$W_j - \sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, W_j = F_j\,,$$

defines an induction relation of order $p + r$ on the sequence (W_j). It is convenient to rewrite this induction relation as an induction of order 1 for an augmented sequence. This is a classical procedure. For $\ell = -r, \dots, p$, we define the matrices

$$\forall\, z \in \mathbb{C} \setminus \{0\}\,, \quad \mathbb{A}_\ell(z) := \delta_{\ell 0}\, I - \sum_{\sigma=0}^{s} z^{-\sigma-1}\, A_{\ell,\sigma}\,, \tag{54}$$

where $\delta_{\ell_1 \ell_2}$ denotes the Kronecker symbol. We also define the matrices

$$\forall\, \ell = 0, \dots, q\,, \quad \forall\, j = 1-r, \dots, 0\,, \quad \forall\, z \in \mathbb{C} \setminus \{0\}\,, \quad \mathbb{B}_{\ell,j}(z) := \sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{\ell,j,\sigma}\,. \tag{55}$$

With these definitions, the reader will easily verify that (37) equivalently reads (use (33))

$$\begin{cases} \displaystyle\sum_{\ell=-r}^{p} \mathbb{A}_\ell(z)\, W_{j+\ell} = F_j\,, & j \geq 1\,, \\ \displaystyle W_j - \sum_{\ell=0}^{q} \mathbb{B}_{\ell,j}(z)\, W_{1+\ell} = g_j\,, & j = 1-r, \dots, 0\,. \end{cases} \tag{56}$$

To rewrite (56) as an induction relation of order 1, we make, as in [10], the following assumption:

Assumption 3.1 (Noncharacteristic discrete boundary)**.** *The matrices $\mathbb{A}_{-r}(z)$ and $\mathbb{A}_p(z)$ are invertible for all $z \in \overline{\mathscr{U}}$, or equivalently for all $z \in \mathbb{C}$ with $|z| > 1 - \varepsilon_0$ for some $\varepsilon_0 \in]0, 1/2]$.*

Let us first consider the case $q < p$. In that case, all the W_j's involved in the boundary conditions for the resolvent equation (56) are coordinates of the augmented vector[9] $\mathscr{W}_1 := (W_p, \dots, W_{1-r}) \in \mathbb{C}^{N(p+r)}$. Using Assumption 3.1, we can define a block companion matrix $\mathbb{M}(z)$ that is holomorphic on some open neighborhood $\mathscr{V} := \{z \in \mathbb{C}\,,\ |z| > 1 - \varepsilon_0\}$ of $\overline{\mathscr{U}}$:

$$\forall\, z \in \mathscr{V}\,, \quad \mathbb{M}(z) := \begin{pmatrix} -\mathbb{A}_p(z)^{-1}\, \mathbb{A}_{p-1}(z) & \dots & \dots & -\mathbb{A}_p(z)^{-1}\, \mathbb{A}_{-r}(z) \\ I & 0 & \dots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & 0 & I & 0 \end{pmatrix} \in \mathscr{M}_{N(p+r)}(\mathbb{C})\,. \tag{57}$$

We also define the matrix that encodes the boundary conditions for (56), namely

$$\forall\, z \in \mathbb{C} \setminus \{0\}\,,\ \mathbb{B}(z) := \begin{pmatrix} 0 & \dots & 0 & -\mathbb{B}_{q,0}(z) & \dots & -\mathbb{B}_{0,0}(z) & I & & 0 \\ \vdots & & \vdots & \vdots & & \vdots & & \ddots & \\ 0 & \dots & 0 & -\mathbb{B}_{q,1-r}(z) & \dots & -\mathbb{B}_{0,1-r}(z) & 0 & & I \end{pmatrix} \in \mathscr{M}_{Nr,N(p+r)}(\mathbb{C})\,, \tag{58}$$

with the $\mathbb{B}_{\ell,j}$'s defined in (55). With such definitions, it is a simple exercise to rewrite the resolvent equation (56) as an induction relation for the augmented vector $\mathscr{W}_j := (W_{j+p-1}, \dots, W_{j-r}) \in \mathbb{C}^{N\,(p+r)}$, $j \geq 1$. This induction relation takes the form

$$\begin{cases} \mathscr{W}_{j+1} = \mathbb{M}(z)\, \mathscr{W}_j + \mathscr{F}_j\,, & j \geq 1\,, \\ \mathbb{B}(z)\, \mathscr{W}_1 = \mathscr{G}\,, \end{cases} \tag{59}$$

where the new source terms $(\mathscr{F}_j), \mathscr{G}$ in (59) are given by:

$$\mathscr{F}_j := (\mathbb{A}_p(z)^{-1}\, F_j, 0, \dots, 0)\,, \quad \mathscr{G} := (g_0, \dots, g_{1-r})\,.$$

Remark 3.2. *It is easy to check that the matrix $\mathbb{B}(z)$ in (58) depends holomorphically on $z \in \mathbb{C}\setminus\{0\}$ and has maximal rank $N\,r$ for all z (just consider the $N\,r \times N\,r$ submatrix formed by the last columns). Consequently, the kernel of $\mathbb{B}(z)$ has dimension $N\,p$ for all $z \in \mathbb{C} \setminus \{0\}$.*

Let us now characterize strong stability for (32) in terms of an estimate for (59). Of course we shall use Theorem 3.4 and the strong relationship between (37) and (59).

Proposition 3.1 (Characterization of strong stability)**.** *Let Assumption 3.1 be satisfied, and let us assume $q < p$. Then the scheme (32) is strongly stable in the sense of Definition 3.1 if and only if for all $R \geq 2$, there exists a constant $C_R > 0$ such that for all $z \in \mathscr{U}$ with $|z| \leq R$, for all $(\mathscr{F}_j) \in \ell^2$ and for all $\mathscr{G} \in \mathbb{C}^{Nr}$, the equation (59) has a unique solution $(\mathscr{W}_j) \in \ell^2$ and this solution satisfies*

$$\frac{|z|-1}{|z|} \sum_{j\geq 1} |\mathscr{W}_j|^2 + |\mathscr{W}_1|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |\mathscr{F}_j|^2 + |\mathscr{G}|^2 \right\}. \tag{60}$$

The main point to keep in mind is that in Proposition 3.1, the source terms $\mathscr{F}_j$ may be arbitrary in $\mathbb{C}^{N(p+r)}$, while when we rewrote (37) under the form (59), only the first coordinate of $\mathscr{F}_j$ was nonzero.

[9] Vectors are now written indifferently in rows or columns in order to simplify the redaction.

Proof of Proposition 3.1. • Let us first assume that the scheme (32) is strongly stable so we can apply Theorem 3.4. Our goal is to show that (59) has a unique solution for all source terms in ℓ^2 and that the estimate (60) holds for a suitable constant C_R. As a warm-up, let us first show that if a solution in ℓ^2 to (59) exists, then it is necessarily unique. By linearity, this amounts to proving that if $(\mathscr{W}_j)_{j\geq 1} \in \ell^2$ satisfies

$$\begin{cases} \mathscr{W}_{j+1} = \mathbb{M}(z)\,\mathscr{W}_j\,, & j \geq 1\,, \\ \mathbb{B}(z)\,\mathscr{W}_1 = 0\,, \end{cases}$$

then $(\mathscr{W}_j)_{j\geq 1}$ is zero (this is a new formulation of the Godunov-Ryabenkii condition). We thus consider such a sequence $(\mathscr{W}_j)_{j\geq 1}$, and we introduce the decomposition $\mathscr{W}_j = (\mathscr{W}_j^{(1)}, \dots, \mathscr{W}_j^{(p+r)})$, where each vector $\mathscr{W}_j^{(k)}$ belongs to $\mathbb{C}^N$. Using the block decomposition of $\mathbb{M}(z)$ - recall the definition (57) - we obtain

$$\forall\, \ell = -r, \dots, p-1\,, \quad \forall\, j \geq 1\,, \quad \mathscr{W}_j^{(p-\ell)} = \mathscr{W}_{j+\ell+r}^{(p+r)}\,.$$

Furthermore, the sequence defined by $W_j := \mathscr{W}_{j+r}^{(p+r)}$, $j \geq 1-r$, satisfies the homogeneous resolvent equation

$$\begin{cases} \displaystyle\sum_{\ell=-r}^{p} \mathbb{A}_\ell(z)\, W_{j+\ell} = 0\,, & j \geq 1\,, \\ \displaystyle W_j - \sum_{\ell=0}^{q} \mathbb{B}_{\ell,j}(z)\, W_{1+\ell} = 0\,, & j = 1-r, \dots, 0\,. \end{cases}$$

The Godunov-Ryabenkii condition (Corollary 3.1) gives $(W_j)_{j\geq 1-r} = 0$, which yields $(\mathscr{W}_j)_{j\geq 1} = 0$. If a solution in ℓ^2 to (59) exists, it is necessarily unique.

Let now $R \geq 2$, let $z \in \mathscr{U}$ satisfy $|z| \leq R$, and let us consider $(\mathscr{F}_j) \in \ell^2$, $\mathscr{G} \in \mathbb{C}^{Nr}$. We wish to construct a solution $(\mathscr{W}_j) \in \ell^2$ to (59). We use again the decomposition $\mathscr{W}_j = (\mathscr{W}_j^{(1)}, \dots, \mathscr{W}_j^{(p+r)})$ as well as the notation $W_j := \mathscr{W}_{j+r}^{(p+r)}$, $j \geq 1-r$. The source terms are also decomposed as $\mathscr{F}_j = (\mathscr{F}_j^{(1)}, \dots, \mathscr{F}_j^{(p+r)})$, $\mathscr{G} = (\mathscr{G}^{(0)}, \dots, \mathscr{G}^{(1-r)})$. Inspecting the system (59) shows that we should necessarily have

$$\forall\, \ell = -r, \dots, p-1\,, \quad \forall\, j \geq 1\,, \quad \mathscr{W}_j^{(p-\ell)} = W_{j+\ell} - \sum_{k=-r}^{\ell-1} \mathscr{F}_{j+\ell-1-k}^{(p-k)}\,. \tag{61}$$

Moreover, the sequence $(W_j)_{j\geq 1-r}$ should be a solution to (56) with source terms $(F_j), g_{1-r}, \dots, g_0$ defined by

$$\forall\, j \geq 1\,, \quad F_j := \sum_{\ell=-r}^{p} \mathbb{A}_\ell(z) \sum_{k=-r}^{\ell-1} \mathscr{F}_{j+\ell-1-k}^{(p-k)}\,, \tag{62}$$

$$\forall\, j = 1-r, \dots, 0\,, \quad g_j := \mathscr{G}^{(j)} + \sum_{k=-r}^{j-2} \mathscr{F}_{j-1-k}^{(p-k)} - \sum_{\ell=0}^{q} \mathbb{B}_{\ell,j}(z) \sum_{k=-r}^{\ell-1} \mathscr{F}_{\ell-k}^{(p-k)}\,. \tag{63}$$

An important remark in view of what follows is that the matrices $\mathbb{A}_\ell$ and $\mathbb{B}_{\ell,j}$ defined by (54), (55) are bounded on $\overline{\mathscr{U}}$. Consequently, it is rather easy to check that the relations (62), (63) define a sequence $(F_j) \in \ell^2$ and vectors $g_{1-r}, \dots, g_0 \in \mathbb{C}^N$ such that, for a given constant M that does not depend on z nor on R, there holds

$$\sum_{j\geq 1} |F_j|^2 \leq M \sum_{j\geq 1} |\mathscr{F}_j|^2\,, \quad \sum_{j=1-r}^{0} |g_j|^2 \leq M \left(\sum_{j\geq 1} |\mathscr{F}_j|^2 + |\mathscr{G}|^2 \right)\,. \tag{64}$$

Applying Theorem 3.4, we know that there exists a unique solution $(W_j) \in \ell^2$ to (56) with the source terms defined in (62), (63), and that for some constant C_R independent of z and of the source

terms, there holds

$$\frac{|z|-1}{|z|}\sum_{j\geq 1-r}|W_j|^2+\sum_{j=1-r}^{p}|W_j|^2\leq C_R\left\{\frac{|z|}{|z|-1}\sum_{j\geq 1}|F_j|^2+\sum_{j=1-r}^{0}|g_j|^2\right\}.$$

The relation (61) defines a sequence $(\mathscr{W}_j)_{j\geq 1}$ in $\mathbb{C}^{N(p+r)}$; it is not difficult to check that this sequence belongs to ℓ^2 and that it is a solution to (59). Moreover, combining (61), (64) and the latter estimate for (W_j), we obtain

$$\frac{|z|-1}{|z|}\sum_{j\geq 1}|\mathscr{W}_j|^2+|\mathscr{W}_1|^2\leq C_R\left\{\frac{|z|}{|z|-1}\sum_{j\geq 1}|\mathscr{F}_j|^2+|\mathscr{G}|^2\right\}.$$

As already shown above, such a solution $(\mathscr{W}_j)\in\ell^2$ to (59) is unique.

• Let us now assume that (59) has a unique solution in ℓ^2 for all source terms $(\mathscr{F}_j),\mathscr{G}$ and that the estimate (60) holds. Let now $R\geq 2$, let $z\in\mathscr{U}$ with $|z|\leq R$, and let us consider some source terms $(F_j)\in\ell^2$, $g_{1-r},\dots,g_0\in\mathbb{C}^N$ for the resolvent equation (37). We define the vectors

$$\mathscr{F}_j:=(\mathbb{A}_p(z)^{-1}F_j,0,\dots,0),\quad \mathscr{G}:=(g_0,\dots,g_{1-r}).$$

The assumption yields the existence of a sequence $(\mathscr{W}_j)\in\ell^2$ to (59), satisfying

$$\frac{|z|-1}{|z|}\sum_{j\geq 1}|\mathscr{W}_j|^2+|\mathscr{W}_1|^2\leq C_R\left\{\frac{|z|}{|z|-1}\sum_{j\geq 1}|\mathscr{F}_j|^2+|\mathscr{G}|^2\right\},$$

with a constant C_R that only depends on R. The above definition of the source terms $(\mathscr{F}_j),\mathscr{G}$ gives[10]

$$\frac{|z|-1}{|z|}\sum_{j\geq 1}|\mathscr{W}_j|^2+|\mathscr{W}_1|^2\leq C_R'\left\{\frac{|z|}{|z|-1}\sum_{j\geq 1}|F_j|^2+\sum_{j=1-r}^{0}|g_j|^2\right\}.$$

Using the decomposition $\mathscr{W}_j=(\mathscr{W}_j^{(1)},\dots,\mathscr{W}_j^{(p+r)})$ as well as the notation $W_j:=\mathscr{W}_{j+r}^{(p+r)}$, $j\geq 1-r$, we can check that $(W_j)\in\ell^2$ is a solution to the resolvent equation (37) and that it satisfies

$$\frac{|z|-1}{|z|}\sum_{j\geq 1-r}|W_j|^2+\sum_{j=1-r}^{p}|W_j|^2\leq C_R'\left\{\frac{|z|}{|z|-1}\sum_{j\geq 1}|F_j|^2+\sum_{j=1-r}^{0}|g_j|^2\right\}.$$

Again, we can also verify that such a solution (W_j) in ℓ^2 is necessarily unique (because the solution to (59) is unique in ℓ^2). The details are left to the reader. Theorem 3.4 completes the argument. □

Remark 3.3. *The result of Proposition 3.1 explains why in Definition 3.1 we have considered the trace estimate*

$$\sum_{n\geq s+1}\sum_{j=1-r}^{p}\Delta t\,\mathrm{e}^{-2\gamma n\Delta t}|U_j^n|^2$$

in the left-hand side of (34). The main purpose for doing so is to obtain the term $|\mathscr{W}_1|^2$ in the left-hand side of the estimate (60) in Proposition 3.1. Obtaining such an estimate is possible only if in the characterization of Theorem 3.3 or Theorem 3.4, the estimate for the resolvent equation involves $|W_{1-r}|^2+\cdots+|W_p|^2$ in the left-hand side and not only $|W_{1-r}|^2+\cdots+|W_0|^2$. If we had kept the definition of strong stability in [10], the left-hand side of (60) would have involved $|\Pi\mathscr{W}_1|^2$ instead of $|\mathscr{W}_1|^2$, where Π would be the projection from $\mathbb{C}^{N(p+r)}$ to $\mathbb{C}^{Nr}$ that retains the last Nr components.

[10]Here we observe that it is crucial to consider a bounded parameter z, because otherwise we could not use a uniform bound for $|\mathbb{A}_p(z)^{-1}|$. This is the main reason why we have proved Theorem 3.4, because Theorem 3.3 would not have been sufficient. It is also crucial that the norm $|\mathbb{A}_p(z)^{-1}|$ remains bounded as z approaches $\mathbb{S}^1$, which amounts to assuming that $\mathbb{A}_p(z)$ is invertible not only on $\mathscr{U}$ but on $\overline{\mathscr{U}}$ (same argument as Lemma 3.4).

Remark 3.4. *The definition of* $\mathbb{M}(z)$ *in* (57) *only depends on the fulfillment of Assumption 3.1 and not on the integer* q*. We could have defined* $\mathbb{M}(z)$ *in the same way even if* q *had not been smaller than* p*.*

We now examine the case $q \geq p$. There is a slight modification to make here. If we wish to rewrite the boundary conditions of (56) as a linear system for some augmented vector $\mathscr{W}_1$, then the coordinates of $\mathscr{W}_1$ should involve $W_{1-r}, \dots, W_{q+1}$, and $q+1 > p$. However we can still write the resolvent equation under a form similar to (59) up to defining

$$\forall\, z \in \mathscr{V}\,, \quad \widetilde{\mathbb{M}}(z) := \begin{pmatrix} -\mathbb{A}_p(z)^{-1}\,\mathbb{A}_{p-1}(z) & \dots & -\mathbb{A}_p(z)^{-1}\,\mathbb{A}_{-r}(z) & 0 & \dots & 0 \\ I & 0 & \dots & \dots & & 0 \\ 0 & & & & 0 & 0 \\ 0 & 0 & \dots & 0 & I & 0 \end{pmatrix} \in \mathscr{M}_{N(q+r+1)}(\mathbb{C})\,. \tag{65}$$

We also define the matrix that encodes the boundary conditions for (56) and the first $q+1-p$ steps in the induction, namely

$$\forall\, z \in \mathbb{C}\setminus\{0\}\,, \quad \widetilde{\mathbb{B}}(z) := \begin{pmatrix} -\mathbb{B}_{q,0}(z) & \dots & -\mathbb{B}_{0,0}(z) & I & & 0 \\ \vdots & & \vdots & & \ddots & \\ -\mathbb{B}_{q,1-r}(z) & \dots & -\mathbb{B}_{0,1-r}(z) & 0 & & I \\ 0 & & \mathbb{A}_p(z) & \dots & \dots & \mathbb{A}_{-r}(z) \\ & \ddots & & & \ddots & \\ \mathbb{A}_p(z) & \dots & \dots & \mathbb{A}_{-r}(z) & & 0 \end{pmatrix} \in \mathscr{M}_{N(q+1-p+r),N(q+1+r)}(\mathbb{C})\,, \tag{66}$$

with the $\mathbb{B}_{\ell,j}$'s defined in (55). With such definitions, it is a simple exercise to rewrite the resolvent equation (56) as an induction relation for the augmented vector $\mathscr{W}_j := (W_{j+q}, \dots, W_{j-r}) \in \mathbb{C}^{N\,(q+1+r)}$, $j \geq 1$. This induction relation takes the form

$$\begin{cases} \mathscr{W}_{j+1} = \widetilde{\mathbb{M}}(z)\,\mathscr{W}_j + \mathscr{F}_j\,, \quad j \geq 1\,, \\ \widetilde{\mathbb{B}}(z)\,\mathscr{W}_1 = \mathscr{G}\,, \end{cases} \tag{67}$$

where the new source terms $(\mathscr{F}_j), \mathscr{G}$ in (67) are given by:

$$\mathscr{F}_j := (\mathbb{A}_p(z)^{-1}\,F_{j+q+1-p}, 0, \dots, 0)\,, \quad \mathscr{G} := (g_0, \dots, g_{1-r}, F_1, \dots, F_{q+1-p})\,.$$

We can then obtain a result that is analogous to Proposition 3.1. The result is just slighlty more complicated because of the definition (66) of the matrix $\widetilde{\mathbb{B}}(z)$ but the proof follows exactly the same arguments.

Proposition 3.2 (Characterization of strong stability)**.** *Let Assumption 3.1 be satisfied, and let us assume* $q \geq p$*. Then the scheme* (32) *is strongly stable in the sense of Definition 3.1 if and only if for all* $R \geq 2$*, there exists a constant* $C_R > 0$ *such that for all* $z \in \mathscr{U}$ *with* $|z| \leq R$*, for all* $(\mathscr{F}_j) \in \ell^2$ *and for all* $\mathscr{G} \in \mathbb{C}^{N(q+1-p+r)}$*, the equation* (67) *has a unique solution* $(\mathscr{W}_j) \in \ell^2$ *and this solution satisfies*

$$\frac{|z|-1}{|z|} \sum_{j\geq 1} |\mathscr{W}_j|^2 + |\mathscr{W}_1|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |\mathscr{F}_j|^2 + \frac{|z|}{|z|-1}\,|\mathscr{G}_{II}|^2 + |\mathscr{G}_I|^2 \right\}, \tag{68}$$

where we use the decomposition $\mathscr{G} = (\mathscr{G}_I, \mathscr{G}_{II})$*,* $\mathscr{G}_I \in \mathbb{C}^{Nr}$*,* $\mathscr{G}_{II} \in \mathbb{C}^{N(q+1-p)}$*.*

Proof of Proposition 3.2. • Let us assume that the scheme (32) is strongly stable, so we can use the result of Theorem 3.4. Let $R \geq 2$, let $z \in \mathscr{U}$ with $|z| \leq R$, and let $(\mathscr{F}_j) \in \ell^2$, $\mathscr{G} \in \mathbb{C}^{N(q+1-p+r)}$. The source terms are decomposed as $\mathscr{F}_j = (\mathscr{F}_j^{(1)}, \dots, \mathscr{F}_j^{(q+1+r)})$, $\mathscr{G}_I = (\mathscr{G}^{(0)}, \dots, \mathscr{G}^{(1-r)}) \in \mathbb{C}^{Nr}$, $\mathscr{G}_{II} = (\mathscr{G}^{(1)}, \dots, \mathscr{G}^{(q+1-p)}) \in \mathbb{C}^{N(q+1-p)}$. We are looking for a solution $\mathscr{W}_j = (\mathscr{W}_j^{(1)}, \dots, \mathscr{W}_j^{(q+1+r)})$

in ℓ^2 to (67). Using the notation $W_j := \mathscr{W}_{j+r}^{(q+1+r)}$, $j \geq 1-r$, we should necessarily have

$$\forall\, \ell = -r, \dots, q\,, \quad \forall\, j \geq 1\,, \quad \mathscr{W}_j^{(q+1-\ell)} = W_{j+\ell} - \sum_{k=-r}^{\ell-1} \mathscr{F}_{j+\ell-1-k}^{(q+1-k)}\,. \tag{69}$$

Moreover, the sequence $(W_j)_{j\geq 1-r}$ should be a solution to (56) with source terms $(F_j), g_{1-r}, \dots, g_0$ defined by

$$\forall\, j \geq q+2-p\,, \quad F_j := \sum_{\ell=0}^{p+r} \mathbb{A}_{p-\ell}(z) \sum_{k=-r}^{q-\ell} \mathscr{F}_{j+p-1-\ell-k}^{(q+1-k)}\,, \tag{70}$$

$$\forall\, j = 1, \dots, q+1-p\,, \quad F_j := \mathscr{G}^{(j)} + \sum_{\ell=-r}^{p} \mathbb{A}_\ell(z) \sum_{k=-r}^{j+\ell-2} \mathscr{F}_{j+\ell-1-k}^{(q+1-k)}\,, \tag{71}$$

$$\forall\, j = 1-r, \dots, 0\,, \quad g_j := \mathscr{G}^{(j)} + \sum_{k=-r}^{j-2} \mathscr{F}_{j-1-k}^{(q+1-k)} - \sum_{\ell=0}^{q} \mathbb{B}_{\ell,j}(z) \sum_{k=-r}^{\ell-1} \mathscr{F}_{\ell-k}^{(q+1-k)}\,. \tag{72}$$

The relations (70), (71), (72) define a sequence $(F_j) \in \ell^2$ and vectors $g_{1-r}, \dots, g_0$ such that, for a given constant M that does not depend on z, there holds

$$\sum_{j\geq 1} |F_j|^2 \leq M \left(\sum_{j\geq 1} |\mathscr{F}_j|^2 + |\mathscr{G}_{II}|^2 \right)\,, \quad \sum_{j=1-r}^{0} |g_j|^2 \leq M \left(\sum_{j\geq 1} |\mathscr{F}_j|^2 + |\mathscr{G}_I|^2 \right)\,. \tag{73}$$

Applying Theorem 3.4, we know that there exists a unique solution $(W_j) \in \ell^2$ to (56) with the source terms defined in (70), (71), (72), and that for some constant C_R independent of z, there holds

$$\frac{|z|-1}{|z|} \sum_{j\geq 1-r} |W_j|^2 + \sum_{j=1-r}^{p} |W_j|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}\,.$$

The relation (69) then defines a sequence $(\mathscr{W}_j)_{j\geq 1} \in \ell^2$ which is a solution to (67). Combining (69), (73) and the estimate of (W_j), we get

$$\frac{|z|-1}{|z|} \sum_{j\geq 1} |\mathscr{W}_j|^2 + \sum_{\ell=-r}^{p-1} \left| \mathscr{W}_1^{(q+1-\ell)} \right|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |\mathscr{F}_j|^2 + \frac{|z|}{|z|-1} |\mathscr{G}_{II}|^2 + |\mathscr{G}_I|^2 \right\}\,.$$

In order to complete the proof, it only remains to estimate the sum

$$\sum_{\ell=p}^{q} \left| \mathscr{W}_1^{(q+1-\ell)} \right|^2\,.$$

This is done with an induction argument based on the relations

$$\forall\, j = 0, \dots, q-p\,, \quad \sum_{\ell=-r}^{p} \mathbb{A}_\ell(z)\, \mathscr{W}_1^{(q+1-\ell-j)} = \mathscr{G}_1^{(1+j)}\,,$$

and the fact that $\mathbb{A}_p(z)$ is invertible for all $z \in \overline{\mathscr{U}}$. The details are left to the reader. Eventually, we obtain an estimate of the form

$$|\mathscr{W}_1|^2 = \sum_{\ell=-r}^{q} \left| \mathscr{W}_1^{(q+1-\ell)} \right|^2 \leq C_R' \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |\mathscr{F}_j|^2 + \frac{|z|}{|z|-1} |\mathscr{G}_{II}|^2 + |\mathscr{G}_I|^2 \right\}\,.$$

The uniqueness of the solution $(\mathscr{W}_j) \in \ell^2$ to (67) is proved by entirely similar arguments to those used in the proof of Proposition 3.1. We feel free at this point to skip the details.

• Let us now assume that (67) has a unique solution in ℓ^2 for all source terms $(\mathscr{F}_j) \in \ell^2$ and $\mathscr{G}$ together with the estimate (68). Let $R \geq 2$, let $z \in \mathscr{U}$ with $|z| \leq R$, and let us consider some source terms $(F_j) \in \ell^2$, $g_{1-r}, \dots, g_0 \in \mathbb{C}^N$ for (37). We define the vectors

$$\mathscr{F}_j := (\mathbb{A}_p(z)^{-1}\, F_{q+1-p+j}, 0, \dots, 0)\,, \quad \mathscr{G} := (g_0, \dots, g_{1-r}, F_1, \dots, F_{q+1-p})\,.$$

The assumption yields the existence of a sequence $(\mathscr{W}_j) \in \ell^2$ to (67), satisfying

$$\frac{|z|-1}{|z|} \sum_{j\geq 1} |\mathscr{W}_j|^2 + |\mathscr{W}_1|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |\mathscr{F}_j|^2 + \frac{|z|}{|z|-1} |\mathscr{G}_{II}|^2 + |\mathscr{G}_I|^2 \right\}.$$

The definition of the source terms $(\mathscr{F}_j), \mathscr{G}$ gives

$$\frac{|z|-1}{|z|} \sum_{j\geq 1} |\mathscr{W}_j|^2 + |\mathscr{W}_1|^2 \leq C'_R \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

Using the decomposition $\mathscr{W}_j = (\mathscr{W}_j^{(1)}, \dots, \mathscr{W}_j^{(q+1+r)})$ as well as the notation $W_j := \mathscr{W}_{j+r}^{(q+1+r)}$, $j \geq 1-r$, $(W_j) \in \ell^2$ is a solution to (37) that satisfies

$$\frac{|z|-1}{|z|} \sum_{j\geq 1-r} |W_j|^2 + \sum_{j=1-r}^{q+1} |W_j|^2 \leq C'_R \left\{ \frac{|z|}{|z|-1} \sum_{j\geq 1} |F_j|^2 + \sum_{j=1-r}^{0} |g_j|^2 \right\}.$$

Such a solution in ℓ^2 to (37) is necessarily unique, and Theorem 3.4 completes the argument. □

Remark 3.5. *Unlike what happened in the case $q < p$ with the definition (58) of the matrix $\mathbb{B}(z)$, it is no longer clear that the matrix $\widetilde{\mathbb{B}}(z)$ in (66) has maximal rank (this was uncorrectly claimed in [4]). However, the result of Proposition 3.2 shows that if the scheme (32) is strongly stable, then $\widetilde{\mathbb{B}}(z)$ should have maximal rank for all $z \in \mathscr{U}$ (use Proposition 3.2 with $\mathscr{F}_j = 0$ for all j and an arbitrary $\mathscr{G}$). A refined version of this result is stated in the following paragraph.*

3.4. **Characterization of strong stability: the main result.** Up to now, we have characterized strong stability in terms of an estimate for the resolvent equation (37), or for the equivalent formulations (59) or (67). We have also seen that a necessary condition for strong stability is the so-called Godunov-Ryabenkii condition of Corollary 3.1, which is an analogue of the Lopatinskii condition for hyperbolic initial boundary value problems. In this paragraph, we make a little more precise this necessary condition for strong stability. It will turn out that this refined necessary condition will also be sufficient for strong stability. Readers who are familiar with the theory of hyperbolic initial boundary value problems will recognize the gap between the Lopatinskii condition and the uniform Lopatinskii condition, see [2, chapter 4]. The gap here between the Godunov-Ryabenkii condition and what we shall call the uniform Kreiss-Lopatinskii condition below is entirely analogous.

Let us begin with a fundamental property of the matrices $\mathbb{M}(z)$ in (57) and $\widetilde{\mathbb{M}}(z)$ in (65). We recall that the operators Q_σ that appear in (32) and whose expression is given in (33) correspond to a discretization of the hyperbolic operator. According to the analysis of Section 2, see in particular Propositions 2.1 and 2.2, stability for the discrete Cauchy problem is encoded in the uniform power boundedness of the amplification matrix $\mathscr{A}(\mathrm{e}^{i\eta})$, $\eta \in \mathbb{R}$. To encompass both situations $s = 0$ and $s \geq 1$, we shall always refer to the discrete Cauchy problem as to problem (14), with the operators Q_σ as in (33) or (15). The amplification matrix $\mathscr{A}$ is then defined in (16) as a (block) companion matrix. When s equals 0, this definition reduces to (11). The fundamental property of $\mathbb{M}(z)$ is stated as follows.

Lemma 3.7 (Stable eigenvalues of $\mathbb{M}(z)$ [12])**.** *Let Assumption 3.1 be satisfied, and let us assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2. Then for all $z \in \mathscr{V}$, the eigenvalues of the matrix $\mathbb{M}(z)$ in (57) are those $\kappa \in \mathbb{C} \setminus \{0\}$ such that*

$$\det\left(\mathscr{A}(\kappa) - z\, I\right) = 0\,.$$

In particular for all $z \in \mathscr{U}$, $\mathbb{M}(z)$ has no eigenvalue on the unit circle $\mathbb{S}^1$ and the number of eigenvalues in $\mathbb{D}$ equals $N\,r$ (eigenvalues are counted with their algebraic multiplicity).

We emphasize that there is no condition on the integer q in Lemma 3.7 because the definition of $\mathbb{M}(z)$ is independent of q, see Remark 3.4.

Proof of Lemma 3.7. The matrix $\mathbb{M}(z)$ in (57) is defined on the open neighborhood $\mathscr{V} = \{z \in \mathbb{C}\,,\ |z| > 1-\varepsilon_0\}$ of $\overline{\mathscr{U}}$. On $\mathscr{V}$, both matrices $\mathbb{A}_{-r}(z)$ and $\mathbb{A}_p(z)$ are invertible thanks to Assumption 3.1. Let now $z \in \mathscr{V}$, and let $X = (X_1, \dots, X_{p+r}) \in \mathbb{C}^{N(p+r)}$ belong to the kernel of $\mathbb{M}(z)$. Using the expression (57) of $\mathbb{M}(z)$, we get

$$X_1 = \cdots = X_{p+r-1} = 0\,, \quad \mathbb{A}_p(z)^{-1}\,\mathbb{A}_{-r}(z)\,X_{p+r} = 0\,,$$

so the kernel of $\mathbb{M}(z)$ is reduced to $\{0\}$. In particular, the eigenvalues of $\mathbb{M}(z)$ are nonzero. We are now going to obtain some more precise information on these eigenvalues.

Applying some standard rules for determinants of block companion matrices (use Schur's complement formula, see e.g. [21]), we obtain for all $z \in \mathscr{V}$ and all $\kappa \neq 0$:

$$\begin{aligned}\det(\mathbb{M}(z) - \kappa\, I) &= (-1)^{N(p+r-1)}\ \det\left[-\sum_{\ell=-r}^{p} \kappa^{\ell+r}\,\mathbb{A}_p(z)^{-1}\,\mathbb{A}_\ell(z) - \kappa^{p+r}\, I\right] \\ &= (-1)^{N(p+r)}\,\kappa^{Nr}\ \det \mathbb{A}_p(z)^{-1}\ \det\left[\sum_{\ell=-r}^{p} \kappa^{\ell}\,\mathbb{A}_\ell(z)\right]. \end{aligned} \tag{74}$$

In the same way, we compute

$$\begin{aligned}\det(\mathscr{A}(\kappa) - z\, I) &= (-1)^{Ns}\ \det\left[\sum_{\sigma=0}^{s} z^{s-\sigma}\,\widehat{Q_\sigma}(\kappa) - z^{s+1}\, I\right] \\ &= (-1)^{N(s+1)}\,z^{N(s+1)}\ \det\left[\sum_{\ell=-r}^{p} \kappa^{\ell}\,\mathbb{A}_\ell(z)\right],\end{aligned}$$

where the amplification matrix $\mathscr{A}$ is defined in (16). In other words, for $z \in \mathscr{V}$ and $\kappa \neq 0$, $\det(\mathbb{M}(z) - \kappa\, I)$ and $\det(\mathscr{A}(\kappa) - z\, I)$ vanish simultaneously. This proves the first part of Lemma 3.7.

Let now $z \in \mathscr{U}$. Let us assume that $\kappa \in \mathbb{S}^1$ is an eigenvalue of $\mathbb{M}(z)$. Then z is an eigenvalue of $\mathscr{A}(\kappa)$. However, stability for the discrete Cauchy problem (14) implies that the von Neumann condition is satisfied, see Corollary 2.1, so the spectral radius of $\mathscr{A}(\kappa)$ is not larger than 1. We are led to a contradiction. By a continuity/connectedness argument, the number of eigenvalues of $\mathbb{M}(z)$ in $\mathbb{D}$ is independent of $z \in \mathscr{U}$. We are now going to show that this number equals $N\,r$. The idea is to study the behavior of eigenvalues of $\mathbb{M}(z)$ as z tends to infinity.

Let us first show that as z tends to infinity, the eigenvalues of $\mathbb{M}(z)$ which belong to $\mathbb{D}$ converge to 0. For otherwise, there would exist $\varepsilon > 0$, a sequence $(z_n)_{n\geq 1}$ with $|z_n| > n$, and a sequence $(\kappa_n)_{n\geq 1}$ such that

$$\forall\, n \geq 1\,, \quad \varepsilon \leq |\kappa_n| < 1\,, \quad \kappa_n \in \mathrm{sp}(\mathbb{M}(z_n))\,.$$

Applying the formula (74), we have

$$\forall\, n \geq 1\,, \quad \det\left[\sum_{\ell=-r}^{p} \kappa_n^{\ell}\,\mathbb{A}_\ell(z_n)\right] = 0\,. \tag{75}$$

Up to extracting a subsequence, we can assume that (κ_n) converges towards κ_∞ which satisfies $\varepsilon \leq |\kappa_\infty| \leq 1$ (in particular, $\kappa_\infty \neq 0$). Recalling the definition (54) and passing to the limit in (75), we obtain $\det I = 0$ which is a contradiction. We have thus proved that for large $|z|$, the eigenvalues of $\mathbb{M}(z)$ which belong to $\mathbb{D}$ are arbitrarily close to 0.

To complete the proof, we introduce the function

$$D(\kappa, Z) := \det\left[\sum_{\ell=-r}^{p} \kappa^{r+\ell}\,\mathbb{A}_\ell(1/Z)\right].$$

According to the definition (54) of the matrices $\mathbb{A}_\ell$, D is a polynomial function of (κ, Z). Moreover, we have $D(\kappa, 0) = \kappa^{Nr}$. This shows that for all $Z \neq 0$ sufficiently small, the polynomial $D(\cdot, Z)$ has exactly $N\,r$ roots (counted with their multiplicity) which are close to 0. (This is a direct application of Rouché's Theorem for holomorphic functions.) Then the formula (74) shows that for large $|z|$, $\mathbb{M}(z)$ has $N\,r$ eigenvalues which are close to 0. Since all eigenvalues of $\mathbb{M}(z)$ in $\mathbb{D}$ must be close to 0, we have proved that for all $z \in \mathscr{U}$, $\mathbb{M}(z)$ has exactly $N\,r$ eigenvalues in $\mathbb{D}$. □

The eigenvalues of $\mathbb{M}(z)$ in $\mathbb{D}$ are called *stable eigenvalues* since they correspond to geometrically decreasing sequences (hence in ℓ^2) that are solutions to the induction relation

$$\mathscr{W}_{j+1} = \mathbb{M}(z)\,\mathscr{W}_j\,, \quad j \geq 1\,.$$

At the opposite, eigenvalues of $\mathbb{M}(z)$ in $\mathscr{U}$ will be called *unstable eigenvalues* since they correspond to sequences whose norm diverges geometrically.

Our proof of Lemma 3.7 follows [12] where the same result is proved in the case $s = 0$. Unlike what is stated in [10], the number of eigenvalues of $\mathbb{M}(z)$ in $\mathbb{D}$ has nothing to do with the boundary conditions in (32). As a matter of fact, the definition of $\mathbb{M}(z)$ only involves the matrices $A_{\ell,\sigma}$ and is completely independent of the matrices $B_{\ell,j,\sigma}$, see (57). In the same way, the definition (65) of $\widetilde{\mathbb{M}}(z)$ only involves the matrices $A_{\ell,\sigma}$.

The matrix $\widetilde{\mathbb{M}}(z)$ defined in (65) and used to rewrite the resolvent equation in the case $q \geq p$ satisfies analogous properties to those stated in Lemma 3.7.

Lemma 3.8 (Stable eigenvalues of $\widetilde{\mathbb{M}}(z)$)**.** *Let Assumption 3.1 be satisfied, let us assume $q \geq p$ and let us further assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2. Then for all $z \in \mathscr{V}$, the eigenvalues of $\widetilde{\mathbb{M}}(z)$ are 0 - with algebraic multiplicity $N\,(q+1-p)$ - and the eigenvalues of the matrix $\mathbb{M}(z)$ (eigenvalues are counted with their algebraic multiplicity).*

In particular for all $z \in \mathscr{U}$, $\widetilde{\mathbb{M}}(z)$ has no eigenvalue on the unit circle $\mathbb{S}^1$ and the number of eigenvalues in $\mathbb{D}$ equals $N\,(q+1-p+r)$.

Proof of Lemma 3.8. With the result of Lemma 3.7, the proof is now straightforward (we recall that Lemma 3.7 holds independently of q). Indeed, for $z \in \mathscr{V}$ and $\kappa \in \mathbb{C}$, we compute

$$\begin{aligned}\det(\widetilde{\mathbb{M}}(z) - \kappa\, I) &= (-1)^{N(q+1+r)}\, \det\left[\sum_{\ell=1}^{p+r} \kappa^{q+1+r-\ell}\, \mathbb{A}_p(z)^{-1}\, \mathbb{A}_{p-\ell}(z) + \kappa^{q+1+r}\, I\right] \\ &= (-1)^{N(q+1+r)}\, \det \mathbb{A}_p(z)^{-1}\, \kappa^{N(q+1-p)}\, \det\left[\sum_{\ell=-r}^{p} \kappa^{r+\ell}\, \mathbb{A}_\ell(z)\right]\,.\end{aligned}$$

Since $\mathbb{A}_{-r}(z)$ is invertible, the latter equality shows that 0 is a root with multiplicity $N\,(q+1-p)$ of the characteristic polynomial of $\widetilde{\mathbb{M}}(z)$. Moreover, the relation (74) shows that the nonzero eigenvalues of $\widetilde{\mathbb{M}}(z)$ are exactly the eigenvalues of $\mathbb{M}(z)$ and the algebraic multiplicities coincide. The result of Lemma 3.8 follows. □

The results of Lemma 3.7 and Lemma 3.8 imply the following necessary conditions for strong stability in the cases $q < p$ and $q \geq p$.

Corollary 3.2 (The uniform Kreiss-Lopatinskii condition in the case $q < p$)**.** *Let Assumption 3.1 be satisfied, let us assume $q < p$ and let us further assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2. If the scheme (32) is strongly stable in the sense of Definition 3.1, then for all $R \geq 2$, there exists a constant $C_R > 0$ such that for all $z \in \mathscr{U}$ with $|z| \leq R$, there holds*

$$\forall\, \mathscr{W} \in \mathbb{E}^s(z)\,, \quad |\mathscr{W}| \leq C_R\, |\mathbb{B}(z)\,\mathscr{W}|\,, \tag{76}$$

where $\mathbb{E}^s(z)$ denotes the generalized eigenspace of the matrix $\mathbb{M}(z)$ associated with eigenvalues in $\mathbb{D}$, and where the matrix $\mathbb{B}(z)$ is defined in (58).

In other words, if the scheme (32) is strongly stable, then the mapping

$$\Phi(z)\ :\ \mathscr{W} \in \mathbb{E}^s(z) \longmapsto \mathbb{B}(z)\,\mathscr{W} \in \mathbb{C}^{Nr}\,,$$

is an isomorphism for all $z \in \mathscr{U}$. Moreover for all $R \geq 2$, the inverse $\Phi(z)^{-1}$ is uniformly bounded with respect to $z \in \mathscr{U}$, $|z| \leq R$.

Proof of Corollary 3.2. The proof is very easy. Let $R \geq 2$, and let $z \in \mathscr{U}$ with $|z| \leq R$. According to the assumptions, we can apply both Propositions 3.1 and Lemma 3.7. Let $\mathscr{W} \in \mathbb{E}^s(z)$. The

sequence $(\mathscr{W}_j)_{j\geq 1}$ defined by

$$\begin{cases} \mathscr{W}_{j+1} = \mathbb{M}(z)\,\mathscr{W}_j\,, \quad j\geq 1\,, \\ \mathscr{W}_1 := \mathscr{W}\,, \end{cases}$$

belongs to ℓ^2 (it converges towards 0 geometrically as j tends to $+\infty$) and it is a solution to

$$\begin{cases} \mathscr{W}_{j+1} = \mathbb{M}(z)\,\mathscr{W}_j\,, \qquad j\geq 1\,, \\ \mathbb{B}(z)\,\mathscr{W}_1 = \mathbb{B}(z)\,\mathscr{W}\,. \end{cases}$$

Then the estimate (60) for solutions to (59) yields (76). Lemma 3.7 shows that the *stable subspace* $\mathbb{E}^s(z)$ has dimension $N\,r$ so the linear mapping $\Phi(z)$ defined in Corollary 3.2 is an isomorphism (it is injective and the spaces have equal dimension). The estimate (76) shows that the norm of $\Phi(z)^{-1}$ remains uniformly bounded as $z\in\mathscr{U}$ approaches the unit circle. □

From Corollary 3.2, we see that the scheme (32) could not have been strongly stable if $\mathbb{B}(z)$ had not had maximal rank. Hopefully, this maximal rank property is obvious here, see Remark 3.2.

There is of course a similar result in the case $q\geq p$. We feel free to skip the proof.

Corollary 3.3 (The uniform Kreiss-Lopatinskii condition in the case $q\geq p$)**.** *Let Assumption 3.1 be satisfied, let us assume $q\geq p$ and let us further assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2. Let us decompose the matrix $\widetilde{\mathbb{B}}(z)$ in (66) as*

$$\forall\, z\in\mathbb{C}\setminus\{0\}\,, \quad \widetilde{\mathbb{B}}(z) = \begin{pmatrix} \mathbb{B}_\sharp(z) \\ \mathbb{B}_\flat(z) \end{pmatrix}\,, \quad \mathbb{B}_\sharp(z)\in\mathscr{M}_{Nr,N(q+1+r)}(\mathbb{C})\,, \quad \mathbb{B}_\flat(z)\in\mathscr{M}_{N(q+1-p),N(q+1+r)}(\mathbb{C})\,.$$

If the scheme (32) is strongly stable in the sense of Definition 3.1, then for all $R\geq 2$, there exists a constant $C_R>0$ such that for all $z\in\mathscr{U}$ with $|z|\leq R$, there holds

$$\forall\,\mathscr{W}\in\widetilde{\mathbb{E}}^s(z)\cap\mathrm{Ker}\;\mathbb{B}_\flat(z)\,, \quad |\mathscr{W}|\leq C_R\,|\mathbb{B}_\sharp(z)\,\mathscr{W}|\,, \tag{77}$$

where $\widetilde{\mathbb{E}}^s(z)$ denotes the generalized eigenspace of the matrix $\widetilde{\mathbb{M}}(z)$ associated with eigenvalues in $\mathbb{D}$.

It is not very hard to show that the space $\widetilde{\mathbb{E}}^s(z)\cap\mathrm{Ker}\;\mathbb{B}_\flat(z)$ is isomorphic to the stable subspace $\mathbb{E}^s(z)$ of $\mathbb{M}(z)$ and thus has dimension $N\,r$ for all $z\in\mathscr{U}$. Moreover, the matrix $\mathbb{B}_\sharp(z)$ has rank $N\,r$ for all $z\in\mathbb{C}\setminus\{0\}$. Hence the estimate (77) is not ruled out by obvious dimensions reasons (for instance if the rank of $\mathbb{B}_\sharp(z)$ had been smaller than $N\,r$).

Let us also observe that if the estimate (77) holds, then the mapping

$$\widetilde{\Phi}(z)\;:\;\mathscr{W}\in\widetilde{\mathbb{E}}^s(z)\longmapsto\widetilde{\mathbb{B}}(z)\,\mathscr{W}\in\mathbb{C}^{N(q+1-p+r)}\,,$$

is injective, so it is an isomorphism. In particular, $\widetilde{\mathbb{B}}(z)$ has maximal rank for all $z\in\mathscr{U}$. Again, this maximal rank property is a necessary condition for strong stability.

Remark 3.6. *We do not know whether the terminology "uniform Kreiss-Lopatinskii condition" is really standard in the context of finite difference schemes (probably "uniform Godunov-Ryabenkii condition" might be more appropriate). Our goal here is to emphasize the link between this condition and the analogous necessary condition for well-posedness for hyperbolic initial boundary value problems.*

As we shall see below, the vector space $\mathbb{E}^s(z)$ varies continuously - and even holomorphically - with respect to $z\in\mathscr{U}$. Another way to rephrase Corollary 3.2 is therefore: for all $z\in\mathscr{U}$, $\mathbb{E}^s(z)\cap\mathrm{Ker}\;\mathbb{B}(z)=\{0\}$, that is, $\mathbb{C}^{N(p+r)}=\mathbb{E}^s(z)\oplus\mathrm{Ker}\;\mathbb{B}(z)$. Moreover, for all $1<R_1\leq R_2$, the quantity

$$\sup_{R_1\leq|z|\leq R_2}\;\sup_{\mathscr{W}\in\mathbb{E}^s(z)\setminus\{0\}}\;\frac{|\mathscr{W}|}{|\mathbb{B}(z)\,\mathscr{W}|}\,,$$

remains bounded as R_1 tends to 1 and R_2 remains fixed.

The Godunov-Ryabenkii condition shows that the latter quantity is finite for all $1<R_1\leq R_2$, but it does not give any information on how this quantity varies as R_1 approaches 1. Some examples for which the uniform Kreiss-Lopatinskii condition is not satisfied show that this quantity may be unbounded as R_1 tends to 1 (see later on in these notes for the case of the Lax-Friedrichs and leap-frog schemes).

The estimate (76), or (77), is a necessary condition for strong stability. The injectivity of the linear mapping $\Phi(z)$ in Corollary 3.2 can be tested by first determining a basis $(e_1(z), \dots, e_{Nr}(z))$ of $\mathbb{E}^s(z)$, and by computing the associated (Lopatinskii) $N\,r \times N\,r$ determinant

$$\Delta(z) := \det \left[\mathbb{B}(z)\, e_1(z) \ \dots \ \mathbb{B}(z)\, e_{Nr}(z) \right].$$

The vanishing of $\Delta(z)$ is independent of the choice of the basis. The Godunov-Ryabenkii condition holds true if and only if Δ does not vanish on $\mathscr{U}$. Some examples of computations of such determinants are given a little further in these notes for the Lax-Friedrichs and leap-frog schemes with various choices of numerical boundary conditions. However, the reader will understand that computing such determinants is not always possible from a practical point of view. For instance, one numerical scheme based on the Runge-Kutta method and presented in Section 2 corresponded to $r = 8$, and it becomes impossible to compute stable eigenvalues in this case. Numerical strategies are necessary to compute the stable subspace and the Lopatinskii determinant.

In the spirit of [10], our main result shows that the uniform Kreiss-Lopatinskii condition (meaning the fulfillment of the estimate (76) or (77) according to the sign of $q - p$) is not only a necessary condition for strong stability but is also a **sufficient** condition. Our result requires however a structural assumption on the operators Q_σ, namely the property of geometric regularity introduced in Section 2. More precisely, our main result in the case $q < p$ reads as follows.

Theorem 3.5 (Main result for $q < p$)**.** *Let Assumption 3.1 be satisfied, let us assume $q < p$ and let us further assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2 and that the operators Q_σ are geometrically regular in the sense of Definition 2.3. For all $z \in \mathscr{U}$, we let $\mathbb{E}^s(z)$ denote the generalized eigenspace of the matrix $\mathbb{M}(z)$ in (57) associated with eigenvalues in $\mathbb{D}$.*

Then the scheme (32) is strongly stable in the sense of Definition 3.1 **if and only if** *for all $R \geq 2$, there exists a constant $C_R > 0$ such that for all $z \in \mathscr{U}$ with $|z| \leq R$, the estimate (76) holds with the matrix $\mathbb{B}(z)$ defined in (58).*

Our main result in the case $q \geq p$ is similar.

Theorem 3.6 (Main result for $q \geq p$)**.** *Let Assumption 3.1 be satisfied, let us assume $q \geq p$ and let us further assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2 and that the operators Q_σ are geometrically regular in the sense of Definition 2.3. For all $z \in \mathscr{U}$, we let $\widetilde{\mathbb{E}}^s(z)$ denote the generalized eigenspace of the matrix $\widetilde{\mathbb{M}}(z)$ in (65) associated with eigenvalues in $\mathbb{D}$.*

Then the scheme (32) is strongly stable in the sense of Definition 3.1 **if and only if** *for all $R \geq 2$, there exists a constant $C_R > 0$ such that for all $z \in \mathscr{U}$ with $|z| \leq R$, the estimate (77) holds with $\mathbb{B}_\sharp(z), \mathbb{B}_\flat(z)$ as in Corollary 3.3.*

We shall give later on a more practical version of Theorems 3.5 and 3.6, where the fulfillment of the estimates (76) or (77) will be replaced by a purely algebraic condition (see Proposition 4.1 below). However, this new formulation will rely on the continuous extension of the stable subspace $\mathbb{E}^s(z)$ to $\mathbb{S}^1$, which is still not known. Let us now give a few details on the strategy of the proof.

The proof of Theorems 3.5 and 3.6 relies on the construction of symmetrizers for the equivalent forms (59) or (67) of the resolvent equation (37). A symmetrizer is a matrix $\mathbb{S}(z)$ such that when one multiplies (59) or (67) by $\mathscr{W}_{j+1}^*\,\mathbb{S}(z)$ and use summation by parts (also known as Abel's transformation), one more or less ends up with the estimate (60) or (68). A precise definition of symmetrizers is given below (see Definitions 4.2 and 4.3). The crucial point is to understand the construction of the symmetrizer when $z \in \mathscr{U}$ is close to $\mathbb{S}^1$. In particular, a crucial issue in the construction is to understand how the stable subspace $\mathbb{E}^s(z)$, or $\widetilde{\mathbb{E}}^s(z)$, behaves as z approaches $\mathbb{S}^1$. The geometric regularity condition will first enable us to prove that $\mathbb{E}^s(z)$ has a limit as $z \in \mathscr{U}$ tends to a point of $\mathbb{S}^1$. We shall then be able to rephrase the uniform Kreiss-Lopatinskii condition in a more convenient way (Proposition 4.1) and to construct a symmetrizer which depends smoothly on z.

In order to clarify the proof of Theorem 3.5, we first devote some paragraphs to the proof of several results that will be intermediate steps for the whole proof. Each step may have its own interest, so we feel that cutting the proof into several "small" pieces is more appropriate. It also

clarifies where the assumptions of Theorem 3.5 are needed. There are more or less four main steps in the proof of Theorem 3.5 (the proof of Theorem 3.6 follows exactly the same strategy):

(1) Reducing the matrix $\mathbb{M}(z)$ in (57) to a convenient block diagonal form, that is, showing that $\mathbb{M}(z)$ satisfies the so-called *discrete block structure condition* defined below (see Definition 4.1). The analysis closely follows [14] and [16, appendix C]. This step is a refined version of the analysis in [10].

(2) Constructing a symmetrizer for each block in the reduction of $\mathbb{M}(z)$. This part of the proof requires the analysis of quite many cases, which correspond to the possible behaviors of eigenvalues for the amplification matrix associated with geometrically regular operators. This is where the analysis and the examples of Section 2 will be useful and this is actually the main reason why we have given so many examples in Section 2. This part of the proof is the main novelty compared with [10] since we are able to cover here all the possible cases while only two of them were allowed in [10]. In particular, the theory developed in [10] could not cover the singular behaviors displayed in Figures 2 and 4.

(3) Showing that the existence of a symmetrizer implies that the stable subspace extends continuously to $z \in \mathbb{S}^1$, and thus reformulating the uniform Kreiss-Lopatinskii condition. This part of the proof is inspired from [15].

(4) Proving energy estimates for the equivalent formulation (59) of the resolvent equation. This part of the proof already appeared in [10] and there is no modification here.

In what follows, we shall deal with the three first steps of the proof as if they were independent problems. The main reason for doing so is to clarify which assumptions are needed for each part of the analysis in view of a future extension to multidimensional problems. To avoid repeating many arguments, we shall only give the proof of Theorem 3.5 and leave the proof of Theorem 3.6 to the interested reader. Most of the arguments are the same, in particular the reduction to the discrete block structure and the construction of symmetrizers. Minor modifications need to be done in the final derivation of the a priori estimate and we hope that the reader will be thrilled to find these subtelties hy himself/herself.

4. Characterization of strong stability: proof of the main results

4.1. **The discrete block structure condition.** The aim of this paragraph is to understand to which extent the resolvent equation (59), resp. (67), can be "diagonalized". The goal is more or less to reduce to a set of scalar equations but this is unfortunately not always possible as we shall see below. We begin with the following

Definition 4.1 (Discrete block structure condition)**.** *Let M be a holomorphic function on some open neighborhood of $\overline{\mathscr{U}}$ with values in $\mathscr{M}_m(\mathbb{C})$ for some integer m. Then M is said to satisfy the discrete block structure condition if the two following conditions are satisfied:*

(1) for all $z \in \mathscr{U}$, $\mathrm{sp}(M(z)) \cap \mathbb{S}^1 = \emptyset$,

(2) for all $\underline{z} \in \overline{\mathscr{U}}$, there exists an open neighborhood $\mathscr{O}$ of $\underline{z}$ in $\mathbb{C}$, and there exists an invertible matrix $T(z)$ that is holomorphic with respect to $z \in \mathscr{O}$ such that

$$\forall\, z \in \mathscr{O}\,, \quad T(z)^{-1}\, M(z)\, T(z) = \mathrm{diag}\,\left(M_1(z), \dots, M_L(z)\right),$$

where the number L of diagonal blocks and the size ν_ℓ of each block M_ℓ do not depend on $z \in \mathscr{O}$, and where each block satisfies one of the following properties:

- *there exists $\delta > 0$ such that for all $z \in \mathscr{O}$, $M_\ell(z)^*\, M_\ell(z) \geq (1+\delta)\, I$,*
- *there exists $\delta > 0$ such that for all $z \in \mathscr{O}$, $M_\ell(z)^*\, M_\ell(z) \leq (1-\delta)\, I$,*
- *$\nu_\ell = 1$, $\underline{z}$ and $M_\ell(\underline{z})$ belong to $\mathbb{S}^1$, and $\underline{z}\, M_\ell'(\underline{z})\, \overline{M_\ell(\underline{z})} \in \mathbb{R} \setminus \{0\}$,*
- *$\nu_\ell > 1$, $\underline{z} \in \mathbb{S}^1$ and $M_\ell(\underline{z})$ has the form*

$$M_\ell(\underline{z}) = \underline{\kappa}_\ell \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & 1 \end{pmatrix}, \quad \underline{\kappa}_\ell \in \mathbb{S}^1\,.$$

Moreover the lower left coefficient m_ℓ of $M_\ell'(\underline{z})$ is such that for all $\theta \in \mathbb{C}$ with $\mathrm{Re}\ \theta > 0$, and for all complex number ζ such that $\zeta^{\nu_\ell} = \overline{\underline{\kappa}_\ell}\, m_\ell\, \underline{z}\, \theta$, then $\mathrm{Re}\ \zeta \neq 0$.

We refer to the blocks M_ℓ in the reduction of M as being of the first, second, third or fourth type.

The discrete block structure condition is more precise than the normal form of [10, Theorem 9.1]. Definition 2.2 clarifies the structure of the blocks associated with eigenvalues in $\mathbb{S}^1$. Such blocks are either scalar, which was not clear in [10], or have a "Jordan structure" (blocks of the fourth type). This clarification will simplify the construction of symmetrizers in the following paragraph. Our goal here is to prove the following

Theorem 4.1 (Characterization of the discrete block structure condition [4])**.** *Let Assumption 3.1 be satisfied. Then $\mathbb{M}$ defined by (57) satisfies the discrete block structure condition if and only if the operators Q_σ in (33) are geometrically regular and the discretization (14) is stable in the sense of Definition 2.2.*

Theorem 4.1 is the analogue for finite difference schemes of Theorem C.3 in [16]. The assumptions of Theorem 4.1 allow more general situations than the cases covered by [10]. In particular, we show that assumptions 5.2 and 5.3 in [10] are not necessary to reduce $\mathbb{M}$ to the discrete block structure. Before proving Theorem 4.1, we recall the basic observation that was already discussed in Section 2: the geometric regularity of the operators Q_σ is not a consequence of the stability of (14) (except in some very specific situations, see Lemma 2.7). However, we have seen that many finite difference schemes used to discretize hyperbolic equations satisfy this geometric regularity condition. We therefore believe that Theorem 4.1 applies more or less to all finite difference discretizations of the form (32).

Proof of Theorem 4.1. • Let us start with the "easy" part of the Theorem. We assume here that $\mathbb{M}$ defined by (57) satisfies the discrete block structure condition. Let us first show that the amplification matrix satisfies the von Neumann condition. Let $\kappa \in \mathbb{S}^1$ and let $z \in \mathrm{sp}(\mathscr{A}(\kappa))$. Let us assume $z \in \mathscr{U}$.

Recalling the definition (16), we obtain (the argument is the same as in the proof of Lemma 3.7)

$$0 = \det(\mathscr{A}(\kappa) - z\, I) = (-1)^{Ns} \det\left[\sum_{\sigma=0}^{s} z^{s-\sigma}\, \widehat{Q_\sigma}(\kappa) - z^{s+1}\, I\right]$$
$$= (-1)^{N(s+1)}\, z^{N(s+1)} \det\left[\sum_{\ell=-r}^{p} \kappa^\ell\, \mathbb{A}_\ell(z)\right]. \tag{78}$$

Since κ is nonzero, the relation (74) shows that $\kappa \in \mathbb{S}^1$ is an eigenvalue of $\mathbb{M}(z)$, and $z \in \mathscr{U}$. This is ruled out by the discrete block structure condition (see condition *(1)* in Definition 4.1). In other words, the eigenvalues of $\mathscr{A}(\kappa)$ belong to $\mathbb{D}$ or $\mathbb{S}^1$, so the von Neumann condition (18) is satisfied.

We are now going to prove that the operators Q_σ are geometrically regular. Let $\underline{\kappa} \in \mathbb{S}^1$ and let us assume that $\underline{z} \in \mathbb{S}^1$ is an eigenvalue of $\mathscr{A}(\underline{\kappa})$ with algebraic multiplicity $\underline{\alpha}$. The same argument as above based on relation (74) shows that $\underline{\kappa}$ is an eigenvalue of $\mathbb{M}(\underline{z})$. We apply property *(2)* of the discrete block structure condition at the point $\underline{z}$: there exists an open neighborhood $\mathscr{O}$ of $\underline{z}$ in $\mathbb{C}$, and there exists an invertible matrix $T(z)$ that depends holomorphically on $z \in \mathscr{O}$ such that

$$\forall\, z \in \mathscr{O}\,, \quad T(z)^{-1}\, \mathbb{M}(z)\, T(z) = \operatorname{diag}\,(\mathbb{M}_1(z), \dots, \mathbb{M}_L(z))\,, \tag{79}$$

where, for some integer $\mu \geq 1$, there holds

$$\underline{\kappa} \in \operatorname{sp}(\mathbb{M}_\ell(\underline{z})) \Longleftrightarrow 1 \leq \ell \leq \mu\,.$$

Moreover, the blocks $\mathbb{M}_1, \dots, \mathbb{M}_L$ are of the first, second, third or fourth type. Since we have $\underline{\kappa} \in \mathbb{S}^1$, it is not difficult to check that the blocks $\mathbb{M}_1, \dots, \mathbb{M}_\mu$ in (79) can only be of the third or fourth type[11]. For all (κ, z) sufficiently close to $(\underline{\kappa}, \underline{z})$, we have

$$\det(\mathbb{M}(z) - \kappa\, I) = \vartheta(\kappa, z) \prod_{\ell=1}^{\mu} \det(\mathbb{M}_\ell(z) - \kappa\, I)\,, \quad \vartheta(\underline{\kappa}, \underline{z}) \neq 0\,,$$

and ϑ is a holomorphic function of (κ, z) near $(\underline{\kappa}, \underline{z})$. Using the relations (78) and (74), which are both valid for (κ, z) close to $(\underline{\kappa}, \underline{z})$, we obtain (for a possibly different function ϑ which is still denoted ϑ)

$$\det(z\, I - \mathscr{A}(\kappa)) = \vartheta(\kappa, z) \prod_{\ell=1}^{\mu} \det(\mathbb{M}_\ell(z) - \kappa\, I)\,, \quad \vartheta(\underline{\kappa}, \underline{z}) \neq 0\,. \tag{80}$$

We now examine each determinant $\det(\mathbb{M}_\ell(z) - \kappa\, I)$ in (80). We recall that $\mathbb{M}_\ell$, $1 \leq \ell \leq \mu$, is either a block of the third or fourth type, and $\underline{\kappa}$ is the unique eigenvalue of $\mathbb{M}_\ell(\underline{z})$. If $\mathbb{M}_\ell$ is a block of the third type, then we have

$$\det(\mathbb{M}_\ell(z) - \kappa\, I) = \mathbb{M}_\ell(z) - \kappa \in \mathbb{C}\,,$$

and[12]

$$\frac{\partial(\mathbb{M}_\ell(z) - \kappa)}{\partial z}\Big|_{(\underline{\kappa}, \underline{z})} = \mathbb{M}_\ell'(\underline{z}) \neq 0\,.$$

If $\mathbb{M}_\ell$ is a block of the fourth type, then we have

$$\mathbb{M}_\ell(\underline{z}) - \underline{\kappa}\, I = \underline{\kappa} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & 0 \end{pmatrix}, \tag{81}$$

and therefore (we use the notation of Definition 4.1 for blocks of the fourth type)

$$\frac{\partial \det(\mathbb{M}_\ell(z) - \kappa\, I)}{\partial z}\Big|_{(\underline{\kappa}, \underline{z})} = (-1)^{\nu_\ell - 1}\, \underline{\kappa}^{\nu_\ell - 1}\, m_\ell \neq 0\,.$$

[11] The eigenvalues of a block of the first type necessarily belong to $\mathscr{U}$, and the eigenvalues of a block of the second type belong to $\mathbb{D}$, see Lemma 4.1 a little further for a refined statement.

[12] Recall from Definition 4.1 that for a block of the third type, $\mathbb{M}_\ell'(\underline{z})$ can not be zero.

Applying the Weierstrass preparation Theorem, for which we refer to [11], for each $\ell = 1, \dots, \mu$, there exists a holomorphic function β_ℓ defined on a suitable neighborhood of $\underline{\kappa}$ and that satisfies

$$\forall\, \ell = 1, \dots, \mu\,, \quad \det(\mathbb{M}_\ell(z) - \kappa\, I) = \vartheta(\kappa, z)\,(z - \beta_\ell(\kappa))\,, \quad \beta_\ell(\underline{\kappa}) = \underline{z}\,, \quad \vartheta(\underline{\kappa}, \underline{z}) \neq 0\,. \tag{82}$$

Using the latter factorization in (80), we obtain

$$\det(z\, I - \mathscr{A}(\kappa)) = \vartheta(\kappa, z) \prod_{\ell=1}^{\mu} (z - \beta_\ell(\kappa))\,, \quad \vartheta(\underline{\kappa}, \underline{z}) \neq 0\,.$$

Evaluating at $\kappa = \underline{\kappa}$, we find that μ equals the multiplicity of $\underline{z}$ as a root of the characteristic polynomial of $\mathscr{A}(\underline{\kappa})$, hence $\mu = \underline{\alpha}$. Going back to Definition 2.3 of geometrically regular operators, we see that it only remains to construct some eigenvectors $e_\ell(\kappa)$ of $\mathscr{A}(\kappa)$ associated with the eigenvalues $\beta_\ell(\kappa)$ and that depend holomorphically on κ.

We now go back to the reduction (79). In what follows, $T_j(z)$ denotes the j-th column vector of the matrix $T(z)$. Let $\ell \in \{1, \dots, \underline{\alpha}\}$. If $\mathbb{M}_\ell(z)$ is a block of the third type, we define

$$E_\ell(\kappa) := T_{j_\ell+1}(\beta_\ell(\kappa))\,, \quad j_\ell := \sum_{\ell'=1}^{\ell-1} \nu_{\ell'}\,,$$

where we use the same notation as in Definition 4.1, that is, ν_k denotes the size of the block $\mathbb{M}_k$ in (79) (this size is independent of z). We also recall that the function β_ℓ satisfies (82). Since $T_{j_\ell+1}(z)$ is an eigenvector of $\mathbb{M}(z)$ associated with the eigenvalue $\mathbb{M}_\ell(z)$, we obtain the relation

$$\mathbb{M}(\beta_\ell(\kappa))\, E_\ell(\kappa) = \kappa\, E_\ell(\kappa)\,,$$

which holds for all κ close to $\underline{\kappa}$, and $E_\ell(\kappa)$ depends holomorphically on κ. Let us now consider the case when $\mathbb{M}_\ell(z)$ is a block of the fourth type. Using the factorization (82), we know that the matrix $\mathbb{M}_\ell(\beta_\ell(\kappa)) - \kappa\, I$ is singular for all κ close to $\underline{\kappa}$. Moreover, the rank of $\mathbb{M}_\ell(\underline{z}) - \underline{\kappa}\, I$ equals $\nu_\ell - 1$, see (81), so the rank of $\mathbb{M}_\ell(\beta_\ell(\kappa)) - \kappa\, I$ is at least $\nu_\ell - 1$ for all κ. Consequently, the kernel of $\mathbb{M}_\ell(\beta_\ell(\kappa)) - \kappa\, I$ is one-dimensional for all κ close to $\underline{\kappa}$, and the last row of $\mathbb{M}_\ell(\beta_\ell(\kappa)) - \kappa\, I$ is a linear combination of the first $\nu_\ell - 1$ rows. We can then construct a vector $\mathbf{e}_\ell(\kappa) \in \mathbb{C}^{\nu_\ell}$ that depends holomorphically on κ and such that[13]

$$\mathbf{e}_\ell(\underline{\kappa}) = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}, \quad (\mathbb{M}_\ell(\beta_\ell(\kappa)) - \kappa\, I)\, \mathbf{e}_\ell(\kappa) = 0\,.$$

It is now not difficult to construct a vector $E_\ell(\kappa)$ that depends holomorphically on κ, that satisfies

$$\mathbb{M}(\beta_\ell(\kappa))\, E_\ell(\kappa) = \kappa\, E_\ell(\kappa)\,, \quad E_\ell(\underline{\kappa}) = T_{j_\ell+1}(\underline{z})\,, \quad j_\ell := \sum_{\ell'=1}^{\ell-1} \nu_{\ell'}\,. \tag{83}$$

Indeed, if we write the vector $\mathbf{e}_\ell(\kappa)$ as $(\gamma_1(\kappa), \dots, \gamma_{\nu_\ell}(\kappa))$, it is sufficient to define

$$E_\ell(\kappa) := \gamma_1(\kappa)\, T_{j_\ell+1}(\beta_\ell(\kappa)) + \dots + \gamma_{\nu_\ell}(\kappa)\, T_{j_\ell+\nu_\ell}(\beta_\ell(\kappa))\,.$$

Eventually, for all $\ell = 1, \dots, \underline{\alpha}$, we have constructed a vector $E_\ell(\kappa)$ satisfying (83) and that depends holomorphically on κ. Relation (83) shows that the $E_\ell(\underline{\kappa})$'s are linearly independent eigenvectors of $\mathbb{M}(\underline{z})$ associated with the eigenvalue $\underline{\kappa}$.

We decompose the vectors $E_\ell(\kappa)$ as $E_\ell(\kappa) = (E_{1,\ell}(\kappa) \dots E_{p+r,\ell}(\kappa))$, where each $E_{k,\ell}$ belongs to $\mathbb{C}^N$. Using (83), we find

$$E_\ell(\underline{\kappa}) = \begin{pmatrix} \underline{\kappa}^{p+r-1}\, E_{p+r,\ell}(\underline{\kappa}) & \dots & \underline{\kappa}\, E_{p+r,\ell}(\underline{\kappa}) & E_{p+r,\ell}(\underline{\kappa}) \end{pmatrix}, \quad \sum_{j=-r}^{p} \kappa^j\, \mathbb{A}_j(\beta_\ell(\kappa))\, E_{p+r,\ell}(\kappa) = 0\,.$$

In particular, the vectors $E_{p+r,\ell}(\underline{\kappa})$, $\ell = 1, \dots, \underline{\alpha}$, are linearly independent in $\mathbb{C}^N$. From the definitions (54) and (16), we obtain

$$\left(\beta_\ell(\kappa)^{s+1}\, I - \sum_{\sigma=0}^{s} \beta_\ell(\kappa)^{s-\sigma}\, \widehat{Q_\sigma}(\kappa) \right) E_{p+r,\ell}(\kappa) = 0\,.$$

[13] To construct $\mathbf{e}_\ell(\kappa)$, it is sufficient to take 1 as its first coordinate, and to determine the last coordinates by solving the linear system formed by the first $\nu_\ell - 1$ rows in the system $(\mathbb{M}_\ell(\beta_\ell(\kappa)) - \kappa\, I)\, \mathbf{e}_\ell(\kappa) = 0$. The last row will be automatically zero as a linear combination of the other rows.

Consequently, the vectors defined by

$$\forall\, \ell = 1, \dots, \underline{\alpha}\,, \quad e_\ell(\kappa) := \big(\beta_\ell(\kappa)^s\, E_{p+r,\ell}(\kappa) \quad \dots \quad \beta_\ell(\kappa)\, E_{p+r,\ell}(\kappa) \quad E_{p+r,\ell}(\kappa)\big) \in \mathbb{C}^{N(s+1)}$$

satisfy

$$\forall\, \ell = 1, \dots, \underline{\alpha}\,, \quad \mathscr{A}(\kappa)\, e_\ell(\kappa) = \beta_\ell(\kappa)\, e_\ell(\kappa)\,.$$

It is straightforward to check that the vectors $e_\ell(\underline{\kappa})$ are linearly independent, so the vectors $e_\ell(\kappa)$ remain linearly independent for κ close to $\underline{\kappa}$. We have thus proved that the operators Q_σ are geometrically regular. Proposition 2.3 shows that the discretization (14) is stable in the sense of Definition 2.2 (because the von Neumann condition is satisfied).

• From now on, we assume that the operators Q_σ are geometrically regular and that the discretization (14) of the Cauchy problem is stable. In particular, Proposition 2.2 shows that the matrix $\mathscr{A}(\kappa)$ is uniformly power bounded for $\kappa \in \mathbb{S}^1$. Our goal is to show that the matrix $\mathbb{M}(z)$ defined by (57) satisfies the discrete block structure condition of Definition 4.1. Since the proof is quite long, we split it in several steps.

Step 1. First of all, condition *(1)* of Definition 4.1 follows from Lemma 3.7. This property immediately implies that the discrete block structure condition is satisfied in the neighborhood of any $\underline{z} \in \mathscr{U}$. More precisely, let $\underline{z} \in \mathscr{U}$. In a small neighborhood $\mathscr{O}$ of $\underline{z}$, the generalized eigenspace associated with eigenvalues of $\mathbb{M}(z)$ in $\mathbb{D}$ and the generalized eigenspace associated with eigenvalues of $\mathbb{M}(z)$ in $\mathscr{U}$ both depend holomorphically on $z \in \mathscr{O}$ (this follows from the Dunford-Taylor formula for projectors, see the proof of Lemma 2.6). We can then reduce $\mathbb{M}(z)$ to a block diagonal form

$$T(z)^{-1}\, \mathbb{M}(z)\, T(z) = \mathrm{diag}\, \big(\mathbb{M}_\flat(z), \mathbb{M}_\sharp(z)\big)\,, \quad \mathbb{M}_\flat(z) \in \mathscr{M}_{Nr}(\mathbb{C})\,, \quad \mathbb{M}_\sharp(z) \in \mathscr{M}_{Np}(\mathbb{C})\,,$$

where the eigenvalues of $\mathbb{M}_\flat(z)$ belong to $\mathbb{D}$ and the eigenvalues of $\mathbb{M}_\sharp(z)$ belong to $\mathscr{U}$. The dimension of each block follows from Lemma 3.7. The invertible matrix $T(z)$ depends holomorphically on $z \in \mathscr{O}$. Then we use the following classical result.

Lemma 4.1. *Let $M \in \mathscr{M}_m(\mathbb{C})$. Then the spectrum of M is included in $\mathbb{D}$ if and only if there exists an invertible matrix P and a positive constant δ such that*

$$(P^{-1}\, M\, P)^*\, (P^{-1}\, M\, P) \leq (1-\delta)\, I\,.$$

Similarly, the spectrum of M is included in $\mathscr{U}$ if and only if there exists an invertible matrix P and a positive constant δ such that

$$(P^{-1}\, M\, P)^*\, (P^{-1}\, M\, P) \geq (1+\delta)\, I\,.$$

Proof of Lemma 4.1. Let $M \in \mathscr{M}_m(\mathbb{C})$ be such that there exists an invertible matrix P and a positive constant δ satisfying

$$(P^{-1}\, M\, P)^*\, (P^{-1}\, M\, P) \leq (1-\delta)\, I\,.$$

Let μ be an eigenvalue of M, and let us consider an eigenvector that we write under the form $P\, X$, with $X \in \mathbb{C}^m$, $|X| = 1$. Then we have $P^{-1}\, M\, P\, X = \mu\, X$, and

$$|\mu|^2 = |(P^{-1}\, M\, P)\, X|^2 \leq 1 - \delta < 1\,,$$

so the spectrum of M is included in $\mathbb{D}$.

Let now $M \in \mathscr{M}_m(\mathbb{C})$ have its spectrum included in $\mathbb{D}$. Let us first choose an invertible matrix P that reduces M to its Jordan form

$$P^{-1}\, M\, P = \begin{pmatrix} \mu_1 & \theta_1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \theta_{m-1} \\ 0 & \dots & 0 & \mu_m \end{pmatrix}\,,$$

with $\mu_j \in \mathbb{D}$ and $\theta_j \in \{0,1\}$. Introducing $P_\varepsilon := \mathrm{diag}\,(1,\varepsilon,\dots,\varepsilon^{m-1})$, $\varepsilon>0$, we have

$$P_\varepsilon^{-1}\,P^{-1}\,M\,P\,P_\varepsilon = \begin{pmatrix} \mu_1 & \varepsilon\,\theta_1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \varepsilon\,\theta_{m-1} \\ 0 & \dots & 0 & \mu_m \end{pmatrix}.$$

Since the matrix $I - \mathrm{diag}\,(|\mu_1|^2,\dots,|\mu_m|^2)$ is positive definite, the matrix

$$I - (P_\varepsilon^{-1}\,P^{-1}\,M\,P\,P_\varepsilon)^*\,(P_\varepsilon^{-1}\,P^{-1}\,M\,P\,P_\varepsilon)$$

is positive definite for $\varepsilon>0$ sufficiently small and the result follows. The analysis in the case of eigenvalues in $\mathscr{U}$ instead of $\mathbb{D}$ is similar. □

Up to a constant change of basis (which modifies $T(z)$ but keeps the holomorphy), we can thus achieve the inequalities

$$\mathbb{M}_\flat(\underline{z})^*\,\mathbb{M}_\flat(\underline{z}) \le (1-2\,\delta)\,I\,,\quad \mathbb{M}_\sharp(\underline{z})^*\,\mathbb{M}_\sharp(\underline{z}) \ge (1+2\,\delta)\,I\,,$$

for some positive constant δ. Thanks to a continuity argument, we can conclude that the discrete block structure condition is satisfied in a sufficiently small neighborhood $\mathscr{O}$ of $\underline{z}\in\mathscr{U}$. The reduction only involves one block of the first type and one block of the second type.

Step 2. We now turn to the case $\underline{z}\in\mathbb{S}^1$. If $\mathbb{M}(\underline{z})$ has no eigenvalue in $\mathbb{S}^1$ then we are reduced to the preceeding case. We thus assume that $\mathbb{M}(\underline{z})$ has some eigenvalues in $\mathbb{S}^1$. More precisely, let $\underline{\kappa}_1,\dots,\underline{\kappa}_k$ denote the elements of $\mathrm{sp}(\mathbb{M}(\underline{z}))\cap\mathbb{S}^1$, and let $\alpha_1,\dots,\alpha_k$ denote the corresponding algebraic multiplicities of these eigenvalues. The generalized eigenspace $\mathrm{Ker}(\mathbb{M}(\underline{z})-\underline{\kappa}_j\,I)^{\alpha_j}$ is denoted $\underline{\mathbb{K}}_j$. For z sufficiently close to $\underline{z}$, we also let $\mathbb{K}_j(z)$ denote the generalized eigenspace of $\mathbb{M}(z)$ associated with its α_j eigenvalues that are close to $\underline{\kappa}_j$. The space $\mathbb{K}_j(z)$ depends holomorphically on z (same argument as in Lemma 2.6) and satisfies $\mathbb{K}_j(\underline{z})=\underline{\mathbb{K}}_j$. Then for z in a small neighborhood $\mathscr{O}$ of $\underline{z}$, we can perform a block diagonalization of $\mathbb{M}(z)$ with a holomorphic change of basis:

$$T(z)^{-1}\,\mathbb{M}(z)\,T(z) = \mathrm{diag}\,(\mathbb{M}_\flat(z),\mathbb{M}_\sharp(z),\mathbb{M}_1(z),\dots,\mathbb{M}_k(z))\,,$$

where the eigenvalues of $\mathbb{M}_\flat(z)$ belong to $\mathbb{D}$, the eigenvalues of $\mathbb{M}_\sharp(z)$ belong to $\mathscr{U}$, and for all $j=1,\dots,k$, the α_j eigenvalues of $\mathbb{M}_j(z)\in\mathscr{M}_{\alpha_j}(\mathbb{C})$ belong to a sufficiently small neighborhood of $\underline{\kappa}_j$. As in the preceeding case, we can always achieve the inequalities

$$\forall\, z\in\mathscr{O}\,,\quad \mathbb{M}_\flat(z)^*\,\mathbb{M}_\flat(z)\le(1-\delta)\,I\,,\quad \mathbb{M}_\sharp(z)^*\,\mathbb{M}_\sharp(z)\ge(1+\delta)\,I\,,$$

for some constant $\delta>0$, so from now on we focus on the blocks $\mathbb{M}_j(z)$. For the sake of clarity, we shall only deal with the first block $\mathbb{M}_1(z)$. This is only to avoid overloaded notations with many indeces. Of course, the analysis below is valid for any of the blocks $\mathbb{M}_j(z)$. We are going to show that in a convenient holomorphic basis of $\mathbb{K}_1(z)$, the block $\mathbb{M}_1(z)$ reduces to a block diagonal form with blocks of the third or fourth type. The proof follows the analysis of [14, 16].

Step 3. Following [14], we first study the characteristic polynomial of $\mathbb{M}_1(z)$. For z close to $\underline{z}$, the α_1 eigenvalues of $\mathbb{M}_1(z)$ are close to $\underline{\kappa}_1$. Combining the relations (74) and (78), we obtain

$$\det(\mathbb{M}_1(z)-\kappa\,I) = \vartheta(\kappa,z)\,\det\,(z\,I-\mathscr{A}(\kappa))\,, \tag{84}$$

where ϑ is holomorphic with respect to (κ,z) and does not vanish on a small neighborhood of $(\underline{\kappa}_1,\underline{z})$. We know that $\underline{z}\in\mathbb{S}^1$ is an eigenvalue of $\mathscr{A}(\underline{\kappa}_1)$ so we can use the geometric regularity of the operators Q_σ. For (κ,z) in a sufficiently small neighborhood of $(\underline{\kappa}_1,\underline{z})$, (84) reads

$$\det(\mathbb{M}_1(z)-\kappa\,I) = \vartheta(\kappa,z)\prod_{j=1}^{\underline{\alpha}}\big(z-\beta_j(\kappa)\big)\,, \tag{85}$$

where $\underline{\alpha}$ is a fixed integer (not necessarily equal to α_1), and the β_j's are holomorphic functions on a neighborhood $\mathscr{W}$ of $\underline{\kappa}_1$ satisfying $\beta_j(\underline{\kappa}_1)=\underline{z}$. Thanks to the uniform power boundedness of the matrices $\mathscr{A}(\kappa)$ for $\kappa\in\mathbb{S}^1$, we know that $|\beta_j(\kappa)|\le 1$ for $\kappa\in\mathbb{S}^1\cap\mathscr{W}$. Using the Taylor expansion

$$\big|\beta_j(\underline{\kappa}_1\,\mathrm{e}^{i\,\xi})\big|^2 = \big|\underline{z}+i\,\underline{\kappa}_1\,\beta_j'(\underline{\kappa}_1)\,\xi+o(\xi)\big|^2 = 1+2\,\mathrm{Re}\,(i\,\overline{\underline{z}}\,\underline{\kappa}_1\,\beta_j'(\underline{\kappa}_1))\,\xi+o(\xi)\,,$$

for $\xi \in \mathbb{R}$ close to 0, we obtain that there exists a real number σ_j such that

$$\underline{\kappa}_1 \, \beta_j'(\underline{\kappa}_1) = \sigma_j \, \underline{z} \,, \quad \sigma_j \in \mathbb{R} \,. \tag{86}$$

Thanks to (85), we can see that $\underline{\kappa}_1$ is a root of finite multiplicity of the holomorphic function $\underline{z} - \beta_j(\cdot)$. (For otherwise, the function $\underline{z} - \beta_j(\kappa)$ would be identically zero for all κ close to $\underline{\kappa}_1$, and this is ruled out by (85).) Consequently there exists an integer $\nu_j \geq 1$ such that

$$\forall \, \nu = 1, \dots, \nu_j - 1 \,, \quad \beta_j^{(\nu)}(\underline{\kappa}_1) = 0 \,, \quad \beta_j^{(\nu_j)}(\underline{\kappa}_1) \neq 0 \,. \tag{87}$$

We can apply the Weierstrass preparation Theorem to the holomorphic function $z - \beta_j(\kappa)$. For all $j = 1, \dots, \underline{\alpha}$, there exists $P_j(\kappa, z)$ that is a unitary polynomial function in κ with degree ν_j, such that for (κ, z) close to $(\underline{\kappa}_1, \underline{z})$, there holds

$$z - \beta_j(\kappa) = \vartheta(\kappa, z) \, P_j(\kappa, z) \,, \quad P_j(\kappa, \underline{z}) = (\kappa - \underline{\kappa}_1)^{\nu_j} \,, \quad \vartheta(\underline{\kappa}_1, \underline{z}) \neq 0 \,. \tag{88}$$

Using (88), (85) reduces to

$$\det(\mathbb{M}_1(z) - \kappa \, I) = \vartheta(\kappa, z) \prod_{j=1}^{\underline{\alpha}} P_j(\kappa, z) \,.$$

For z close to $\underline{z}$, the polynomial $P_j(\cdot, z)$ has ν_j roots, and these roots are close to $\underline{\kappa}_1$. Consequently, the size of the block $\mathbb{M}_1(z)$ equals $\nu_1 + \cdots + \nu_{\underline{\alpha}}$. We also know that the size of this block equals α_1, the algebraic multiplicity of $\underline{\kappa}_1$ as an eigenvalue of $\mathbb{M}(\underline{z})$. Up to reordering the terms, there exists an integer $\underline{\mu}$ (possibly zero) such that

$$\nu_1 = \cdots = \nu_{\underline{\mu}} = 1 \,, \quad \nu_{\underline{\mu}+1}, \dots, \nu_{\underline{\alpha}} \geq 2 \,.$$

For $j = 1, \dots, \underline{\mu}$, we have $\beta_j'(\underline{\kappa}_1) \neq 0$, see (87), or equivalently $\sigma_j \neq 0$ in (86). Therefore β_j is a biholomorphic homeomorphism from a neighborhood $\mathscr{W}$ of $\underline{\kappa}_1$ to a neighborhood $\mathscr{O}$ of $\underline{z}$. We let m_j denote its (holomorphic) inverse. With such notation, we obtain $P_j(\kappa, z) = \kappa - m_j(z)$ for all $j = 1, \dots, \underline{\mu}$.

Using the relation (88), we also obtain $\partial_z P_j(\underline{\kappa}_1, \underline{z}) \neq 0$. Then Puiseux's expansions theory shows that for z close to $\underline{z}$ and $z \neq \underline{z}$, the ν_j roots of $P_j(\cdot, z)$ are simple, see for instance [1]. More precisely, Puiseux's expansions theory shows that the ν_j roots of $P_j(\cdot, z)$ behave asymptotically, at the leading order in $(z - \underline{z})$ as the roots of

$$(\kappa - \underline{\kappa}_1)^{\nu_j} + \partial_z P_j(\underline{\kappa}_1, \underline{z}) \, (z - \underline{z}) = 0 \,,$$

when z is close to $\underline{z}$.

Step 4. For each eigenvalue $\beta_j(\kappa)$, $j = 1, \dots, \underline{\alpha}$ and κ close to $\underline{\kappa}_1$, we know that $\mathscr{A}(\kappa)$ has a holomorphic eigenvector $e_j(\kappa) \in \mathbb{C}^{N(s+1)}$. Using the definition (16) of $\mathscr{A}$, we find that $e_j(\kappa)$ reads

$$\forall \, j = 1, \dots, \underline{\alpha} \,, \quad e_j(\kappa) = \begin{pmatrix} \beta_j(\kappa)^s \, \mathbf{e}_j(\kappa) \\ \vdots \\ \beta_j(\kappa) \, \mathbf{e}_j(\kappa) \\ \mathbf{e}_j(\kappa) \end{pmatrix} \,, \quad \mathbf{e}_j(\kappa) \in \mathbb{C}^N \,, \quad \sum_{\ell=-r}^{p} \kappa^\ell \, \mathbb{A}_\ell(\beta_j(\kappa)) \, \mathbf{e}_j(\kappa) = 0 \,.$$

The vectors $\mathbf{e}_1(\underline{\kappa}_1), \dots, \mathbf{e}_{\underline{\alpha}}(\underline{\kappa}_1)$ are linearly independent in $\mathbb{C}^N$ because $e_1(\underline{\kappa}_1), \dots, e_{\underline{\alpha}}(\underline{\kappa}_1)$ are linearly independent in $\mathbb{C}^{N(s+1)}$. Therefore when κ is close to $\underline{\kappa}_1$, the vectors $\mathbf{e}_1(\kappa), \dots, \mathbf{e}_{\underline{\alpha}}(\kappa)$ remain linearly independent. We define

$$\forall \, j = 1, \dots, \underline{\alpha} \,, \quad E_j(\kappa) := \begin{pmatrix} \kappa^{p+r-1} \, \mathbf{e}_j(\kappa) \\ \vdots \\ \kappa \, \mathbf{e}_j(\kappa) \\ \mathbf{e}_j(\kappa) \end{pmatrix} \in \mathbb{C}^{N(p+r)} \,.$$

These vectors depend holomorphically on κ, they are linearly independent in $\mathbb{C}^{N(p+r)}$ for κ close to $\underline{\kappa}_1$, and $E_j(\kappa)$ is an eigenvector of $\mathbb{M}(\beta_j(\kappa))$ associated with the eigenvalue κ:

$$\forall \, j = 1, \dots, \underline{\alpha} \,, \quad \big(\mathbb{M}(\beta_j(\kappa)) - \kappa \, I \big) \, E_j(\kappa) = 0 \,. \tag{89}$$

In particular, for $j = 1, \dots, \underline{\mu}$ and for z in a neighborhood $\mathscr{O}$ of $\underline{z}$, we have

$$\forall\, j = 1, \dots, \underline{\mu}\,, \quad \forall\, z \in \mathscr{O}\,, \quad \big(\mathbb{M}(z) - m_j(z)\, I\big)\, E_j(m_j(z)) = 0\,. \tag{90}$$

Let us recall that m_j is the holomorphic inverse of β_j for $j = 1, \dots, \underline{\mu}$, that is when $\beta_j'(\underline{\kappa}_1) \neq 0$. For all $j = 1, \dots, \underline{\mu}$, we have thus constructed a holomorphic eigenvalue $m_j(z)$ and a holomorphic eigenvector $E_j(m_j(z))$ of $\mathbb{M}(z)$. Moreover, we have $m_j'(\underline{z}) = 1/\beta_j'(\underline{\kappa}_1)$ so we get

$$\forall\, j = 1, \dots, \underline{\mu}\,, \quad m_j(\underline{z}) = \underline{\kappa}_1 \in \mathbb{S}^1\,, \quad \underline{z}\, m_j'(\underline{z})\, \overline{m_j(\underline{z})} = \frac{1}{\sigma_j} \in \mathbb{R} \setminus \{0\}\,.$$

Step 5. We now turn to the most difficult case $j = \underline{\mu} + 1, \dots, \underline{\alpha}$ (that is, $\sigma_j = 0$). We start from the relation (89), differentiate this relation $\nu_j - 1$ times with respect to κ, and evaluate the result at $\kappa = \underline{\kappa}_1$. This yields

$$\begin{aligned}
&\big(\mathbb{M}(\underline{z}) - \underline{\kappa}_1\, I\big)\, E_j(\underline{\kappa}_1) = 0\,,\\
&- E_j(\underline{\kappa}_1) + \big(\mathbb{M}(\underline{z}) - \underline{\kappa}_1\, I\big)\, E_j'(\underline{\kappa}_1) = 0\,,\\
&\qquad\vdots\\
&- (\nu_j - 1)\, E_j^{(\nu_j - 2)}(\underline{\kappa}_1) + \big(\mathbb{M}(\underline{z}) - \underline{\kappa}_1\, I\big)\, E_j^{(\nu_j - 1)}(\underline{\kappa}_1) = 0\,.
\end{aligned}$$

Then for all $j = \underline{\mu} + 1, \dots, \underline{\alpha}$, we define the following vectors:

$$\big(\underline{E}_{j,1}, \dots, \underline{E}_{j,\nu_j}\big) := \Big(E_j(\underline{\kappa}_1), \frac{\underline{\kappa}_1}{1!}\, E_j'(\underline{\kappa}_1), \dots, \frac{\underline{\kappa}_1^{\nu_j - 1}}{(\nu_j - 1)!}\, E_j^{(\nu_j - 1)}(\underline{\kappa}_1)\Big)\,, \tag{91}$$

that satisfy

$$\big(\mathbb{M}(\underline{z}) - \underline{\kappa}_1\, I\big)\, \underline{E}_{j,1} = 0\,, \quad \forall\, \nu = 2, \dots, \nu_j\,, \quad \big(\mathbb{M}(\underline{z}) - \underline{\kappa}_1\, I\big)\, \underline{E}_{j,\nu} = \underline{\kappa}_1\, \underline{E}_{j,\nu-1}\,. \tag{92}$$

Using the relations (90) and (92), we can show that the vectors

$$E_1(\underline{\kappa}_1), \dots, E_{\underline{\mu}}(\underline{\kappa}_1), \quad \underline{E}_{\underline{\mu}+1,1}, \dots, \underline{E}_{\underline{\mu}+1,\nu_{\underline{\mu}+1}}, \quad \dots, \quad \underline{E}_{\underline{\alpha},1}, \dots, \underline{E}_{\underline{\alpha},\nu_{\underline{\alpha}}},$$

are linearly independent. Moreover, these α_1 vectors span the generalized eigenspace $\underline{\mathbb{K}}_1$ of $\mathbb{M}(\underline{z})$ associated with the eigenvalue $\underline{\kappa}_1$ (they all belong to this space and they are linearly independent so they form a basis). So far we have thus obtained a basis of $\underline{\mathbb{K}}_1$ in which the block $\mathbb{M}_1(\underline{z})$ reads

$$\mathbb{M}_1(\underline{z}) = \operatorname{diag}\,\big(\underline{\kappa}_1, \dots, \underline{\kappa}_1, \underline{M}_{\underline{\mu}+1}, \dots, \underline{M}_{\underline{\alpha}}\big)\,, \quad \underline{M}_j := \underline{\kappa}_1 \begin{pmatrix} 1 & 1 & 0 & 0\\ 0 & \ddots & \ddots & 0\\ \vdots & \ddots & \ddots & 1\\ 0 & \dots & 0 & 1 \end{pmatrix} \in \mathscr{M}_{\nu_j}(\mathbb{C})\,.$$

In the next step of the analysis, we are going to extend the definition of the vectors $\underline{E}_{j,\nu}$ to a neighborhood of $\underline{z}$.

Step 6. Let us recall that for all $j = 1, \dots, \underline{\alpha}$, the polynomial $P_j(\cdot, z)$ is defined by (88). We can choose $r > 0$ such that for z in a neighborhood $\mathscr{O}$ of $\underline{z}$, the ν_j roots of $P_j(\cdot, z)$ belong to the disc of center $\underline{\kappa}_1$ and radius $r/2$. Then for all $z \in \mathscr{O}$, for all $j = \underline{\mu} + 1, \dots, \underline{\alpha}$ and for all $\nu = 1, \dots, \nu_j$, we define a vector $E_{j,\nu}(z)$ by the formula

$$E_{j,\nu}(z) := \frac{\underline{\kappa}_1^{\nu-1}\, (\nu_j - \nu)!}{2\, i\, \pi\, \nu_j!} \int_{|\kappa - \underline{\kappa}_1| = r} \frac{\partial_\kappa^\nu P_j(\kappa, z)}{P_j(\kappa, z)}\, E_j(\kappa)\, \mathrm{d}\kappa\,.$$

Cauchy's formula shows that for $z = \underline{z}$, $E_{j,\nu}(\underline{z})$ coincides with the vector $\underline{E}_{j,\nu}$ defined by (91). Moreover, $E_{j,\nu}(z)$ depends holomorphically on $z \in \mathscr{O}$. In particular we can choose the neighborhood $\mathscr{O}$ such that for all $z \in \mathscr{O}$, the vectors

$$E_1(m_1(z)), \dots, E_{\underline{\mu}}(m_{\underline{\mu}}(z)), \quad E_{\underline{\mu}+1,1}(z), \dots, E_{\underline{\mu}+1,\nu_{\underline{\mu}+1}}(z), \quad \dots, \quad E_{\underline{\alpha},1}(z), \dots, E_{\underline{\alpha},\nu_{\underline{\alpha}}}(z),$$

are linearly independent. We are now going to show that these vectors span the invariant subspace $\mathbb{K}_1(z)$, and that in this basis of $\mathbb{K}_1(z)$, the matrix $\mathbb{M}_1(z)$ is in block diagonal form with blocks of the third and fourth type (the proof will be almost finished then !).

For z close to $\underline{z}$ and $j = \underline{\mu}+1, \dots, \underline{\alpha}$, we let $\mathbb{F}_j(z)$ denote the vector space spanned by the linearly independent vectors $E_{j,1}(z), \dots, E_{j,\nu_j}(z)$. For $j = 1, \dots, \underline{\mu}$, we let $\mathbb{F}_j(z)$ denote the one-dimensional vector space spanned by $E_j(m_j(z))$. Then for all j, the dimension of $\mathbb{F}_j(z)$ is ν_j. Moreover the sum of the $\mathbb{F}_j(z)$ is direct and has dimension α_1. We already know that for $j = 1, \dots, \underline{\mu}$, $E_j(m_j(z))$ is an eigenvector of $\mathbb{M}(z)$ for the eigenvalue $m_j(z)$, see (90). Consequently, $\mathbb{F}_j(z)$ is stable by the matrix $\mathbb{M}(z)$ and $\mathbb{F}_j(z) \subset \mathbb{K}_1(z)$ for $j = 1, \dots, \underline{\mu}$. We are now going to show that the same properties hold true for $j = \underline{\mu}+1, \dots, \underline{\alpha}$.

For $z = \underline{z}$, thanks to (92), we know that $\mathbb{F}_j(\underline{z})$ is stable by $\mathbb{M}(\underline{z})$ and $\mathbb{F}_j(\underline{z}) \subset \underline{\mathbb{K}}_1$. From now on we thus consider a fixed $z \in \mathscr{O} \setminus \{\underline{z}\}$. For all $j = \underline{\mu}+1, \dots, \underline{\alpha}$, we let $\kappa_{j,1}, \dots, \kappa_{j,\nu_j}$ denote the ν_j disctinct roots of the polynomial $P_j(\cdot, z)$. (We recall that these roots are distinct thanks to Puiseux's expansions theory.) These roots belong to the disc of center $\underline{\kappa}_1$ and radius $r/2$. Therefore, using the residue Theorem, we obtain

$$E_{j,\nu}(z) = \sum_{m=1}^{\nu_j} \omega_{j,\nu,m}\, E_j(\kappa_{j,m})\,,$$

for some suitable complex numbers $\omega_{j,\nu,m}$. Therefore $\mathbb{F}_j(z)$ is contained in the vector space $\widetilde{\mathbb{F}}_j(z)$ spanned by the vectors $E_j(\kappa_{j,1}), \dots, E_j(\kappa_{j,\nu_j})$. Because the dimension of $\mathbb{F}_j(z)$ is ν_j, we can conlude that the dimension of $\widetilde{\mathbb{F}}_j(z)$ is also ν_j and $\mathbb{F}_j(z) = \widetilde{\mathbb{F}}_j(z)$. Let us now show that $\widetilde{\mathbb{F}}_j(z)$ is stable by $\mathbb{M}(z)$. We know that $P_j(\kappa_{j,m}, z) = 0$ so $z = \beta_j(\kappa_{j,m})$. Using (89) we see that $E_j(\kappa_{j,m})$ is an eigenvector of $\mathbb{M}(z)$ for the eigenvalue $\kappa_{j,m}$ that is close to $\underline{\kappa}_1$. Consequently the vector space $\widetilde{\mathbb{F}}_j(z)$ is stable by $\mathbb{M}(z)$ and $\widetilde{\mathbb{F}}_j(z) \subset \mathbb{K}_1(z)$. Since $\mathbb{F}_j(z) = \widetilde{\mathbb{F}}_j(z)$, we have proved that for all $j = 1, \dots, \underline{\alpha}$, $\mathbb{F}_j(z)$ is stable by $\mathbb{M}(z)$ and $\mathbb{F}_j(z) \subset \mathbb{K}_1(z)$. Using a dimension argument, we have obtained

$$\mathbb{K}_1(z) = \mathbb{F}_1(z) \oplus \cdots \oplus \mathbb{F}_{\underline{\alpha}}(z)\,,$$

and each $\mathbb{F}_j(z)$ is a stable vector space for $\mathbb{M}(z)$. Moreover, the characteristic polynomial of the restriction of $\mathbb{M}(z)$ to $\mathbb{F}_j(z)$ is $P_j(\cdot, z)$. We have thus constructed a holomorphic basis of $\mathbb{K}_1(z)$ in which the matrix $\mathbb{M}_1(z)$ reads

$$\mathbb{M}_1(z) = \operatorname{diag}\left(m_1(z), \dots, m_{\underline{\mu}}(z), M_{\underline{\mu}+1}(z), \dots, M_{\underline{\alpha}}(z)\right).$$

We also know that the characteristic polynomial of $M_j(z)$ is $P_j(\cdot, z)$ for $j = \underline{\mu}+1, \dots, \underline{\alpha}$, and $M_j(\underline{z})$ is the Jordan block $\underline{M}_j$ defined above (same expression as in Definition 4.1). The size of each block in the reduction of $\mathbb{M}_1(z)$ is independent of z.

Step 7. The only remaining task is to obtain the property stated in Definition 4.1 for the lower left corner coefficient m_j of $M_j'(\underline{z})$, $j = \underline{\mu}+1, \dots, \underline{\alpha}$. We know that $P_j(\kappa, z)$ is the characteristic polynomial of $M_j(z)$, and (88) gives $\partial_z P_j(\underline{\kappa}_1, \underline{z}) \neq 0$. According to the form of $M_j(\underline{z}) = \underline{M}_j$, we also have

$$\partial_z P_j(\underline{\kappa}_1, \underline{z}) = \det \begin{pmatrix} \star & -\underline{\kappa}_1 & & 0 \\ \vdots & 0 & \ddots & \\ \star & \vdots & \ddots & -\underline{\kappa}_1 \\ -m_j & 0 & \dots & 0 \end{pmatrix} = -\underline{\kappa}_1^{\nu_j - 1}\, m_j\,.$$

Hence m_j is not zero. Let $\theta \in \mathbb{C}$ satisfy $\operatorname{Re}\theta > 0$. For $\varepsilon > 0$, we define $z_\varepsilon := \underline{z}\,(1 + \varepsilon\,\theta) \in \mathscr{U}$. The eigenvalues of $M_j(z_\varepsilon)$ are the roots of $P_j(\cdot, z_\varepsilon)$. According to Puiseux's expansions theory, the eigenvalues $\kappa_1(\varepsilon), \dots, \kappa_{\nu_j}(\varepsilon)$ of $M_j(z_\varepsilon)$ have an asymptotic expansion of the form

$$\kappa_\nu(\varepsilon) = \underline{\kappa}_1 \left(1 + \varepsilon^{1/\nu_j}\, \zeta_\nu + O(\varepsilon^{2/\nu_j})\right), \tag{93}$$

where the complex numbers ζ_ν are such that

$$\begin{aligned} 0 = P_j(\kappa_\nu(\varepsilon), z_\varepsilon) &= (\kappa_\nu(\varepsilon) - \underline{\kappa}_1)^{\nu_j} - \underline{\kappa}_1^{\nu_j - 1}\, m_j\, (z_\varepsilon - \underline{z}) + o(\varepsilon) \\ &= \left(\underline{\kappa}_1^{\nu_j}\, \zeta_\nu^{\nu_j} - \underline{\kappa}_1^{\nu_j - 1}\, m_j\, \underline{z}\, \theta\right) \varepsilon + o(\varepsilon)\,. \end{aligned}$$

In other words, the ζ_ν's are the roots of the equation

$$\zeta^{\nu_j} = \underline{\kappa}_1^{-1}\, m_j\, \underline{z}\, \theta\,,$$

and the ν_j roots of this equation are simple. Our goal is to show that none of these roots is purely imaginary. Let us argue by contradiction and let us therefore assume that, say, ζ_1 is purely imaginary. We write $\zeta_1 = i\,\xi_1$. Then some simple Taylor expansions (recall (93)) yield

$$\begin{aligned} &\frac{\kappa_1(\varepsilon)}{\underline{\kappa}_1} - \mathrm{e}^{i\,\xi_1\,\varepsilon^{1/\nu_j}} = O(\varepsilon^{2/\nu_j})\,, \\ \forall\,\nu = 2,\dots,\nu_j\,, \quad &\frac{\kappa_\nu(\varepsilon)}{\underline{\kappa}_1} - \mathrm{e}^{i\,\xi_1\,\varepsilon^{1/\nu_j}} = O(\varepsilon^{1/\nu_j})\,, \end{aligned}$$

and we get

$$\left|\det\left(M_j(z_\varepsilon) - \underline{\kappa}_1\,\mathrm{e}^{i\,\xi_1\,\varepsilon^{1/\nu_j}}\,I\right)\right| = \prod_{\nu=1}^{\nu_j}\left|\kappa_\nu(\varepsilon) - \underline{\kappa}_1\,\mathrm{e}^{i\,\xi_1\,\varepsilon^{1/\nu_j}}\right| = O\big(\varepsilon^{1+1/\nu_j}\big)\,. \tag{94}$$

To complete the proof, we need the following

Lemma 4.2 ([10]). *Let Assumption 3.1 be satisfied, and let us assume that the discretization (14) is stable in the sense of Definition 2.2. Then there exists a constant $C > 0$ such that for all $z \in \mathscr{U}$ and for all $\kappa \in \mathbb{S}^1$, there holds*

$$|(\mathbb{M}(z) - \kappa\,I)^{-1}| \leq C\,\frac{|z|}{|z|-1}\,.$$

Let us assume for the moment that Lemma 4.2 holds. Then using the block diagonalization of $\mathbb{M}(z)$ in the neighborhood of $\underline{z} \in \mathbb{S}^1$, we find that there exists a constant $C > 0$ and a neighborhood $\mathscr{O}$ of $\underline{z}$ such that for all $z \in \mathscr{O} \cap \mathscr{U}$ and for all $\kappa \in \mathbb{S}^1$, there holds

$$|(T(z)^{-1}\,\mathbb{M}(z)\,T(z) - \kappa\,I)^{-1}| \leq \frac{C}{|z|-1}\,.$$

In particular, for all $\varepsilon > 0$ sufficiently small, and all $\kappa \in \mathbb{S}^1$, there holds (recall $z_\varepsilon = \underline{z}\,(1+\varepsilon\,\theta)$ and $\mathrm{Re}\,\theta > 0$)

$$|(M_j(z_\varepsilon) - \kappa\,I)^{-1}| \leq \frac{C}{\varepsilon}\,.$$

This inequality is uniform with respect to κ, so we can use it for $\kappa = \underline{\kappa}_1\,\mathrm{e}^{i\,\xi_1\,\varepsilon^{1/\nu_j}}$. Using (94), and the classical formula $P^{-1} = \mathrm{Com}(P)^T/\det(P)$ for an invertible matrix P, we obtain that the comatrix of $M_j(\underline{z}) - \underline{\kappa}_1\,I$ vanishes. However, this is impossible because the rank of $M_j(\underline{z}) - \underline{\kappa}_1\,I$ is $\nu_j - 1$. We have thus obtained that all the roots ζ_ν have nonzero real part. □

Proof of Lemma 4.2. We first apply Proposition 2.2 and the Kreiss matrix Theorem (Theorem 2.1): since the amplification matrix $\mathscr{A}(\kappa)$ is uniformly power bounded for $\kappa \in \mathbb{S}^1$, there exists a constant $C > 0$ such that

$$\forall\,\kappa \in \mathbb{S}^1\,, \quad \forall\,z \in \mathscr{U}\,, \quad |(\mathscr{A}(\kappa) - z\,I)^{-1}| \leq \frac{C}{|z|-1}\,. \tag{95}$$

Let $z \in \mathscr{U}$, $\kappa \in \mathbb{S}^1$, and let $Y = (y, 0, \dots, 0) \in \mathbb{C}^{N(s+1)}$ with $y \in \mathbb{C}^N$. We are going to compute the vector $(\mathscr{A}(\kappa) - z\,I)^{-1}\,Y$. Indeed, let us denote $X = (x_0, \dots, x_s) \in \mathbb{C}^{N(s+1)}$ the unique solution to the linear system $(\mathscr{A}(\kappa) - z\,I)\,X = Y$. We have

$$\begin{aligned} &\forall\,\sigma = 0,\dots,s, \quad x_\sigma = z^{s-\sigma}\,x_s\,, \\ &\left(I - \sum_{\sigma=0}^{s} z^{-\sigma-1}\,\widehat{Q_\sigma}(\kappa)\right)\,x_s = -z^{-s-1}\,y\,. \end{aligned}$$

The inequality (95) gives $(|z|-1)\,|X| \leq C\,|y|$ so in particular, we have $(|z|-1)\,|x_0| \leq C\,|y|$. Using the relation $x_0 = z^s\,x_s$, we get the estimate

$$|x_s| \leq \frac{C\,|z|^{-s}}{|z|-1}\,|y|\,, \quad \text{where } x_s = -z^{-s-1}\left(I - \sum_{\sigma=0}^{s} z^{-\sigma-1}\,\widehat{Q_\sigma}(\kappa)\right)^{-1} y\,.$$

The latter matrix is invertible for otherwise, $\mathscr{A}(\kappa) - z\,I$ would have a nontrivial kernel. Taking the supremum over $y \in \mathbb{C}^N$, we obtain that there exists a constant $C > 0$ such that

$$\forall\, \kappa \in \mathbb{S}^1\,, \quad \forall\, z \in \mathscr{U}\,, \quad \left| \left(I - \sum_{\sigma=0}^{s} z^{-\sigma-1}\, \widehat{Q_\sigma}(\kappa) \right)^{-1} \right| \leq C\, \frac{|z|}{|z|-1}\,.$$

Using the relation (this relation already appeared earlier in the proof of Theorem 4.1)

$$I - \sum_{\sigma=0}^{s} z^{-\sigma-1}\, \widehat{Q_\sigma}(\kappa) = \sum_{\ell=-r}^{p} \kappa^\ell\, \mathbb{A}_\ell(z)\,,$$

we have just proved that there exists a constant $C > 0$ such that

$$\forall\, \kappa \in \mathbb{S}^1\,, \quad \forall\, z \in \mathscr{U}\,, \quad \left| \left(\sum_{\ell=-r}^{p} \kappa^\ell\, \mathbb{A}_\ell(z) \right)^{-1} \right| \leq C\, \frac{|z|}{|z|-1}\,. \tag{96}$$

We now consider a vector $b = (b_p, \dots, b_{1-r}) \in \mathbb{C}^{N(p+r)}$ and we let $X = (x_{p-1}, \dots, x_{-r})$ denote the unique solution to the linear system $(\mathbb{M}(z) - \kappa\, I)\, X = b$ (Lemma 3.7 shows that the matrix $\mathbb{M}(z) - \kappa\, I$ is invertible). From the definition (57), we obtain the relations

$$\forall\, \ell = 1-r, \dots, p-1, \quad x_\ell = \kappa^{r+\ell}\, x_{-r} + \sum_{j=0}^{\ell+r-1} \kappa^j\, b_{\ell-j}\,,$$

$$\kappa^r \left(\sum_{\ell=-r}^{p} \kappa^\ell\, \mathbb{A}_\ell(z) \right) x_{-r} = -\widetilde{b}(\kappa, z)\,,$$

with a vector $\widetilde{b}(\kappa, z)$ defined by

$$\widetilde{b}(\kappa, z) := \mathbb{A}_p(z)\, b_p + \sum_{\ell=1-r}^{p-1} \mathbb{A}_\ell(z) \sum_{j=0}^{\ell+r-1} \kappa^j\, b_{\ell-j} + \kappa\, \mathbb{A}_p(z) \sum_{j=0}^{p+r-2} \kappa^j\, b_{p-1-j}\,.$$

For $z \in \mathscr{U}$ and $\kappa \in \mathbb{S}^1$, we have a uniform bound

$$|\widetilde{b}(\kappa, z)| \leq C_0\, |b|\,,$$

because the matrices $\mathbb{A}_\ell(z)$ are uniformly bounded for $z \in \mathscr{U}$, see (54). We then use the estimate (96) to obtain the upper bound

$$|x_{-r}| \leq C\, \frac{|z|}{|z|-1}\, |b|\,,$$

with a constant C that is uniform with respect to $\kappa \in \mathbb{S}^1$ and $z \in \mathscr{U}$. The other components $x_{1-r}, \dots, x_{p-1}$ of x are easily estimated in terms of x_{-r} and b. We have thus proved that there exists a constant $C > 0$ such that for all $z \in \mathscr{U}$ and for all $\kappa \in \mathbb{S}^1$, we have

$$|(\mathbb{M}(z) - \kappa\, I)^{-1}\, b| \leq C\, \frac{|z|}{|z|-1}\, |b|\,.$$

The proof of Lemma 4.2 is complete. □

Theorem 4.1 shows that under the assumptions of Theorem 3.5, the matrix $\mathbb{M}(z)$ satisfies the discrete block structure condition. We are now interested in constructing a symmetrizer for $\mathbb{M}(z)$. Rather than working on $\mathbb{M}(z)$ directly, we shall work on this partially diagonalized form of $\mathbb{M}(z)$ and eventually go back to $\mathbb{M}(z)$ by changing basis.

4.2. **The construction of symmetrizers.** The following terminology was borrowed from [15] and adapted to the context of finite difference schemes in [4].

Definition 4.2 (K-symmetrizer)**.** *Let $\underline{z} \in \overline{\mathscr{U}}$, and let M be a function defined on some neighborhood $\mathscr{O}$ of $\underline{z}$ with values in $\mathscr{M}_m(\mathbb{C})$ for some integer m. Then M is said to admit a K-symmetrizer at $\underline{z}$ if there exists a decomposition*

$$\mathbb{C}^m = \underline{\mathbb{E}}^s \oplus \underline{\mathbb{E}}^u ,$$

with associated projectors $(\underline{\pi}^s, \underline{\pi}^u)$, such that for all $K \geq 1$, there exists a neighborhood $\mathscr{O}_K$ of $\underline{z}$, there exists a $\mathscr{C}^\infty$ function S_K on $\mathscr{O}_K$ with values in $\mathscr{H}_m$, and there exists a constant $c_K > 0$ such that the following properties hold for all $z \in \mathscr{O}_K \cap \overline{\mathscr{U}}$:

- $M(z)^* \, S_K(z) \, M(z) - S_K(z) \geq c_K \, (|z| - 1)/|z| \, I$,
- *for all $W \in \mathbb{C}^m$, $W^* \, S_K(z) \, W \geq K^2 \, |\underline{\pi}^u \, W|^2 - |\underline{\pi}^s \, W|^2$.*

If M is a function defined on a neighborhood $\mathscr{O}$ of $\overline{\mathscr{U}}$ with values in $\mathscr{M}_m(\mathbb{C})$ for some integer m, then M is said to admit a K-symmetrizer if it admits a K-symmetrizer at all points of $\overline{\mathscr{U}}$.

We recall that in Definition 4.2, $\mathscr{H}_m$ denotes the set of Hermitian matrices of size m.

A few remarks should be made. In the decomposition as a direct sum of $\mathbb{C}^m$, $\underline{\mathbb{E}}^s$ should be thought of as the stable subpsace of $M(\underline{z})$, meaning the generalized eigenspace associated with eigenvalues in $\mathbb{D}$, and $\underline{\mathbb{E}}^u$ should be thought of as the unstable subpsace of $M(\underline{z})$, meaning the generalized eigenspace associated with eigenvalues in $\mathscr{U}$, see Lemma 4.3 below. The main difficulty arises when there are also eigenvalues on $\mathbb{S}^1$ so that one needs to determine whether such neutral eigenvalues should be counted as stable or unstable.

The goal of the symmetrizer is basically to make the matrix $M(z)^* \, S_K(z) \, M(z) - S_K(z)$ positive definite by putting a large positive weight K^2 on the unstable components and the negative weight -1 on the stable components. As explained below, this is rather easy when stable and unstable eigenvalues decouple. This decoupling occurs either when $M(\underline{z})$ has no eigenvalue on $\mathbb{S}^1$ or more generally when there is no "singular" crossing of stable and unstable eigenvalues on $\mathbb{S}^1$. The construction of the symmetrizer becomes much more involved when $M(\underline{z})$ has at least one eigenvalue on $\mathbb{S}^1$ that corresponds to such a crossing, because then one needs a precise description of how the spectrum of $M(z)$ behaves when z is close to $\underline{z}$. For the stability analysis of finite difference schemes, the reduction of $\mathbb{M}$ to the discrete block structure (Theorem 4.1) was precisely performed in order to give the information required for this construction.

Before stating the main result of this paragraph, which is Theorem 4.2 below, let us give a rather elementary result which explains some necessary properties for the existence of a K-symmetrizer.

Lemma 4.3. *Let $\underline{z} \in \mathscr{U}$, and let M be a function defined on some neighborhood $\mathscr{O}$ of $\underline{z}$ with values in $\mathscr{M}_m(\mathbb{C})$ for some integer m. If M admits a K-symmetrizer at $\underline{z}$, then $M(\underline{z})$ has no eigenvalue on $\mathbb{S}^1$. Furthermore, the vector space $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^m$ contains the generalized eigenspace associated with eigenvalues of $M(\underline{z})$ in $\mathbb{D}$.*

Lemma 4.3 shows that in the "interior" case $\underline{z} \in \mathscr{U}$ there is more or less no choice for $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^m$. For dimension reasons, the vector space $\underline{\mathbb{E}}^s$ will be chosen to be exactly the generalized eigenspace associated with eigenvalues in $\mathbb{D}$ (stable eigenvalues). There is more freedom in the choice of $\underline{\mathbb{E}}^u$ but the most natural choice will be the generalized eigenspace associated with eigenvalues in $\mathscr{U}$ (unstable eigenvalues). The limit case $\underline{z} \in \mathbb{S}^1$ will be analyzed by a continuity argument.

Proof of Lemma 4.3. Under the assumption of the Lemma, we know (apply Definition 4.2 with $K = 1$) that there exists a Hermitian matrix $\underline{S}$ such that $M(\underline{z})^* \, \underline{S} \, M(\underline{z}) - \underline{S}$ is positive definite. Here we have used the assumption $|\underline{z}| > 1$. If X is an eigenvector for $M(\underline{z})$ associated with an eigenvalue $\kappa \in \mathbb{S}^1$, we have

$$X^* \, (M(\underline{z})^* \, \underline{S} \, M(\underline{z}) - \underline{S}) \, X = (|\kappa|^2 - 1) \, X^* \, \underline{S} \, X = 0 .$$

Since $M(\underline{z})^* \, \underline{S} \, M(\underline{z}) - \underline{S}$ is positive definite, this implies $X = 0$. Hence $M(\underline{z})$ has no eigenvalue on $\mathbb{S}^1$.

Let us now consider a vector W in the generalized eigenspace of $M(\underline{z})$ associated with eigenvalues in $\mathbb{D}$. We then define a sequence $(W_j) \in \ell^2$ by the iterative formula

$$W_1 := W\,, \qquad W_{j+1} = M(\underline{z})\, W_j\,,\ j \geq 1\,.$$

For $K \geq 1$, the point $\underline{z}$ belongs to the set $\mathscr{O}_K$ on which the mapping S_K is defined. For all $j \geq 1$, there holds

$$W_j^* \left(M(\underline{z})^*\, S_K(\underline{z})\, M(\underline{z}) \right) W_j = \left(M(\underline{z})\, W_j \right)^* S_K(\underline{z})\, M(\underline{z})\, W_j = W_{j+1}^*\, S_K(\underline{z})\, W_{j+1}\,.$$

We thus get the following relations for all integer $J \geq 1$:

$$\begin{aligned} 0 &= \sum_{j=1}^{J} W_j^* \left(M(\underline{z})^*\, S_K(\underline{z})\, M(\underline{z}) \right) W_j - W_{j+1}^*\, S_K(\underline{z})\, W_{j+1} \\ &= W_1^*\, S_K(\underline{z})\, W_1 - W_{J+1}^*\, S_K(\underline{z})\, W_{J+1} + \sum_{j=1}^{J} W_j^* \Big(M(\underline{z})^*\, S_K(\underline{z})\, M(\underline{z}) - S_K(\underline{z}) \Big) W_j\,. \end{aligned}$$

Observing that the matrix $M(\underline{z})^*\, S_K(\underline{z})\, M(\underline{z}) - S_K(\underline{z})$ is positive definite and that W_{J+1} tends to 0 as J tends to infinity, we can pass to the limit with respect to J and obtain

$$W_1^*\, S_K(\underline{z})\, W_1 \leq 0\,.$$

We now use the second property of the symmetrizer S_K, see Definition 4.2, and we have thus obtained

$$|\underline{\pi}^u\, W| \leq \frac{1}{K}\, |\underline{\pi}^s\, W|\,.$$

Since the latter inequality holds for all $K \geq 1$, and the vector W as well as the projectors are independent of K, we can pass to the limit and obtain $W \in \underline{\mathbb{E}}^s$. The proof of Lemma 4.3 is complete. □

Our main result in this paragraph reads as follows. This result was partly achieved in [4] and completed in [5].

Theorem 4.2 (Existence of a K-symmetrizer [4, 5])**.** *Let Assumption 3.1 be satisfied, and let $\mathbb{M}$ defined by (57) satisfy the discrete block structure assumption. Then $\mathbb{M}$ admits a K-symmetrizer and at each point $\underline{z} \in \overline{\mathscr{U}}$, the dimension of the vector space $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^{N(p+r)}$ equals $N\,r$.*

We emphasize that at this stage, no assumption on the numerical boundary conditions has been made. More precisely, Theorem 4.1 characterizes the block structure condition by means of some properties of the operators Q_σ used in the discretization of the hyperbolic operator. According to Theorem 4.2, the existence of a K-symmetrizer is completely independent of the numerical boundary conditions used in (32). In the following paragraphs, we shall see how the result of Theorem 4.2 can be used to obtain the existence of a Kreiss symmetrizer (the terminology is introduced below). As in [15], the Kreiss symmetrizer is the main tool in showing strong stability for the numerical scheme (32). It will be obtained by using the result of Theorem 4.2 with a large enough parameter K, provided that the uniform Kreiss-Lopatinskii condition holds (see the following paragraphs for more details).

Proof of Theorem 4.2. We start the proof of Theorem 4.2 by showing two rather elementary results, the proof of which relies on some manipulations of Definition 4.2.

Lemma 4.4. *Let $\underline{z} \in \overline{\mathscr{U}}$, and let M_1, resp. M_2, be a function defined on some neighborhood $\mathscr{O}$ of $\underline{z}$ with values in $\mathscr{M}_{m_1}(\mathbb{C})$, resp. $\mathscr{M}_{m_2}(\mathbb{C})$, for some integer m_1, resp. m_2. Assume that both M_1 and M_2 admit a K-symmetrizer at $\underline{z}$ with corresponding vector spaces $\underline{\mathbb{E}}^s_1, \underline{\mathbb{E}}^s_2$ of dimension μ_1, μ_2.*

Then the block diagonal matrix $\mathrm{diag}(M_1, M_2) \in \mathscr{M}_{m_1+m_2}(\mathbb{C})$ admits a K-symmetrizer at $\underline{z}$ with a vector space $\underline{\mathbb{E}}^s$ of dimension $\mu_1 + \mu_2$.

Proof of Lemma 4.4. For all vector $W \in \mathbb{C}^{m_1+m_2}$, we let $W_1 \in \mathbb{C}^{m_1}$ denote the vector formed by the m_1 first coordinates of W and $W_2 \in \mathbb{C}^{m_2}$ the vector formed by the m_2 last coordinates of W. Then we set

$$\underline{\mathbb{E}}^s := \{W \in \mathbb{C}^{m_1+m_2} \,/\, (W_1, W_2) \in \underline{\mathbb{E}}^s_1 \times \underline{\mathbb{E}}^s_2\}\,, \quad \underline{\mathbb{E}}^u := \{W \in \mathbb{C}^{m_1+m_2} \,/\, (W_1, W_2) \in \underline{\mathbb{E}}^u_1 \times \underline{\mathbb{E}}^u_2\}\,.$$

It is straightforward to check that $\underline{\mathbb{E}}^s$ and $\underline{\mathbb{E}}^u$ are complementary vector spaces in $\mathbb{C}^{m_1+m_2}$ and that $\underline{\mathbb{E}}^s$ has dimension $\mu_1 + \mu_2$. The projectors $\underline{\pi}^s, \underline{\pi}^u$ satisfy

$$\forall\, W \in \mathbb{C}^{m_1+m_2}\,, \quad \underline{\pi}^s\, W = \begin{pmatrix} \underline{\pi}^s_1\, W_1 \\ \underline{\pi}^s_2\, W_2 \end{pmatrix}\,, \quad \underline{\pi}^u\, W = \begin{pmatrix} \underline{\pi}^u_1\, W_1 \\ \underline{\pi}^u_2\, W_2 \end{pmatrix}\,.$$

Let $K \geq 1$, and let $\mathscr{O}_K$ denote a neighborhood of $\underline{z}$ on which both mappings $S_{K,1}$, $S_{K,2}$ respectively symmetrizing M_1, M_2, are defined. For $z \in \mathscr{O}_K$, we define $S_K(z) := \mathrm{diag}(S_{K,1}(z), S_{K,2}(z)) \in \mathscr{H}_{m_1+m_2}$, and it is now a simple exercise to check that S_K satisfies all the properties required for a symmetrizer. The proof of Lemma 4.4 is therefore complete. □

Lemma 4.5. *Let $\underline{z} \in \overline{\mathscr{U}}$, and let M be a function defined on some neighborhood $\mathscr{O}$ of $\underline{z}$ with values in $\mathscr{M}_m(\mathbb{C})$ for some integer m. Assume that there exists a $\mathscr{C}^\infty$ function T defined on $\mathscr{O}$ with values in $Gl_m(\mathbb{C})$ such that $T^{-1}\, M\, T$ admits a K-symmetrizer at $\underline{z}$ with a vector space $\underline{\widetilde{\mathbb{E}}}^s$ of dimension μ.*

Then M admits a K-symmetrizer at $\underline{z}$ with a vector space $\underline{\mathbb{E}}^s$ of dimension μ.

Proof of Lemma 4.5. The proof is slightly more subtle than the proof of Lemma 4.4 but remains quite simple. First of all, since T is smooth, there is no loss of generality (up to restricting $\mathscr{O}$) in assuming that there exists a constant $c > 0$ such that for all $z \in \mathscr{O}$, there holds

$$\forall\, W \in \mathbb{C}^m\,, \quad c\,|W| \leq |T(z)^{-1}\, W| \leq \frac{1}{c}\,|W|\,. \tag{97}$$

We define the complementary vectors spaces

$$\underline{\mathbb{E}}^s := T(\underline{z})\,\underline{\widetilde{\mathbb{E}}}^s\,, \quad \underline{\mathbb{E}}^u := T(\underline{z})\,\underline{\widetilde{\mathbb{E}}}^u\,,$$

where $\underline{\widetilde{\mathbb{E}}}^s, \underline{\widetilde{\mathbb{E}}}^u$ are the complementary vector spaces given by the existence of a K-symmetrizer for $T^{-1}\, M\, T$.

Let now $K \geq 1$. We fix $\widetilde{K} \geq 1$ such that

$$\frac{1}{2}\, c^4\, \widetilde{K}^2 \geq K^2 + \frac{1}{2}\,. \tag{98}$$

For such a $\widetilde{K}$, that only depends on K, there exist a neighborhood $\mathscr{O}_K$ of $\underline{z}$, a constant $\tilde{c}_K > 0$ and a $\mathscr{C}^\infty$ mapping $\widetilde{S}_K$ defined on $\mathscr{O}_K$ with values in $\mathscr{H}_m$ such that

$$\forall\, z \in \mathscr{O}_K \cap \overline{\mathscr{U}}\,, \quad (T^{-1}\, M\, T)(z)^*\, \widetilde{S}_K(z)\, (T^{-1}\, M\, T)(z) - \widetilde{S}_K(z) \geq \tilde{c}_K\, \frac{|z|-1}{|z|}\, I\,,$$

$$\forall\, W \in \mathbb{C}^m\,, \quad W^* \widetilde{S}_K(\underline{z})\, W \geq \widetilde{K}^2\, |\underline{\widetilde{\pi}}^u\, W|^2 - |\underline{\widetilde{\pi}}^s\, W|^2\,.$$

For $z \in \mathscr{O}_K$, we define

$$S_K(z) := \frac{c^2}{2}\, (T^{-1}(z))^*\, \widetilde{S}_K(z)\, T^{-1}(z)\,,$$

and we are going to show that S_K symmetrizes M. Let $W \in \mathbb{C}^m$ be decomposed as $W = W^s + W^u$ according to the decomposition $\mathbb{C}^m = \underline{\mathbb{E}}^s \oplus \underline{\mathbb{E}}^u$. Then $T^{-1}(\underline{z})\, W^s$ and $T^{-1}(\underline{z})\, W^u$ are the components of the vector $T^{-1}(\underline{z})\, W$ according to the decomposition $\mathbb{C}^m = \underline{\widetilde{\mathbb{E}}}^s \oplus \underline{\widetilde{\mathbb{E}}}^u$. Consequently, we have

$$\begin{aligned} W^*\, S_K(\underline{z})\, W = \frac{c^2}{2}\, (T^{-1}(\underline{z})\, W)^*\, \widetilde{S}_K(\underline{z})\, T^{-1}(\underline{z})\, W &\geq \frac{c^2}{2}\, \widetilde{K}^2\, |\underline{\widetilde{\pi}}^u\, T^{-1}(\underline{z})\, W|^2 - \frac{c^2}{2}\, |\underline{\widetilde{\pi}}^s\, T^{-1}(\underline{z})\, W|^2 \\ &= \frac{c^2}{2}\, \widetilde{K}^2\, |T^{-1}(\underline{z})\, W^u|^2 - \frac{c^2}{2}\, |T^{-1}(\underline{z})\, W^s|^2\,. \end{aligned}$$

Using the estimate (97), we end up with

$$W^*\, S_K(\underline{z})\, W \geq \frac{c^4}{2}\, \widetilde{K}^2\, |W^u|^2 - \frac{1}{2}\, |W^s|^2 \geq \left(K^2 + \frac{1}{2}\right)\, |W^u|^2 - \frac{1}{2}\, |W^s|^2\,,$$

where in the end we have used the inequality (98). By continuity, up to restricting the neighborhood $\mathscr{O}_K$, there holds

$$W^* S_K(z) W \geq K^2 |W^u|^2 - |W^s|^2 ,$$

for all $z \in \mathscr{O}_K$, and therefore for all $z \in \mathscr{O}_K \cap \overline{\mathscr{U}}$. Let us now check the second property for S_K. If $z \in \mathscr{O}_K \cap \overline{\mathscr{U}}$, we have

$$\begin{aligned} M(z)^* S_K(z) M(z) - S_K(z) & \\ &= \frac{c^2}{2} \Big(M(z)^* (T^{-1}(z))^* \widetilde{S}_K(z) T^{-1}(z) M(z) - (T^{-1}(z))^* \widetilde{S}_K(z) T^{-1}(z) \Big) \\ &= \frac{c^2}{2} T^{-1}(z)^* \Big((T^{-1} M T)(z)^* \widetilde{S}_K(z) (T^{-1} M T)(z) - \widetilde{S}_K(z) \Big) T^{-1}(z) \\ &\geq \frac{c^2 \tilde{c}_K}{2} \frac{|z|-1}{|z|} T^{-1}(z)^* T^{-1}(z) \geq \frac{c^4 \tilde{c}_K}{2} \frac{|z|-1}{|z|} I , \end{aligned}$$

where we have used (97) again. The proof of Lemma 4.5 is thus complete. □

We now turn to the proof of Theorem 4.2. First of all, Lemma 4.5, combined with Lemma 4.4, shows that it is sufficient to construct a K-symmetrizer for each block of the first, second, third or fourth type arising in the discrete block structure, see Definition 4.1. If we wish the corresponding vector space $\underline{\mathbb{E}}^s$ to have dimension $N r$, it is sufficient to show that for each block $\mathbb{M}_\ell$, the corresponding vector space $\underline{\mathbb{E}}^s_\ell$ arising in the K-symmetrizer decomposition has a dimension equal to the number of stable eigenvalues of the block. More precisely, let us consider a block $\mathbb{M}_\ell(z)$ defined in the neighborhood of $\underline{z} \in \overline{\mathscr{U}}$ and occurring in the discrete block structure of $\mathbb{M}(z)$. There is no restriction in assuming that $\mathbb{M}_\ell$ is defined on the open disk $B(\underline{z}, r)$ centered at $\underline{z}$ and of radius r. In particular, the set $B(\underline{z}, r) \cap \mathscr{U}$ is connected. On $B(\underline{z}, r) \cap \mathscr{U}$, $\mathbb{M}_\ell(z)$ has no eigenvalue in $\mathbb{S}^1$ so there is no ambiguity in defining an integer μ_ℓ equal to the number of eigenvalues of $\mathbb{M}_\ell(z)$ in $\mathbb{D}$ when z belongs to $B(\underline{z}, r) \cap \mathscr{U}$ (this number is independent of z). The number μ_ℓ is called the number of stable eigenvalues of the block $\mathbb{M}_\ell$, and is made explicit below for each type of block. Lemma 3.7 shows that the sum of the μ_ℓ's equals $N r$.

• Blocks of the first type. Let $\underline{z} \in \overline{\mathscr{U}}$, and let us consider a block $\mathbb{M}_\ell(z)$ of size m_ℓ defined on a neighborhood $\mathscr{O}$ of $\underline{z}$ and satisfying $\mathbb{M}_\ell(z)^* \mathbb{M}_\ell(z) \geq (1+\delta) I$ for some constant $\delta > 0$ that is independent of z. Lemma 4.1 shows that all eigenvalues of $\mathbb{M}_\ell(z)$ belong to $\mathscr{U}$ so the number of stable eigenvalues of such a block equals zero. Let $K \geq 1$, and let us define $\underline{\mathbb{E}}^s_\ell := \{0\}$, $\underline{\mathbb{E}}^u_\ell := \mathbb{C}^{m_\ell}$. (Observe that the dimension of $\underline{\mathbb{E}}^s_\ell$ equals the number of stable eigenvalues of the block.) We also define the symmetrizer S_K as $S_K(z) := K^2 I$ independently of z. With these definitions, the relation

$$W^* S_K(z) W = K^2 |W|^2 = K^2 |\underline{\pi}^u_\ell W|^2 - |\underline{\pi}^s_\ell W|^2 , \tag{99}$$

is obvious. Moreover, there holds

$$\mathbb{M}_\ell(z)^* S_K(z) \mathbb{M}_\ell(z) - S_K(z) = K^2 \big(\mathbb{M}_\ell(z)^* \mathbb{M}_\ell(z) - I \big) \geq K^2 \delta I \geq K^2 \delta \frac{|z|-1}{|z|} I .$$

We have thus shown the existence of a K-symmetrizer at $\underline{z}$ for a block $\mathbb{M}_\ell$ of the first type.

• Blocks of the second type. Let $\underline{z} \in \overline{\mathscr{U}}$, and let us consider a block $\mathbb{M}_\ell(z)$ of size m_ℓ defined on a neighborhood $\mathscr{O}$ of $\underline{z}$ and satisfying $\mathbb{M}_\ell(z)^* \mathbb{M}_\ell(z) \leq (1-\delta) I$ for some $\delta > 0$ that is independent of z. Lemma 4.1 shows again that all eigenvalues of $\mathbb{M}_\ell(z)$ belong to $\mathbb{D}$ so the number of stable eigenvalues of such a block equals m_ℓ. Let $K \geq 1$, and let us define $\underline{\mathbb{E}}^s_\ell := \mathbb{C}^{m_\ell}$, $\underline{\mathbb{E}}^u_\ell := \{0\}$. We also define the symmetrizer S_K as $S_K(z) := -I$ independently of z, and the reader can easily adapt the argument developed for blocks of the first type to show that S_K satisfies all the properties required for a symmetrizer. We observe again that the dimension of $\underline{\mathbb{E}}^s_\ell$ equals the number of stable eigenvalues of the block.

• Blocks of the third type (part I). We recall from Definition 4.1 that blocks of the third type are scalar and can only occur for $\underline{z} \in \mathbb{S}^1$. We thus consider a holomorphic function $\mathbb{M}_\ell$ defined on a neighborhood $\mathscr{O}$ of $\underline{z} \in \mathbb{S}^1$ and satisfying $\mathbb{M}_\ell(\underline{z}) \in \mathbb{S}^1$, $\underline{z}\, \mathbb{M}'_\ell(\underline{z}) \overline{\mathbb{M}_\ell(\underline{z})} > 0$. (According to Definition

4.1, $\underline{z}\,\mathbb{M}_\ell'(\underline{z})\,\overline{\mathbb{M}_\ell(\underline{z})}$ is a nonzero real number so we first consider the case where this number is positive.) Let us first show that there is no stable eigenvalue in that case. For $\varepsilon > 0$ small enough, $(1+\varepsilon)\,\underline{z}$ belongs to $\mathscr{O} \cap \mathscr{U}$ and Taylor's expansion reads

$$\frac{\mathbb{M}_\ell((1+\varepsilon)\,\underline{z})}{\mathbb{M}_\ell(\underline{z})} = 1 + \underline{z}\,\mathbb{M}_\ell'(\underline{z})\,\overline{\mathbb{M}_\ell(\underline{z})}\,\varepsilon + O(\varepsilon^2)\,.$$

In particular, the modulus of $\mathbb{M}_\ell((1+\varepsilon)\,\underline{z})$ is larger than 1 for $\varepsilon > 0$ small enough and there is no stable eigenvalue for such a scalar block. Unsurprisingly, we thus define $\mathbb{E}_\ell^s := \{0\}$, $\mathbb{E}_\ell^u := \mathbb{C}$, and $S_K(z) := K^2$ independently of z. This symmetrizer trivially satisfies the property (99). Following the analysis performed above for blocks of the first type, the result relies on a lower bound of the quantity $|\mathbb{M}_\ell(z)|^2 - 1$ for $z \in \mathscr{O} \cap \overline{\mathscr{U}}$. This lower bound is derived in the following Lemma which we state separately for the sake of clarity.

Lemma 4.6. *Let f be a holomorphic function defined on a disk $B(1,r)$ centered at 1 and of radius $r > 0$, verifying $f(1) = 1$, $\mathrm{Re}\, f'(1) > 0$, and*

$$\forall\, z \in B(1,r) \cap \mathbb{S}^1\,, \quad |f(z)| \geq 1\,.$$

Then there exists a constant $c > 0$ such that, up to diminishing r, there holds

$$\forall\, z \in B(1,r) \cap \overline{\mathscr{U}}\,, \quad |f(z)|^2 - 1 \geq c\,(|z| - 1)\,.$$

Proof of Lemma 4.6. For τ in a sufficiently small neighborhood of 0, we define:

$$h(\tau) := \ln f(\mathrm{e}^\tau)\,,$$

where ln denotes the standard complex logarithm defined on $\mathbb{C} \setminus \mathbb{R}^-$. We have $h'(0) = f'(1)$, and $h(\tau)$ has nonnegative real part when τ is purely imaginary. Using the notation $\tau = x + i\,y$, a direct Taylor expansion yields

$$\begin{aligned}\mathrm{Re}\; h(\tau) = \mathrm{Re}\; h(i\,y) + \mathrm{Re}\; (h(\tau) - h(i\,y)) \geq \mathrm{Re}\; (h(\tau) - h(i\,y)) &= \mathrm{Re}\; (h'(i\,y)\,x) + o(x) \\ &= (\mathrm{Re}\; f'(1))\,x + o(x)\,,\end{aligned}$$

where the last equality holds for sufficiently small r (and the smallness condition only depends on f). We have thus shown the estimate

$$\mathrm{Re}\; h(\tau) \geq \frac{\mathrm{Re}\; f'(1)}{2}\;\; \mathrm{Re}\; \tau\,,$$

for all τ of nonnegative real part close to 0. The estimate for $|f(z)|^2$ for $z \in B(1,r) \cap \overline{\mathscr{U}}$ easily follows:

$$|f(z)|^2 - 1 = (|f(z)| + 1)\,(|f(z)| - 1) = (|f(z)| + 1)\left(\mathrm{e}^{\mathrm{Re}\; h(\ln z)} - 1\right) \geq \frac{\mathrm{Re}\; f'(1)}{2}\;\; \mathrm{Re}\;\; \ln z\,.$$

□

Remark 4.1. *The assumption $|f(z)| \geq 1$ for all $z \in B(1,r) \cap \mathbb{S}^1$ is absolutely necessary in Lemma 4.6, and it is no consequence of the assumption $\mathrm{Re}\, f'(1) > 0$. The reader may for instance consider the example*

$$f(z) := 1 + (z-1) + \left(\frac{1}{2} + i\right)(z-1)^2\,,$$

which satisfies $f(1) = 1$, $f'(1) = 1$. However, if one considers the points $z_\alpha := 1 + i\,\alpha$, with $\alpha > 0$ small enough, there holds $|f(z_\alpha)|^2 - 1 < 0$ and $z_\alpha \in \mathscr{U}$. This prevents f from verifying the conclusion of Lemma 4.6.

More generally, the property $|f(z)| \geq 1$ for all $z \in B(1,r) \cap \mathbb{S}^1$ can not follow from any information on a finite number of derivatives of f at 1. In general, this property can only follow from the full series expansion of f at 1.

We can apply Lemma 4.6 to the function $w \mapsto \mathbb{M}_\ell(\underline{z}\,w)/\mathbb{M}_\ell(\underline{z})$. Indeed, we know that $\mathbb{M}_\ell(z)$ belongs to $\mathscr{U}$ for all $z \in \mathscr{O} \cap \mathscr{U}$. By continuity, this implies $\mathbb{M}_\ell(z) \in \overline{\mathscr{U}}$ for all $z \in \mathscr{O} \cap \overline{\mathscr{U}}$. We therefore obtain the estimate

$$\mathbb{M}_\ell(z)^*\, S_K(z)\, \mathbb{M}_\ell(z) - S_K(z) = K^2\left(|\mathbb{M}_\ell(z)|^2 - 1\right) \geq c\,K^2\,(|z| - 1) \geq c\,K^2\,\frac{|z| - 1}{|z|}\,,$$

for all $z \in \mathscr{O} \cap \overline{\mathscr{U}}$ sufficiently close to $\underline{z}$. We have proved that S_K satisfies all the properties of a symmetrizer, and the dimension of $\underline{\mathbb{E}}^s_\ell$ coincides with the number of stable eigenvalues of the block.

• Blocks of the third type (part II). We now turn to the case $\underline{z} \in \mathbb{S}^1$, $\mathbb{M}_\ell(\underline{z}) \in \mathbb{S}^1$, $\underline{z}\,\mathbb{M}'_\ell(\underline{z})\,\overline{\mathbb{M}_\ell(\underline{z})} < 0$. Unsurprisingly, the reader will easily verify that there is one stable eigenvalue and that the symmetrizer S_K can be chosen as $S_K(z) := -1$ independently of z. The argument relies on the following analogue of Lemma 4.6, which we feel free to use without proof.

Lemma 4.7. *Let f be a holomorphic function defined on a disk $B(1,r)$ centered at 1 and of radius $r > 0$, verifying $f(1) = 1$, $\mathrm{Re}\, f'(1) < 0$, and*

$$\forall\, z \in B(1,r) \cap \mathbb{S}^1\,, \quad |f(z)| \leq 1\,.$$

Then there exists a constant $c > 0$ such that, up to diminishing r, there holds

$$\forall\, z \in B(1,r) \cap \overline{\mathscr{U}}\,, \quad |f(z)|^2 - 1 \leq -c\,(|z| - 1)\,.$$

• Blocks of the fourth type. This is by far the most difficult case. A complete analysis of the construction of the symmetrizer is performed in [5]. The analysis is unfortunately very long, and involves a generalization of the original construction performed in [13]. In order to keep the length of these notes reasonable, we shall not detail the construction of the symmetrizer for blocks of the fourth type and we shall rather refer to [5, Theorem 3.4]. In particular, the dimension of the corresponding vector space $\underline{\mathbb{E}}^s_\ell$ equals the number of stable eigenvalues of the block. This number can be explicitly determined from the size ν_ℓ of the block and the lower left coefficient m_ℓ of $\mathbb{M}'_\ell(\underline{z})$, see [5] for more details.

We just emphasize for the interested reader the new main difficulty compared with [13]. In the analysis of [13], which is devoted to boundary value problems for hyperbolic systems of partial differential equations, the construction of the symmetrizer relies on the fact[14] that for $z \in \mathbb{S}^1$ close to $\underline{z}$, all eigenvalues of the block belong to $\mathbb{S}^1$. This is a very strong property which implies that some coefficients in the matrices are either real or purely imaginary. In our framework, there is a lot more freedom because we only know that for $z = \underline{z}$, $\mathbb{M}_\ell(\underline{z})$ has one eigenvalue on $\mathbb{S}^1$. When z varies on $\mathbb{S}^1$ close to $\underline{z}$, the eigenvalues of $\mathbb{M}_\ell(z)$ usually do not stay on $\mathbb{S}^1$. This phenomenon can be checked by hand on the following elementary example[15]:

$$\underline{z} = \underline{\kappa} = 1\,, \quad M(z) := \begin{pmatrix} 1 & 1 \\ z-1 & 1 \end{pmatrix}.$$

Other examples of this behavior occur for discretizations of the hyperbolic operator whose amplification matrix displays some eigenvalues curves with singular points on $\mathbb{S}^1$. Examples of such discretizations were given in Section 2. As a matter of fact, when singular points in $\mathbb{S}^1$ occur for eigenvalues of the amplification matrix $\mathscr{A}(\kappa)$, this gives rise in the reduction of $\mathbb{M}$ to blocks of the fourth type, see the proof of Theorem 4.1. Unless the behavior of the eigenvalues corresponds to that of the leap-frog scheme, see Figure 1, the eigenvalues of the block in the reduction of $\mathbb{M}$ can have a much more complex behavior than just remaining on $\mathbb{S}^1$ for $z \in \mathbb{S}^1$. This led us in [5] to introducing an integer which we called the *dissipation index* and that gave a description of the singularity for the eigenvalue curve for $\mathscr{A}$. The construction of the symmetrizer for a block of the fourth type depends both on the size of the block and of the dissipation index (there are approximately ten cases to deal with). Even though we shall not reproduce the complete analysis here, we strongly encourage the reader to go through [5] since we believe that this new construction is basically the first step towards a full treatment of the analogous problem for multidimensional problems. This extension is postponed to a future work. □

The symmetrizer construction performed in this paragraph will be crucial for the proof of Theorems 3.5 and 3.6. However, before giving the proof of Theorem 3.5, we need one last technical - though crucial - point about the behavior of the stable subspace $\mathbb{E}^s(z)$ when $z \in \mathscr{U}$ tends to a point of $\mathbb{S}^1$.

[14] We slightly adapt the result of [13] to our framework but there is no difficulty to pass from one to the other thanks to the exponential function.

[15] On this example, the reader can check that the eigenvalues of $M(\mathrm{e}^{i\,\varepsilon})$, $\varepsilon > 0$ small, do not belong to $\mathbb{S}^1$.

4.3. **Extending the stable subspace.** The main result of this paragraph is the following.

Theorem 4.3 (Continuous extension of the stable subspace [4])**.** *Let Assumption 3.1 be satisfied, and let us assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2. Let us also assume that the matrix $\mathbb{M}$ defined by (57) admits a K-symmetrizer where, at each point $\underline{z} \in \overline{\mathscr{U}}$, the dimension of the vector space $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^{N(p+r)}$ equals $N\,r$.*

Then the stable subspace $\mathbb{E}^s(z)$ of $\mathbb{M}(z)$, which is well-defined for $z \in \mathscr{U}$ according to Lemma 3.7, defines a holomorphic vector bundle over $\mathscr{U}$ that can be extended in a unique way as a continuous vector bundle over $\overline{\mathscr{U}}$.

In all what follows, we shall let $\mathbb{E}^s(z)$ denote the continuous extension of the stable subspace for $z \in \mathbb{S}^1 (= \partial\mathscr{U})$. In general, for $z \in \mathbb{S}^1$, the matrix $\mathbb{M}(z)$ may have eigenvalues on $\mathbb{S}^1$, so the number of eigenvalues in $\mathbb{D}$ can be less than $N\,r$. As was already pointed out in the proof of Theorem 4.2, the difficulty consists in determining whether eigenvalues on $\mathbb{S}^1$ should count as stable or unstable eigenvalues, and this is determined by a perturbation argument, that is by slightly moving z towards the open set $\mathscr{U}$ and by studying whether the eigenvalues move towards $\mathbb{D}$ or towards $\mathscr{U}$. The cases of the Lax-Friedrichs and leap-frog schemes are detailed below.

Proof of Theorem 4.3. Lemma 3.7 shows that the stable subspace $\mathbb{E}^s(z)$ of $\mathbb{M}(z)$ has constant dimension $N\,r$ for all $z \in \mathscr{U}$. The holomorphic dependence of $\mathbb{M}(z)$ on z implies that $\mathbb{E}^s(z)$ also varies holomorphically with z on $\mathscr{U}$. (Here we use the same arguments as in the proof of Lemma 2.6 and Theorem 4.1: the spectral projector on $\mathbb{E}^s(z)$ is given by the Dunford-Taylor formula, which shows that the projector depends holomorphically on z. We can then construct a basis of $\mathbb{E}^s(z)$ that depends holomorphically on z in the neighborhood of any point of $\mathscr{U}$. In other words, $\mathbb{E}^s$ defines a holomorphic vector bundle over $\mathscr{U}$.)

Let $\underline{z} \in \mathbb{S}^1$ and let us first show that $\mathbb{E}^s(z)$ has a limit as $z \in \mathscr{U}$ tends to $\underline{z}$. We consider the decomposition $\mathbb{C}^{N(p+r)} = \underline{\mathbb{E}}^s \oplus \underline{\mathbb{E}}^u$ given by the existence of a K-symmetrizer at $\underline{z}$. From the assumption of Theorem 4.3, we know that the dimension of $\underline{\mathbb{E}}^s$ equals $N\,r$. Let now $K > 2$, and let us consider a neighborhood $\mathscr{O}_K$ of $\underline{z}$ and a symmetrizer S_K defined on $\mathscr{O}_K$ and satisfying the properties given in definition 4.2. Let $z \in \mathscr{O}_K \cap \mathscr{U}$ and let $W \in \mathbb{E}^s(z)$. We define the sequence:

$$W_1 := W\,, \qquad W_{j+1} = \mathbb{M}(z)\,W_j\,, \quad j \geq 1\,.$$

Using the exact same method as in the proof of Lemma 4.3, we end up with the inequality $W_1^*\,S_K(z)\,W_1 \leq 0$, which in turn yields:

$$\forall\, z \in \mathscr{O}_K \cap \mathscr{U}\,, \quad \forall\, W \in \mathbb{E}^s(z)\,, \quad K\,|\underline{\pi}^u\,W| \leq |\underline{\pi}^s\,W|\,.$$

The rest of the analysis follows [15]. Writing $\underline{\pi}^s\,W = W - \underline{\pi}^u\,W$, we get (use the triangle inequality)

$$\forall\, z \in \mathscr{O}_K \cap \mathscr{U}\,, \quad \forall\, W \in \mathbb{E}^s(z)\,, \quad (K-1)\,|\underline{\pi}^u\,W| \leq |W|\,. \tag{100}$$

The estimate (100) shows that the mapping

$$\begin{aligned} \Phi(z)\ :\ \mathbb{E}^s(z) &\longrightarrow \underline{\mathbb{E}}^s \\ W &\longmapsto \underline{\pi}^s\,W\,, \end{aligned}$$

which is defined for $z \in \mathscr{O}_K \cap \mathscr{U}$, is injective. (If W belongs to the kernel of $\Phi(z)$, then W belongs to $\mathbb{E}^s(z) \cap \underline{\mathbb{E}}^u$ and (100) gives $(K-1)\,|W| \leq |W|$ so W is zero because K is larger than 2.) Since the dimensions of $\mathbb{E}^s(z)$ and $\underline{\mathbb{E}}^s$ are the same, $\Phi(z)$ is an isomorphism. We can write the inverse mapping $\Phi(z)^{-1}$ in the following way

$$\begin{aligned} \Phi(z)^{-1}\ :\quad \underline{\mathbb{E}}^s &\longrightarrow \mathbb{E}^s(z) \\ \underline{W} &\longmapsto \underline{W} + \varphi(z)\,\underline{W}\,, \end{aligned}$$

where $\varphi(z)$ is a linear mapping from $\underline{\mathbb{E}}^s$ to $\underline{\mathbb{E}}^u$. This may look suprising but we only decompose the vector $\Phi(z)^{-1}\,\underline{W}$ along the direct sum $\underline{\mathbb{E}}^s \oplus \underline{\mathbb{E}}^u$ and we observe that the component on $\underline{\mathbb{E}}^s$ equals $\underline{W}$ itself (use the definition of $\Phi(z)$). Using (100) once again, we obtain

$$\forall\, z \in \mathscr{O}_K \cap \mathscr{U}\,, \quad \forall\, \underline{W} \in \underline{\mathbb{E}}^s\,, \quad |\varphi(z)\,\underline{W}| \leq \frac{1}{K-2}\,|\underline{W}|\,. \tag{101}$$

Indeed, (100) shows that for all $\underline{W} \in \underline{\mathbb{E}}^s$, there holds

$$(K-1)\,|\varphi(z)\,\underline{W}| = (K-1)\,|\underline{\pi}^u\,(\underline{W}+\varphi(z)\,\underline{W})| \leq |\underline{W}+\varphi(z)\,\underline{W}| \leq |\underline{W}| + |\varphi(z)\,\underline{W}|\,,$$

and (101) follows (use $K>2$).

We now have all the ingredients in order to show that $\mathbb{E}^s(z)$ tends to $\underline{\mathbb{E}}^s$ as $z \in \mathscr{U}$ tends to $\underline{z}$. We consider a basis $(\underline{e}_1,\dots,\underline{e}_{Nr})$ of $\underline{\mathbb{E}}^s$ and we fix $\varepsilon>0$. Let us choose $K>2$ such that $|\underline{e}_j|/(K-2) \leq \varepsilon$ for all $j=1,\dots,N\,r$. The above analysis shows that the estimate (101) holds for all $z \in \mathscr{O}_K \cap \mathscr{U}$. In particular, we have

$$\forall\, z \in \mathscr{O}_K \cap \mathscr{U}\,, \quad \forall\, j=1,\dots,N\,r\,, \quad |\underline{e}_j - \Phi(z)^{-1}\,\underline{e}_j| \leq \varepsilon\,.$$

We know that $\Phi(z)^{-1}$ is an isomorphism so the family $(\Phi(z)^{-1}\,\underline{e}_1,\dots,\Phi(z)^{-1}\,\underline{e}_{Nr})$ is a basis of $\mathbb{E}^s(z)$. We have thus proved that for $z \in \mathscr{U}$ sufficiently close to $\underline{z}$, there exists a basis of $\mathbb{E}^s(z)$ whose elements are ε-close to the elements of a basis of $\underline{\mathbb{E}}^s$. In other words, we have shown that $\mathbb{E}^s(z)$ tends to $\underline{\mathbb{E}}^s$ as $z \in \mathscr{U}$ tends to $\underline{z}$. This means that the vector bundle $\mathbb{E}^s$ can be extended to $\overline{\mathscr{U}}$, and it remains to show that this extended bundle is continuous over $\overline{\mathscr{U}}$. This is not straightforward because continuity at $\underline{z} \in \mathbb{S}^1$ now requires to consider the limit of $\mathbb{E}^s(z)$ when $z \in \overline{\mathscr{U}}$ tends to $\underline{z}$, while before we have only studied the limit of $\mathbb{E}^s(z)$ when $z \in \mathscr{U}$ tends to $\underline{z}$.

Let us observe that the above argument shows that for $\underline{z} \in \mathbb{S}^1$, the vector space $\underline{\mathbb{E}}^s$ of dimension $N\,r$ in the decomposition of $\mathbb{C}^{N(p+r)}$ is necessarily unique since it is the limit of $\mathbb{E}^s((1+\varepsilon)\,\underline{z})$ as $\varepsilon>0$ tends to 0.

Let us now prove that the bundle $\mathbb{E}^s$, which has been extended to $\partial\mathscr{U}$, is continuous over $\overline{\mathscr{U}}$. It is obviously continuous over $\mathscr{U}$ since it is holomorphic, and we thus only check the continuity of $\mathbb{E}^s$ at any point of $\mathbb{S}^1$. We follow [15] again and perform more or less the same analysis as above. We use the convention introduced above and let $\mathbb{E}^s(z)$ denote the continuous extension of the stable subspace for $z \in \mathbb{S}^1 (= \partial\mathscr{U})$. Let $\underline{z} \in \mathbb{S}^1$, and let $K>2$. With the above argument, we already have the estimate (100). Furthermore, there is no loss of generality in assuming that the neighborhood $\mathscr{O}_K$ of $\underline{z}$ is an open disk $B(\underline{z}, r_K)$, $r_K>0$.

Let us consider a point $\underline{z}' \in \mathscr{O}_K \cap \mathbb{S}^1$. Since $\mathscr{O}_K$ is an open neighborhood of $\underline{z}'$, there exists a sequence (z_n) in $\mathscr{O}_K \cap \mathscr{U}$ that converges towards $\underline{z}'$. In particular, the above analysis shows that $\mathbb{E}^s(z_n)$ converges towards $\mathbb{E}^s(\underline{z}')$. This means that any element $W' \in \mathbb{E}^s(\underline{z}')$ can be written as the limit - in $\mathbb{C}^{N(p+r)}$ - of a sequence (W_n) where for each integer n, W_n belongs to $\mathbb{E}^s(z_n)$. Applying (100) and passing to the limit as n tends to infinity, we get the inequality $(K-1)\,|\underline{\pi}^u\,W'| \leq |W'|$ for all $W' \in \mathbb{E}^s(\underline{z}')$. In other words, we have obtained

$$\forall\, z \in \mathscr{O}_K \cap \overline{\mathscr{U}}\,, \quad \forall\, W \in \mathbb{E}^s(z)\,, \quad (K-1)\,|\underline{\pi}^u\,W| \leq |W|\,. \tag{102}$$

(Observe the slight, though important, difference between (100) and (102).) At this point, the exact same argument as above shows that $\mathbb{E}^s(z)$ tends to $\mathbb{E}^s(\underline{z})$ as $z \in \overline{\mathscr{U}}$ tends to $\underline{z}$. The only difference is that we are now allowed to consider some $z \in \mathscr{O}_K$ that belong to $\mathbb{S}^1$ and use (102) while before we were only allowed to consider some $z \in \mathscr{O}_K$ that belonged to $\mathscr{U}$ and use (100). Eventually, we have proved that $\mathbb{E}^s$ is continuous at any point of $\mathbb{S}^1$. □

Here we have followed the approach of [15], which gives an "analytical" and somehow simple proof of the continuous extension of the stable bundle. As observed in [15], the nice point is that constructing a symmetrizer seems to be necessary to deal with the derivation of a priori estimates for solutions to the resolvent equation. In the original approach by Kreiss [13], see also the books [2, 3], the first step consisted in first showing through mostly "algebraic" arguments that the stable subspace could be continuously extended and then in constructing a symmetrizer. The alternative approach introduced in [15] bypasses the algebraic part of the proof and focuses on the symmetrizer construction. The continuous extension of the stable bundle appears as a corollary of the existence of a symmetrizer (which itself relies on the block structure). From our point of view, this alternative approach clarifies one of the main technical and difficult points of the theory. The main remaining difficulties are the (i) reduction of the symbol $\mathbb{M}$ to the discrete block structure and (ii) the construction of the symmetrizer. This technical simplification gives us hope to deal with multidimensional problems in a near future.

4.4. **Proof of Theorem 3.5.** We first give a new formulation of the Uniform Kreiss-Lopatinskii Condition in the framework of Theorem 3.5.

Proposition 4.1 (Reformulation of the UKLC)**.** *Under the assumptions of Theorem 3.5, the UKLC holds if and only if*

$$\forall\, z \in \overline{\mathscr{U}}\,, \quad \operatorname{Ker} \mathbb{B}(z) \cap \mathbb{E}^s(z) = \{0\}\,,$$

where $\mathbb{E}^s(z)$ denotes the generalized eigenspace of $\mathbb{M}(z)$ associated with eigenvalues in $\mathbb{D}$, which is defined in Lemma 3.7 for $z \in \mathscr{U}$ and is continuously extended to $z \in \mathbb{S}^1$.

We observe again that the UKLC is compatible with the dimensions of the vector spaces: $\mathbb{E}^s(z)$ has dimension $N\,r$, while $\mathbb{B}(z) \in \mathscr{M}_{Nr,N(p+r)}(\mathbb{C})$ has maximal rank (see the expression (58)) so its kernel has dimension $N\,p$. Hence there is no obstruction for $\operatorname{Ker} \mathbb{B}(z)$ and $\mathbb{E}^s(z)$ to be complementary in $\mathbb{C}^{N(p+r)}$.

Proof of Proposition 4.1. Let us first verify that the stable subspace $\mathbb{E}^s$ can be continuously extended to the boundary $\mathbb{S}^1$ of $\mathscr{U}$. Applying first Theorem 4.1, we know that the matrix $\mathbb{M}$ defined by (57) satisfies the discrete block structure condition. We can then apply Theorem 4.2: $\mathbb{M}$ admits a K-symmetrizer where, at each point of $\overline{\mathscr{U}}$, the dimension of the vector space $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^{N(p+r)}$ equals $N\,r$. Eventually Theorem 4.3 shows that the stable subspace extends continuously to $\mathbb{S}^1$, and the extended bundle is continuous over $\overline{\mathscr{U}}$.

• We now prove the result of Proposition 4.1. We first assume that the UKLC is satisfied, meaning that for all $R \geq 2$, there exists a constant $C_R > 0$ such that for all $z \in \mathscr{U}$ with $|z| \leq R$, the estimate (76) holds with the matrix $\mathbb{B}(z)$ defined in (58). We let C_2 denote the corresponding constant for $R = 2$. It is already clear that $\mathbb{E}^s(z)$ does not intersect the kernel of $\mathbb{B}(z)$ for $z \in \mathscr{U}$ (this is the Godunov-Ryabenkii condition). We thus consider $z_0 \in \mathbb{S}^1$. The space $\mathbb{E}^s(z_0)$ is the limit, as $\varepsilon > 0$ tends to 0, of $\mathbb{E}^s((1+\varepsilon)\,z_0)$. Any vector $\mathscr{W} \in \mathbb{E}^s(z_0)$ can thus be written as the limit, as $\varepsilon > 0$ tends to 0, of a sequence of vectors $\mathscr{W}_\varepsilon \in \mathbb{E}^s((1+\varepsilon)\,z_0)$. Passing to the limit in the inequality

$$\forall\, \varepsilon \in]0,1]\,, \quad |\mathscr{W}_\varepsilon| \leq C_2\, |\mathbb{B}((1+\varepsilon)\,z_0)\,\mathscr{W}_\varepsilon|\,,$$

we obtain the inequality $|\mathscr{W}| \leq C_2\,|\mathbb{B}(z_0)\,\mathscr{W}|$ for all $\mathscr{W} \in \mathbb{E}^s(z_0)$. This property implies that $\mathbb{E}^s(z)$ does not intersect the kernel of $\mathbb{B}(z)$ for all $z \in \mathbb{S}^1$.

• We now assume that $\mathbb{E}^s(z)$ does not intersect the kernel of $\mathbb{B}(z)$ for all $z \in \overline{\mathscr{U}}$ and we are going to show that the UKLC holds. Let $R \geq 2$. For $z \in \overline{\mathscr{U}}$ with $|z| \leq R$, we consider the quantity

$$m(z) := \inf_{\mathscr{W} \in \mathbb{E}^s(z), |\mathscr{W}|=1} \left|\mathbb{B}(z)\,\mathscr{W}\right|.$$

The quantity $m(z)$ is positive for all z, and m depends continuously on z because both the vector space $\mathbb{E}^s(z)$ and the matrix $\mathbb{B}(z)$ depend continuously on z. Since the annulus $\{z \in \mathbb{C}\,,\, 1 \leq |z| \leq R\}$ is compact, m is bounded from below by a positive constant $c_R > 0$ on this annulus. In other words, we have shown the inequality

$$\forall\, \mathscr{W} \in \mathbb{E}^s(z)\,, \quad |\mathscr{W}| \leq \frac{1}{c_R}\,|\mathbb{B}(z)\,\mathscr{W}|\,,$$

as long as $1 \leq |z| \leq R$. Consequently the UKLC is satisfied. □

We introduce the following terminology.

Definition 4.3 (Kreiss symmetrizer)**.** *Let $\mathbb{M}$ be defined by (57), and let $\mathbb{B}$ be defined by (58). The pair $(\mathbb{M}, \mathbb{B})$ is said to admit a Kreiss symmetrizer if for all $R \geq 2$, there exists a constant $c_R > 0$ and there exists a $\mathscr{C}^\infty$ function S on the annulus $\{z \in \mathbb{C}\,,\, 1 \leq |z| \leq R\}$ with values in $\mathscr{H}_{N(p+r)}$ such that the following properties hold for all z in the annulus:*

- $\mathbb{M}(z)^*\, S(z)\, \mathbb{M}(z) - S(z) \geq c_R\,(|z|-1)/|z|\, I$,
- *for all $\mathscr{W} \in \mathbb{C}^{N(p+r)}$, $\mathscr{W}^*\, S(z)\, \mathscr{W} \geq c_R\, |\mathscr{W}|^2 - c_R^{-1}\, |\mathbb{B}(z)\,\mathscr{W}|^2$.*

We can now prove a refined version of Theorem 3.5.

Theorem 4.4 (Existence of a Kreiss symmetrizer and strong stability)**.** *Let Assumption 3.1 be satisfied, let us assume $q < p$ and let us further assume that the discretization of the Cauchy problem (14) is stable in the sense of Definition 2.2 and that the operators Q_σ are geometrically regular in the sense of Definition 2.3.*

If the UKLC holds, then the pair $(\mathbb{M}, \mathbb{B})$ admits a Kreiss symmetrizer and the scheme (32) is strongly stable.

The assumptions of Theorem 4.4 are exactly the same as the assumptions of Theorem 3.5. It should be rather clear at this point that Theorem 4.4 yields the result of Theorem 3.5. Indeed, Theorem 4.4 shows that the UKLC is a sufficient condition for strong stability (it even shows that the UKLC is a sufficient condition for the existence of a Kreiss symmetrizer). In the meantime, Corollary 3.2 shows that the UKLC is a necessary condition for strong stability. We thus focus on the proof of Theorem 4.4.

Proof of Theorem 4.4. • We first show that under the assumptions of Theorem 4.4, the pair $(\mathbb{M}, \mathbb{B})$ admits a Kreiss symmetrizer. Following the same arguments as in the proof of Proposition 4.1, we already know that $\mathbb{M}$ admits a K-symmetrizer where, at each point of $\overline{\mathscr{U}}$, the dimension of the vector space $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^{N(p+r)}$ equals $N\,r$. Lemma 4.3 and Lemma 3.7 show that at each point $\underline{z} \in \mathscr{U}$, the vector space $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^{N(p+r)}$ coincides with $\mathbb{E}^s(\underline{z})$. Furthermore, the proof of Theorem 4.2 shows that this property holds true also on the boundary $\mathbb{S}^1$ of $\mathscr{U}$. Summarizing, $\mathbb{M}$ admits a K-symmetrizer in the sense of Definition 4.2 where, at each point $\underline{z} \in \overline{\mathscr{U}}$, the vector space $\underline{\mathbb{E}}^s$ in the decomposition of $\mathbb{C}^{N(p+r)}$ equals $\mathbb{E}^s(\underline{z})$.

Let $R \geq 2$, and let $\underline{z} \in \overline{\mathscr{U}}$ with $|\underline{z}| \leq R$. We are going to show that the pair $(\mathbb{M}, \mathbb{B})$ admits a Kreiss symmetrizer in the neighborhood of $\underline{z}$. More precisely, since the UKLC holds, Proposition 4.1 shows that there exists a constant $\underline{c} > 0$ such that

$$\forall\, \mathscr{W} \in \mathbb{E}^s(\underline{z})\,, \quad \underline{c}\,|\mathscr{W}| \leq |\mathbb{B}(\underline{z})\,\mathscr{W}|\,.$$

We fix a parameter $K \geq 1$ by choosing $K^2 := 1 + 4\,|\mathbb{B}(\underline{z})|^2/\underline{c}^2$. Applying Theorem 4.2, we know that $\mathbb{M}$ admits a K-symmetrizer at $\underline{z}$ so there exists a neighborhood $\mathscr{O}$ of $\underline{z}$, a constant $c > 0$, and a $\mathscr{C}^\infty$ function S on $\mathscr{O}$ with values in $\mathscr{H}_{N(p+r)}$ such that for all $z \in \mathscr{O} \cap \overline{\mathscr{U}}$, there holds

$$\mathbb{M}(z)^*\, S(z)\, \mathbb{M}(z) - S(z) \geq c\,(|z| - 1)/|z|\, I\,, \tag{103}$$

and

$$\forall\, \mathscr{W} \in \mathbb{C}^{N(p+r)}\,, \quad \mathscr{W}^*\, S(\underline{z})\, \mathscr{W} \geq K^2\, |\underline{\pi}^u\, \mathscr{W}|^2 - |\underline{\pi}^s\, \mathscr{W}|^2\,.$$

In particular, we have

$$\begin{aligned}
\mathscr{W}^*\, S(\underline{z})\, \mathscr{W} &\geq |\underline{\pi}^s\, \mathscr{W}|^2 + K^2\, |\underline{\pi}^u\, \mathscr{W}|^2 - 2\,|\underbrace{\underline{\pi}^s\, \mathscr{W}}_{\in \mathbb{E}^s(\underline{z})}|^2 \\
&\geq |\underline{\pi}^s\, \mathscr{W}|^2 + K^2\, |\underline{\pi}^u\, \mathscr{W}|^2 - \frac{2}{\underline{c}^2}\, \big|\mathbb{B}(\underline{z})\,(\mathscr{W} - \underline{\pi}^u\, \mathscr{W})\big|^2 \\
&\geq |\underline{\pi}^s\, \mathscr{W}|^2 + K^2\, |\underline{\pi}^u\, \mathscr{W}|^2 - \frac{4}{\underline{c}^2}\, \Big(|\mathbb{B}(\underline{z})\, \mathscr{W}|^2 + |\mathbb{B}(\underline{z})\, \underline{\pi}^u\, \mathscr{W}|^2\Big)\,.
\end{aligned}$$

With our choice of the parameter K, we get

$$\mathscr{W}^*\, S(\underline{z})\, \mathscr{W} \geq |\underline{\pi}^s\, \mathscr{W}|^2 + |\underline{\pi}^u\, \mathscr{W}|^2 - \frac{4}{\underline{c}^2}\, |\mathbb{B}(\underline{z})\, \mathscr{W}|^2 \geq \frac{1}{2}\, |\mathscr{W}|^2 - \frac{4}{\underline{c}^2}\, |\mathbb{B}(\underline{z})\, \mathscr{W}|^2\,.$$

In particular, the matrix $S(\underline{z}) + 4\,\underline{c}^{-2}\, \mathbb{B}(\underline{z})^*\, \mathbb{B}(\underline{z}) - I/4$ is positive definite so, by a continuity argument, for all z sufficiently close to $\underline{z}$, there holds

$$\forall\, \mathscr{W} \in \mathbb{C}^{N(p+r)}\,, \quad \mathscr{W}^*\, S(z)\, \mathscr{W} \geq c\,|\mathscr{W}|^2 - \frac{1}{c}\, |\mathbb{B}(z)\, \mathscr{W}|^2\,, \tag{104}$$

with a suitable constant $c > 0$ that is independent of z. To summarize, we have proved that for all $\underline{z}$ in the annulus $\{z \in \mathbb{C}\,,\, 1 \leq |z| \leq R\}$, there exists a neighborhood $\mathscr{O}$ of $\underline{z}$, there exists a constant $c > 0$, and there exists a $\mathscr{C}^\infty$ function S on $\mathscr{O}$ with values in $\mathscr{H}_{N(p+r)}$ such that (103) and (104) hold for all $z \in \mathscr{O} \cap \overline{\mathscr{U}}$. (Actually, the reader may observe that (104) holds not only for $z \in \mathscr{O} \cap \overline{\mathscr{U}}$ but for all $z \in \mathscr{O}$, but this will not play any role in what follows.)

We now make the construction of the Kreiss symmetrizer "global" by a compactness argument. The annulus $\{z \in \mathbb{C}\,,\, 1 \leq |z| \leq R\}$ is covered by a finite number $\mathscr{O}_1, \ldots, \mathscr{O}_J$ of such neighborhoods. We consider a partition of unity $\chi_1, \ldots, \chi_J$ that is subordinated to this covering. In other words, χ_j is a nonnegative $\mathscr{C}^\infty$ function with support in $\mathscr{O}_j$ for every j, and there holds

$$\forall\, z \in \overline{\mathscr{U}}\,, \quad |z| \leq R\,, \quad \sum_{j=1}^{J} \chi_j(z) = 1\,.$$

We define

$$\forall\, z \in \overline{\mathscr{U}}\,, \quad |z| \leq R\,, \quad S(z) := \sum_{j=1}^{J} \chi_j(z)\, S_j(z) \in \mathscr{H}_{N(p+r)}\,.$$

If c_j denotes the constant associated with the neighborhood $\mathscr{O}_j$ and if $c > 0$ denotes the minimum of the c_j's, then it is not so difficult to check the property

$$\forall\, z \in \overline{\mathscr{U}}\,, \quad |z| \leq R\,, \quad \mathbb{M}(z)^*\, S(z)\, \mathbb{M}(z) - S(z) \geq c\,(|z|-1)/|z|\, I\,,$$

(just multiply (103) on $\mathscr{O}_j$ by $\chi_j(z)$ and sum with respect to j), as well as

$$\forall\, z \in \overline{\mathscr{U}}\,, \quad |z| \leq R\,, \quad \forall\, \mathscr{W} \in \mathbb{C}^{N(p+r)}\,, \quad \mathscr{W}^*\, S(z)\, \mathscr{W} \geq c\, |\mathscr{W}|^2 - \frac{1}{c}\, |\mathbb{B}(z)\, \mathscr{W}|^2\,.$$

In other words, the pair $(\mathbb{M}, \mathbb{B})$ admits a Kreiss symmetrizer.

• We now show that the existence of a Kreiss symmetrizer is a sufficient condition for strong stability. Let $R \geq 2$, and let us consider a Kreiss symmetrizer S on the annulus $\{z \in \mathbb{C}\,,\, 1 \leq |z| \leq R\}$. We consider a point z in this annulus and a sequence $(\mathscr{W}_j) \in \ell^2$. The source terms $(\mathscr{F}_j)$, $\mathscr{G}$ are defined such that (59) holds. The a priori estimate of $(\mathscr{W}_j)$ follows from computations that are rather similar to what we have already done. More precisely, we multiply the induction relation in (59) by $(S(z)\, \mathscr{W}_{j+1})^*$ and use the fact that $S(z)$ is hermitian to obtain

$$\sum_{j=1}^{J} \mathrm{Re}\; \mathscr{W}_{j+1}^*\, S(z)\, \mathbb{M}(z)\, \mathscr{W}_j - \sum_{j=2}^{J+1} \mathscr{W}_j^*\, S(z)\, \mathscr{W}_j + \sum_{j=1}^{J} \mathrm{Re}\; \mathscr{W}_{j+1}^*\, S(z)\, \mathscr{F}_j = 0\,.$$

Using the induction relation again and substituting the expression of $\mathscr{W}_{j+1}$, we get

$$\mathscr{W}_1^*\, S(z)\, \mathscr{W}_1 - \mathscr{W}_{J+1}^*\, S(z)\, \mathscr{W}_{J+1} + \sum_{j=1}^{J} \mathscr{W}_j^* \left(\mathbb{M}(z)^*\, S(z)\, \mathbb{M}(z) - S(z)\right) \mathscr{W}_j$$
$$= -\mathrm{Re} \sum_{j=1}^{J} (\mathscr{W}_{j+1} + \mathbb{M}(z)\, \mathscr{W}_j)^*\, S(z)\, \mathscr{F}_j\,.$$

We let J tend to $+\infty$ and use the properties of the Kreiss symmetrizer, which yields

$$c_R\, \frac{|z|-1}{|z|} \sum_{j \geq 1} |\mathscr{W}_j|^2 + c_R\, |\mathscr{W}_1|^2 - \frac{1}{c_R}\, |\mathscr{G}|^2 \leq -\mathrm{Re} \sum_{j=1}^{J} (\mathscr{W}_{j+1} + \mathbb{M}(z)\, \mathscr{W}_j)^*\, S(z)\, \mathscr{F}_j\,.$$

Using some uniform bounds for $S(z)$ and $\mathbb{M}(z)$ on the annulus and the Cauchy-Schwarz inequality, we end up with

$$\frac{|z|-1}{|z|} \sum_{j \geq 1} |\mathscr{W}_j|^2 + |\mathscr{W}_1|^2 \leq C_R \left\{ \frac{|z|}{|z|-1} \sum_{j \geq 1} |\mathscr{F}_j|^2 + |\mathscr{G}|^2 \right\},$$

with a constant $C_R > 0$ that does not depend on $z \in \overline{\mathscr{U}}$, $|z| \leq R$.

It remains to show that the resolvent equation (59) admits a unique solution in ℓ^2 for all source terms (up to now we have only proved an a priori estimate for the solution). This final part of the proof follows from applying Lemma 3.3 and Lemma 3.4 again. More precisely, Lemma 3.3 shows that the resolvent equation (37) is uniquely solvable for $|z|$ large enough. There is no difficulty to show that the equivalent formulation (59) is also uniquely solvable for $|z|$ large enough. Then we can

apply Lemma 3.4 on every annulus $\{z \in \mathbb{C}\,,\ 1+2^{-\nu} \leq |z| \leq 2^{\nu}\}$, $\nu \in \mathbb{N}$ large enough. Eventually, Proposition 3.1 shows that the scheme (32) is strongly stable. □

4.5. **Some examples: the Lax-Friedrichs and leap-frog schemes.** The aim of this paragraph is to show how the theory developed in the proof of Theorem 3.6 applies in the case of some elementary numerical schemes. We shall test various discretized boundary conditions and compute the associated Lopatinskii determinants. For simplicity, we restrict in this paragraph to the case of a single scalar transport equation

$$\partial_t u + a\,\partial_x u = F(t,x)\,, \quad (t,x) \in \mathbb{R}^+ \times \mathbb{R}^+\,, \quad u|_{t=0} = 0\,. \tag{105}$$

For $a < 0$, there is no boundary condition to prescribe on $\{x = 0\}$, while for $a > 0$ the transport equation (105) should be supplemented with a Dirichlet boundary condition on $\{x = 0\}$.

4.5.1. *The Lax-Friedrichs scheme.* The Lax-Friedrichs discretization of the transport equation is given by (20) (here $N = 1$ and $A = a$ is a real number). We have seen in Section 2 that this scheme is stable in the sense of Definition 2.1 if and only if $\lambda\,|a| \leq 1$, and the corresponding operator Q_{LF} is geometrically regular. From the general definition (54), we obtain

$$\mathbb{A}_{-1}(z) = -\frac{1+\lambda\,a}{2\,z}\,, \quad \mathbb{A}_0(z) = 1\,, \quad \mathbb{A}_1(z) = -\frac{1-\lambda\,a}{2\,z}\,.$$

Consequently, Assumption 3.1 holds if and only if $\lambda\,|a| < 1$, which we assume from now on. It is not so surprising that the limit case $\lambda\,|a| = 1$ is excluded by the theory because in that case the Lax-Friedrichs scheme "degenerates" and becomes the upwind scheme which does not involve the same number of grid points (either p or r is zero while $p = r = 1$ when $\lambda\,|a| < 1$).

The matrix $\mathbb{M}(z)$ in (57) reads[16]

$$\mathbb{M}(z) = \begin{pmatrix} \dfrac{2\,z}{1-\lambda\,a} & -\dfrac{1+\lambda\,a}{1-\lambda\,a} \\ 1 & 0 \end{pmatrix}\,,$$

and we are going to check in an easy and direct way that $\mathbb{M}$ satisfies the discrete block structure condition. The eigenvalues of $\mathbb{M}(z)$ are the roots to the polynomial equation

$$\kappa^2 - \frac{2\,z}{1-\lambda\,a}\,\kappa + \frac{1+\lambda\,a}{1-\lambda\,a} = 0\,.$$

In particular, the matrix $\mathbb{M}(2)$ has two real eigenvalues: one belongs to the interval $]0,1[$ and the other one belongs to $]1,+\infty[$. Moreover, $\mathbb{M}(z)$ has an eigenvalue on $\mathbb{S}^1$ if and only if z belongs to the curve $\{\cos\eta - i\,\lambda\,a\,\sin\eta\,,\ \eta \in \mathbb{R}\}$. Since $\lambda\,|a| < 1$, the latter curve is included in the closed unit disk and its contact points with $\mathbb{S}^1$ are ± 1. Applying a continuity/connectedness argument, we are led to the following conclusion: for all $z \in \overline{\mathscr{U}} \setminus \{\pm 1\}$, the matrix $\mathbb{M}(z)$ has a unique eigenvalue $\kappa_s(z)$ in $\mathbb{D}$ and a unique eigenvalue in $\mathscr{U}$. The eigenvalue κ_s depends holomorphically on z near any point of $\overline{\mathscr{U}} \setminus \{\pm 1\}$, and $\mathbb{M}$ is holomorphically diagonalizable near any point of $\overline{\mathscr{U}} \setminus \{\pm 1\}$.

For $z \in \overline{\mathscr{U}} \setminus \{\pm 1\}$, the stable subspace $\mathbb{E}^s(z)$ of $\mathbb{M}(z)$ has dimension 1 - this is compatible with Lemma 3.7 because $N = r = 1$ in this example - and is given by

$$\forall\, z \in \overline{\mathscr{U}} \setminus \{\pm 1\}\,, \quad \mathbb{E}^s(z) = \mathrm{Span}\,\begin{pmatrix} \kappa_s(z) \\ 1 \end{pmatrix}\,.$$

In particular, the continuous extension of $\mathbb{E}^s$ to $\mathbb{S}^1$ proved in Theorem 3.5 is trivial here (it is even a holomorphic extension !), except possibly at the points ± 1 which we examine right now. From the expression of $\mathbb{E}^s$, we see that $\mathbb{E}^s(z)$ will have a limit at ± 1 if we can prove that the eigenvalue κ_s has a limit at ± 1.

The eigenvalues of $\mathbb{M}(1)$ are 1 and $(1+\lambda\,a)/(1-\lambda\,a)$. In the case $a < 0$, there holds $(1+\lambda\,a)/(1-\lambda\,a) \in]0,1[$, so this is another trivial case of continuous extension of the stable eigenvalue and we have $\kappa_s(1) = (1+\lambda\,a)/(1-\lambda\,a)$. In the case $a > 0$, there holds $(1+\lambda\,a)/(1-\lambda\,a) \in \mathscr{U}$ so the only possible extension of κ_s at the point 1 is 1. For z close to 1, $\mathbb{M}(z)$ has a unique eigenvalue close to 1

[16] Observe that in this special case, $\mathbb{M}$ is a holomorphic function on $\mathbb{C}$ and not only on a neighborhood of $\overline{\mathscr{U}}$.

that depends holomorphically on z. If we consider the points $z_\varepsilon := 1+\varepsilon \in \mathscr{U}$, $\varepsilon > 0$ small enough, the expansion of the eigenvalue of $\mathbb{M}(z_\varepsilon)$ close to 1 reads

$$1 - \frac{1}{\lambda\, a}\,\varepsilon + o(\varepsilon)\,,$$

so this eigenvalue belongs to $\mathbb{D}$ for $\varepsilon > 0$ small enough. By uniqueness of the stable eigenvalue, we can conlude that κ_s extends holomorphically to a whole neighborhood of 1 and $\kappa_s(1) = 1 \in \mathbb{S}^1$ when $a > 0$. The situation at $z = -1$ is examined in exactly the same way and we obtain the following conclusion: κ_s admits a holomorphic extension to a whole neighborhood of -1, and $\kappa_s(-1) = -(1+\lambda\, a)/(1-\lambda\, a) \in \mathbb{D}$ if $a < 0$, $\kappa_s(-1) = -1$ if $a > 0$.

The discrete block structure condition is very easy to verify because of the spectral splitting satisfied by the matrix $\mathbb{M}$: $\mathbb{M}$ has two distinct eigenvalues at every point of $\overline{\mathscr{U}}$ and is therefore diagonalizable (with a holomorphic change of basis) in the neighborhood of any point of $\overline{\mathscr{U}}$. The reduction near any point of $\overline{\mathscr{U}} \setminus \{\pm 1\}$ involves one (scalar) block of the first type and one (scalar) block of the second type. If $a < 0$, the reduction near ± 1 involves one (scalar) block of the second type and one (scalar) block of the third type. If $a > 0$, the reduction near ± 1 involves one (scalar) block of the first type and one (scalar) block of the third type.

Let us now verify whether the UKLC is satisfied for various types of discretized boundary conditions. We begin with the Dirichlet boundary condition. In this case, the numerical scheme reads

$$\begin{cases} U_j^{n+1} = \dfrac{U_{j-1}^n + U_{j+1}^n}{2} - \dfrac{\lambda\, a}{2}\,(U_{j+1}^n - U_{j-1}^n) + \Delta t\, F_j^n\,, & j \geq 1\,, \quad n \geq 0\,, \\ U_0^{n+1} = g^{n+1}\,, & n \geq 0\,, \\ U_j^0 = 0\,, & j \geq 0\,. \end{cases}$$

In this case, one has $q = 0$, $B_{0,0} = B_{0,-1} = 0$ and the matrix $\mathbb{B}(z)$, whose abstract definition is (58), reads

$$\forall\, z \in \mathbb{C} \setminus \{0\}\,, \quad \mathbb{B}(z) = \begin{pmatrix} 0 & 1 \end{pmatrix}\,.$$

It is easily checked that the UKLC is satisfied, whatever the sign of a. Indeed, the intersection of $\mathbb{E}^s(z)$ with $\operatorname{Ker} \mathbb{B}(z)$ is non-trivial provided that the Lopatinskii determinant

$$\Delta(z) := \mathbb{B}(z) \begin{pmatrix} \kappa_s(z) \\ 1 \end{pmatrix}\,,$$

vanishes. Here this determinant equals 1 for all $z \in \overline{\mathscr{U}}$ so the UKLC is satisfied. From a practical point of view, it is interesting to test the Dirichlet boundary condition for an outgoing transport equation. Let us therefore consider the transport equation (105) with $a < 0$ and $F = 0$. In that case, the solution to (105) is 0. To approximate this solution, we use the numerical scheme

$$\begin{cases} U_j^{n+1} = \dfrac{U_{j-1}^n + U_{j+1}^n}{2} - \dfrac{\lambda\, a}{2}\,(U_{j+1}^n - U_{j-1}^n)\,, & j \geq 1\,, \quad n \geq 0\,, \\ U_0^{n+1} = g^{n+1}\,, & n \geq 0\,, \\ U_j^0 = 0\,, & j \geq 0\,, \end{cases}$$

with a nonzero source terme (g^n) on the boundary. The numerical computations are run with $a = -1$, $\lambda = 0.9$, and $g^n = 1$ for all $n \geq 1$. The result of the computation is shown in Figure 5 at two different time steps. The space interval is $[0, 1]$ and the number of grid points is 100. By finite speed of propagation, we know that both the exact solution and the numerical solution vanish at the right end of the computation interval, so we impose a homogeneous Dirichlet condition at 1. This is relevant provided that the computations are run up to a certain number of time steps. The observed numerical solution is very small, which is justified by Theorem 3.5 and our verification of the UKLC.

We go on with the Neumann boundary condition. The corresponding numerical scheme reads

$$\begin{cases} U_j^{n+1} = \dfrac{U_{j-1}^n + U_{j+1}^n}{2} - \dfrac{\lambda\, a}{2}\,(U_{j+1}^n - U_{j-1}^n) + \Delta t\, F_j^n\,, & j \geq 1\,, \quad n \geq 0\,, \\ U_0^{n+1} = U_1^{n+1} + g^{n+1}\,, & n \geq 0\,, \\ U_j^0 = 0\,, & j \geq 0\,. \end{cases}$$

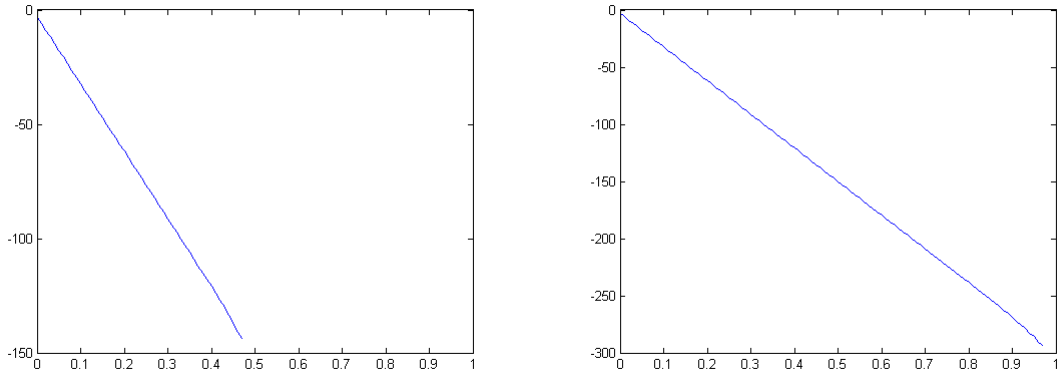

FIGURE 5. The Lax-Friedrichs scheme for an outgoing transport equation with a non-homogeneous boundary condition $g^n = 1$ at various time steps. The numerical scheme should approximate the solution zero. The solution is represented on a log scale, and the space interval is $[0, 1]$.

For this scheme, we still have $q = 0$, and in the notation of (32), $B_{0,0} = 0$, $B_{0,-1} = \mathbf{T}^0$. The corresponding matrix $\mathbb{B}(z)$ reads

$$\mathbb{B}(z) = \begin{pmatrix} -1 & 1 \end{pmatrix},$$

so the Lopatinskii determinant reads

$$\Delta(z) = 1 - \kappa_s(z).$$

If $a < 0$, we have seen that $\kappa_s(z)$ belongs to $\mathbb{D}$ for all $z \in \overline{\mathscr{U}}$. In particular, $\kappa_s(z) \neq 1$ and the UKLC holds. When one wishes to discretize the outgoing transport equation (105), for which no boundary condition is required, one can therefore use the stronlgy stable (and consistent !) scheme

$$\begin{cases} U_j^{n+1} = \dfrac{U_{j-1}^n + U_{j+1}^n}{2} - \dfrac{\lambda\, a}{2}\,(U_{j+1}^n - U_{j-1}^n) + \Delta t\, F(n\,\Delta t, j\,\Delta x)\,, & j \geq 0\,, \quad n \geq 0\,, \\ U_0^{n+1} = U_1^{n+1}\,, & n \geq 0\,, \\ U_j^0 = 0\,, & j \geq 0\,. \end{cases} \tag{106}$$

To observe the strong stability of the latter numerical scheme, one can use the same test as the one reported in Figure 5 for the Dirichlet boundary condition (that is, no source term in the interior and a constant source term equal to 1 on the boundary). The results are entirely similar with either the Dirichlet or the Neumann boundary condition.

If $a > 0$, we know that $\kappa_s(z)$ belongs to $\mathbb{D}$ for all $z \in \overline{\mathscr{U}} \setminus \{\pm 1\}$ so Δ does not vanish on this set. Since $\kappa_s(\pm 1) = \pm 1$, we also find that Δ vanishes at 1 and does not vanish at -1. In the incoming case $a > 0$, the Neumann boundary condition does not satisfy the UKLC and the corresponding numerical scheme is not strongly stable. What can we observe and conclude in such a situation ? We report on a very simple numerical test which shows that the violation of strong stability is a serious obstacle for convergence of the numerical solution. We consider the incoming transport equation (105) with $a = 1$ and $F = 0$. We impose the homogeneous boundary condition $u(t, 0) = 0$ so the exact solution to the transport equation is 0. Since $u(t, 0) = 0$, we have $\partial_t u|_{x=0} = 0$ and (105) gives $\partial_x u|_{x=0} = 0$. This may suggest to use a homogeneous Neumann condition at the boundary instead of the Dirichlet boundary condition. We thus consider the numerical scheme (106). When the source term (F_j^n) vanishes, the numerical solution is 0 and it reproduces the exact solution. We perturb this situation by choosing $F_1^0 = 1/\Delta t$ and all other F_j^n vanish[17] The solution is represented in Figure

[17]This perturbation is a classical test for stability. First, it is easy to use since it localized on a single mesh of the grid, and even though its L^∞-norm is large, the $L^2(t, x)$-norm of this perturbation is of order 1, independently of Δt. The second reason why it is useful is that because of space localization, its Fourier transform triggers more or less all frequencies so if one frequency is amplified by the scheme, there is a reasonable chance to observe this phenomenon with this perturbation.

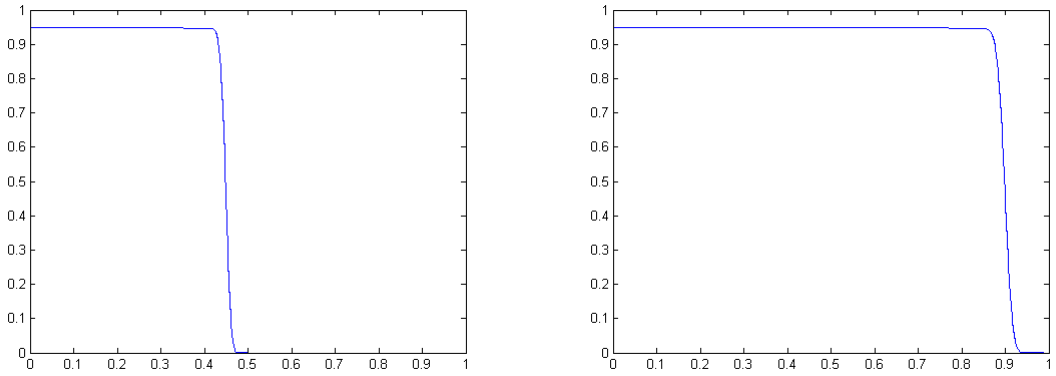

FIGURE 6. The Lax-Friedrichs scheme for an incoming transport equation with a homogeneous Neumann boundary condition at various time steps. The interior source term vanishes except F_1^0 which we choose equal to $1/\Delta t$.

6 at two various iterations, where we have chosen $\lambda = 0.9$ again. The number of grid points is 1000 and the space interval is $[0, 1]$. The numerical solution is some kind of traveling wave propagating to the right and connecting a state $\overline{U} > 0$ to 0. In particular, the exact boundary condition is not approximated at all by the numerical scheme (even though the homogeneous discrete Neumann condition was imposed!). Though we have not pushed any rigorous investigation further than a few numerical tests, we believe that this specific choice of source term may be a good candidate for showing rigorously that the energy estimate (34) is not satisfied.

4.5.2. *The leap-frog scheme.* We consider the leap-frog approximation (22) for the transport equation (105). We still restrict to the scalar case $N = 1$, $A = a \in \mathbb{R}$. For this scheme, there holds $p = r = 1$, and the definition (54) reads

$$\mathbb{A}_{-1}(z) = -\frac{\lambda\, a}{z}\,, \quad \mathbb{A}_0(z) = 1 - \frac{1}{z^2}\,, \quad \mathbb{A}_1(z) = \frac{\lambda\, a}{z}\,.$$

Assumption 3.1 is thus satisfied as long as $a \neq 0$. (When $a = 0$, the scheme degenerates and involves only one point.) We have seen in Section 2 that both stability in the sense of Definition 2.2 and geometric regularity hold as long as $\lambda\,|a| < 1$. We thus assume $\lambda\,|a| < 1$ and $a \neq 0$ from now on.

The matrix $\mathbb{M}(z)$ in (57) reads

$$\mathbb{M}(z) = \begin{pmatrix} \dfrac{1 - z^2}{\lambda\, a\, z} & 1 \\ 1 & 0 \end{pmatrix}\,,$$

so the eigenvalues of $\mathbb{M}(z)$ are the roots to the polynomial equation

$$\kappa^2 + \frac{z^2 - 1}{\lambda\, a\, z}\,\kappa - 1 = 0\,.$$

The matrix $\mathbb{M}(2)$ has two real eigenvalues: one of them belongs to the interval $]-\infty, -1[$ and the second one belongs to $]0, 1[$. Moreover, $\mathbb{M}(z)$ has no eigenvalue on $\mathbb{S}^1$ when z belongs to $\mathscr{U}$ so we can conclude, as in Lemma 3.7, that $\mathbb{M}(z)$ has a unique eigenvalue $\kappa_s(z) \in \mathbb{D}$ and a unique eigenvalue in $\mathscr{U}$ for all $z \in \mathscr{U}$. Of course, κ_s depends holomorphically on $z \in \mathscr{U}$. The stable subspace $\mathbb{E}^s(z)$ has dimension 1 and is given by

$$\forall\, z \in \mathscr{U}\,, \quad \mathbb{E}^s(z) = \mathrm{Span}\, \begin{pmatrix} \kappa_s(z) \\ 1 \end{pmatrix}\,.$$

This is exactly the same expression as for the Lax-Friedrichs scheme, which is not surprising because $\mathbb{M}$ is still a companion matrix[18]. Our goal is now to study the continuous extension of the stable

[18]The reader will find in the following paragraph an extension of this remark where the structure of $\mathbb{M}$ will be fully used. This will help us proving the so-called Goldberg-Tadmor Lemma.

eigenvalue κ_s to the boundary $\mathbb{S}^1$ of $\mathscr{U}$ and to verify that $\mathbb{M}$ satisfies the discrete block structure condition. The situation is slightly more complicated but in some sense much more interesting than for the Lax-Friedrichs scheme.

Computing the discriminant of the characteristic polynomial of $\mathbb{M}(z)$, we first observe that $\mathbb{M}$ has a double eigenvalue if and only if z is one of the points $\pm(\sqrt{1-\lambda^2\,a^2}+i\,\lambda\,a)$ or their conjugates. These four points are located on $\mathbb{S}^1$, and unsurprisingly they correspond exactly to the singular points of the eigenvalues curves for the leap-frog scheme (see the right picture in Figure 1). We can already conclude that $\mathbb{M}$ can be holomorphically diagonalized in the neighborhood of any points $\underline{z}\in\mathbb{S}^1$ which is not one of these four points and that the stable eigenvalue κ_s admits a holomorphic extension to the neighborhood of any such "non-exceptional" point. The continuous - and even holomorphic extension - of the stable subspace is clear in this case. Let us now focus on the points where $\mathbb{M}$ has a double eigenvalue. We consider for instance the point $\underline{z}:=\sqrt{1-\lambda^2\,a^2}+i\,\lambda\,a$ (the three other cases are entirely similar). There holds

$$\mathbb{M}(\underline{z})=\begin{pmatrix}-2\,i & 1\\ 1 & 0\end{pmatrix},$$

so $\mathbb{M}(\underline{z})$ is similar to a Jordan block with the eigenvalue $-i$. More precisely, if we introduce the invertible matrix

$$\underline{T}:=\begin{pmatrix}1 & 1\\ i & 0\end{pmatrix},$$

we have

$$\underline{T}^{-1}\,\mathbb{M}(\underline{z})\,\underline{T}=-i\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}.$$

In view of Definition 4.1, the constant matrix $\underline{T}$ is a good candidate for reducing $\mathbb{M}$ to the discrete block structure condition. Let us check this property in full details. We compute

$$\underline{T}^{-1}\,\mathbb{M}(z)\,\underline{T}=\begin{pmatrix}-i & -i\\ \dfrac{1-z^2}{\lambda\,a\,z}+2\,i & \dfrac{1-z^2}{\lambda\,a\,z}+i\end{pmatrix}. \tag{107}$$

In order to check that the discrete block structure condition holds, we only need to compute the derivative at $z=\underline{z}$ of the lower left coefficient of the latter matrix. We obtain

$$\frac{\partial}{\partial z}\left(\frac{1-z^2}{\lambda\,a\,z}+2\,i\right)\Big|_{z=\underline{z}}=-\frac{2\sqrt{1-\lambda^2\,a^2}}{\lambda\,a\,\underline{z}}.$$

Let now $\theta\in\mathbb{C}$ with $\mathrm{Re}\,\theta>0$. We consider the roots ζ to the equation

$$\zeta^2=\overline{(-i)}\,\frac{-2\sqrt{1-\lambda^2\,a^2}}{\lambda\,a\,\underline{z}}\,\underline{z}\,\theta=-\frac{2\,i\sqrt{1-\lambda^2\,a^2}}{\lambda\,a}\,\theta.$$

The roots ζ cannot be purely imaginary, for otherwise $i\,\theta$ would be a real number. According to Definition 4.1, the derivative of the lower left coefficient in (107) satisfies the property required in the definition of the discrete block structure condition. This reduction involves a single 2×2 block of the fourth type. We have even shown that the change of basis can be chosen to be independent of z in the neighborhood of $\underline{z}$. The continuous extension of κ_s, and therefore of $\mathbb{E}^s$, to $\underline{z}$ follows from the continuity of the roots of the characteristic polynomial of $\mathbb{M}(z)$.

Let us now check whether the UKLC is satisfied for various types of numerical boundary conditions. As before, we first consider the Dirichlet boundary conditions. In other words, we consider the numerical scheme

$$\begin{cases}U_j^{n+1}=U_j^{n-1}-\lambda\,a\,(U_{j+1}^n-U_{j-1}^n)+\Delta t\,F_j^n, & j\geq 1, \quad n\geq 1,\\ U_0^{n+1}=g^{n+1}, & n\geq 1,\\ U_j^1=U_j^0=0, & j\geq 0.\end{cases}$$

The matrix $\mathbb{B}(z)$ defined in (58) reads

$$\mathbb{B}(z)=\begin{pmatrix}0 & 1\end{pmatrix},$$

and the associated Lopatinskii determinant equals 1 for all $z \in \overline{\mathscr{U}}$. This shows, as for the Lax-Friedrichs scheme, that the Dirichlet boundary condition satisfies the UKLC for the leap-frog scheme. We emphasize that this result is independent of the sign of a. Numerical tests as the one reported in Figure 5 can be performed and give rather good results in the outgoing case (meaning that the numerical solution is rather close to the exact solution even in the case of non-homogeneous Dirichlet boundary conditions).

The reader can also check that the leap-frog scheme combined with the Neumann condition at the boundary always satisfies the Godunov-Ryabenkii condition, but always violates the UKLC. Again, this result is independent of the sign of a. If one performs the same kind of test as the one reported in Figure 6, the numerical solution has similar features, meaning that it looks like a traveling wave propagating to the right and connecting some state $\overline{U} > 0$ to 0.

We now study another type of discrete boundary condition which is obtained by using backward integration along the characteristics. More precisely, for $a < 0$, the transport equation (105) is outgoing. On the boundary mesh of index $j = 0$, we apply the so-called upwind scheme, which amounts to considering the scheme

$$\begin{cases} U_j^{n+1} = U_j^{n-1} - \lambda\, a\, (U_{j+1}^n - U_{j-1}^n) + \Delta t\, F_j^n\,, & j \geq 1\,, \quad n \geq 1\,, \\ U_0^{n+1} = U_0^n - \lambda\, a\, (U_1^n - U_0^n) + g^{n+1}\,, & n \geq 1\,, \\ U_j^1 = U_j^0 = 0\,, & j \geq 0\,. \end{cases} \tag{108}$$

This numerical procedure seems to be a somehow reasonable discretization for $a < 0$ since we use a stable approximation of the Cauchy problem in the interior domain and a rather precise approximation of the solution at the boundary. It seems much less reasonable in the case $a > 0$ for in that case, the upwind discretization "on the right" is known to be unstable for the Cauchy problem (one should use the discretization "on the left"). We are going to examine the strong stability of (108) according to the sign of a.

The careful reader may have observed that the discrete boundary condition in (108) involves not only U_1^n but also U_0^n, which does not exactly fall into the framework of (32). However, we could have equally considered boundary operators $B_{j,\sigma}$ in (32) of the form

$$B_{j,-1} = \sum_{\ell=0}^{q} B_{\ell,j,-1}\, \mathbf{T}^\ell\,, \quad j = 1-r, \dots, 0\,,$$

$$B_{j,\sigma} = \sum_{\ell=-r-j}^{q} B_{\ell,j,\sigma}\, \mathbf{T}^\ell\,, \quad j = 1-r, \dots, 0\,, \quad \sigma = 0, \dots, s\,.$$

For such boundary operators, the reader can verify that the values U_j^{n+1}, $j = 1-r, \dots, 0$, are obtained as linear combinations of some $U_j^{n-s}, \dots, U_j^n$, which are already known from the previous iteration steps, and of some U_j^{n+1}, $j \geq 1$, which are also known because they are obtained from the "interior" discretization. Hence the numerical scheme is explicit and well-defined. There is a slight difference in the definition of the matrix $\mathbb{B}(z)$ in (58), and we leave as an exercise to the reader to go through the derivation of the resolvent equation (59) in the case of (108). The associated matrix $\mathbb{B}(z)$ is

$$\mathbb{B}(z) = \begin{pmatrix} \dfrac{\lambda\, a}{z} & 1 - \dfrac{1 + \lambda\, a}{z} \end{pmatrix},$$

and with the above parametrization of the stable subspace, the Lopatinskii determinant reads

$$\Delta(z) = \frac{\lambda\, a}{z}\, \kappa_s(z) + 1 - \frac{1 + \lambda\, a}{z}\,.$$

Our goal is therefore to determine whether there exists some $z \in \overline{\mathscr{U}}$ such that

$$z - 1 = \lambda\, a\, (1 - \kappa_s(z))\,, \tag{109}$$

knowing that $\kappa_s(z)$ satisfies the relation

$$\kappa_s(z)\, \big(z^2 - 1\big) = \lambda\, a\, z\, \big(1 - \kappa_s(z)^2\big)\,. \tag{110}$$

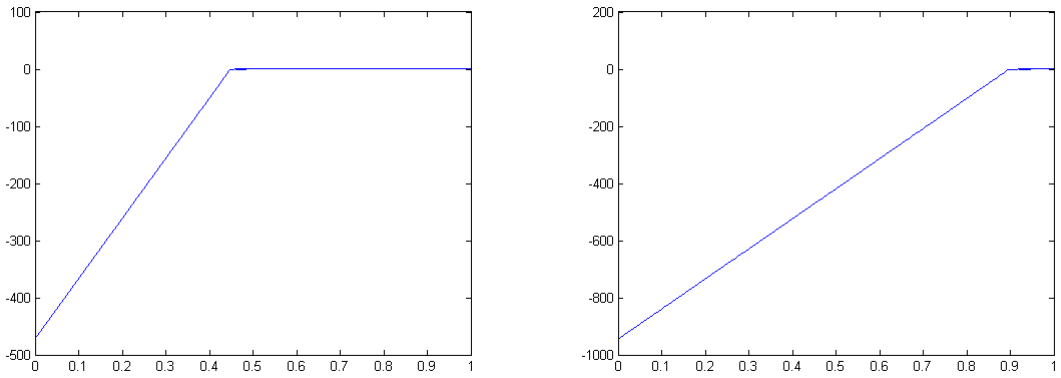

FIGURE 7. The leap-frog scheme (108) for an incoming equation ($a = 1$) with backward integration along the characteristic at the boundary. The interior source term vanishes except F_1^1 which we choose equal to $1/\Delta t$.

If $z \in \overline{\mathscr{U}} \setminus \{1\}$, the only possibility for $\Delta(z)$ to vanish is to have $\kappa_s(z) \neq 1$ and we can then divide both left and right hand side terms in (110) by the corresponding expression in (109). We then obtain $\kappa_s(z) = z$. In other words, for $z \neq 1$, the only possibility for $\Delta(z)$ to vanish is to have $\kappa_s(z) = z$ but then (109) gives $\lambda\, a = -1$. This is obviously in contradiction with our stability assumption for the discrete Cauchy problem. Hence Δ can only vanish at the point 1. In particular, the Godunov-Ryabenkii condition holds for (108) whatever the sign of a. Moreover, the above expression of Δ yields $\Delta(1) = \lambda\, a\, (\kappa_s(1) - 1)$ so Δ vanishes at 1 if and only if $\kappa_s(1)$ equals 1. The eigenvalues of $\mathbb{M}(1)$ are ± 1 so it is not clear at first sight whether $\kappa_s(1)$ equals 1 or -1. Considering the sequence of points $z_\varepsilon := 1 + \varepsilon$ with $\varepsilon > 0$ going to 0, we can compute the asymptotic expansions of both eigenvalues of $\mathbb{M}(z_\varepsilon)$. We then obtain $\kappa_s(1) = 1$ if $a > 0$ and $\kappa_s(1) = -1$ if $a < 0$. Consequently, we find $\Delta(1) = 0$ if $a > 0$ and $\Delta(1) \neq 0$ if $a < 0$. The numerical scheme (108) satisfies the UKLC and is strongly stable if $a < 0$, while it is not strongly stable if $a > 0$. We can go a little further. In the previous paragraph, when we have shown that the Lax-Friedrichs scheme with the Neumann condition on the boundary is not strongly stable, we have shown that 1 is a root of the Lopatinskii determinant. In that case, the reader can check that Δ extends holomorphically to a neighborhood of 1 and that 1 is a simple root of Δ. The situation is a little more singular for (108) when $a > 0$: the Lopatinskii determinant Δ also extends holomorphically to a neighborhood of 1, but here 1 is at least a double root of Δ. Indeed, we can differentiate Δ with respect to z and obtain

$$\Delta'(1) = \lambda\, a\, \kappa_s'(1) + 1 + \lambda\, a\, (\kappa_s(1) - 1) = \lambda\, a\, \kappa_s'(1) + 1\,.$$

In the meantime, we can differentiate (110) with respect to z, use $\kappa_s(1) = 1$ (here we use $a > 0$), and get $\kappa_s'(1) = -1/(\lambda\, a)$. In other words, $\Delta'(1)$ vanishes and 1 is at least a double root of Δ.

We report on the numerical simulation of (108) in the unstable case $a = 1$. The space interval is $[0, 1]$, we choose 1000 grid points, $\lambda = 0.9$, the source term g^n on the boundary equals zero for all $n \geq 2$, while $F_j^n = 0$ for all j, n except $F_1^1 = 1/\Delta t$. The numerical solution is represented at two different time steps in Figure 7. The instability is of a different kind than the one reported in the case of the Lax-Friedrichs scheme with Neumann boundary condition, but it is not as violent as an exponential growth. Anyway, the exact boundary condition is not approximated at all since U_0^n seems to grow linearly in n. The exact same numerical test can be performed in the strongly stable case $a = -1$ (we do not change any other parameter). The results are shown on Figure 8 on a log scale: the numerical solution is small, as predicted by the strong stability estimate.

4.6. **Goldberg-Tadmor's Lemma for Dirichlet boundary conditions.** The aim of this paragraph is to understand why in all above examples the Dirichlet boundary conditions lead to strongly stable numerical schemes. This result if first due to Goldberg and Tadmor [7] and we show that it holds in our more general framework. The result is the following.

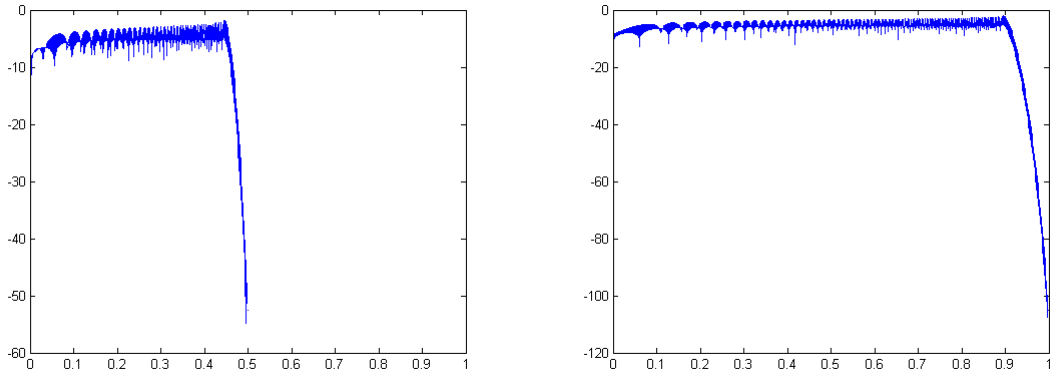

FIGURE 8. The leap-frog scheme (108) for an outgoing equation ($a = -1$) with backward integration along the characteristic at the boundary. The interior source term vanishes except F_1^1 which we choose equal to $1/\Delta t$. The numerical solution is represented on a log scale.

Proposition 4.2 (Goldberg-Tadmor)**.** *Let us consider the scalar case $N = 1$, with a numerical scheme (14) that is stable for the discrete Cauchy problem. Then the numerical scheme*

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} Q_\sigma \, U_j^{n-\sigma} + \Delta t \, F_j^n \,, & j \geq 1 \,, \quad n \geq s \,, \\ U_j^{n+1} = g_j^{n+1} \,, & j = 1-r, \dots, 0 \,, \quad n \geq s \,, \\ U_j^n = 0 \,, & j \geq 1-r \,, \quad n = 0, \dots, s \,, \end{cases} \tag{111}$$

is strongly stable in the sense of Definition 3.1.

Proposition 4.2 shows that in the scalar case, there exists at least one way to impose numerical boundary conditions and to obtain strong stability. The reader may observe that this is far from clear when one considers the characterization in Theorems 3.5 and 3.6. What may look surprising at first glance is that, in general, the Dirichlet boundary condition is not consistent in the L^∞-norm (just think of an outgoing transport equation with a bump propagating towards the left, which does not satisfy the homogeneous Dirichlet boundary condition at all !). From a numerical point of view, the Dirichlet boundary condition may give rise to boundary layers, and one way to reformulate Proposition 4.2 is to say that in the scalar case, such numerical boundary layers are stable. We emphasize that Proposition 4.2 is independent of the underlying transport equation that is approximated by the operators Q_σ, meaning that these operators may be obtained by discretizing either an incoming or an outgoing transport equation.

Before proving Proposition 4.2, we state and prove two preliminary results that will be useful later on.

Lemma 4.8. *Let $M \in \mathscr{M}_m(\mathbb{C})$ and let λ be an eigenvalue of M with algebraic multiplicity p. If* Ker $(M - \lambda I)$ *has dimension* 1*, then for all $k = 1, \dots, p$,* Ker $(M - \lambda I)^k$ *has dimension k.*

Proof of Lemma 4.8. There is nothing to prove if p equals 1, so we assume $p \geq 2$. The result is proved by induction on k. Let us assume that the result holds up to the index k. If $k = p$ then the proof is complete, so we further assume $k \leq p-1$. We already know that Ker $(M-\lambda I)^{k+1}$ contains Ker $(M-\lambda I)^k$. The dimension of Ker $(M-\lambda I)^{k+1}$ can not be equal to k for otherwise, there would hold Ker $(M - \lambda I)^k =$ Ker $(M - \lambda I)^{k+1}$ and this implies Ker $(M - \lambda I)^k =$ Ker $(M - \lambda I)^{k+j}$ for all integer j. In particular, Ker $(M - \lambda I)^p$ would have dimension $k < p$ and this is impossible.

Let us now assume that the dimension Ker $(M - \lambda I)^{k+1}$ equals at least $k + 2$. In particular, there exist two linearly independent vectors X_1, X_2 in Ker $(M - \lambda I)^{k+1} \setminus$ Ker $(M - \lambda I)^k$. Since $(M - \lambda I)^k X_i$, $i = 1, 2$ belong to the one-dimensional space Ker $(M - \lambda I)$, there exists a non-trivial linear combination $\mu_1 X_1 + \mu_2 X_2$ that belongs to Ker $(M - \lambda I)^k$ but this is excluded by the

construction of X_1, X_2. We are led to a contradiction. The only remaining possibility is to have Ker $(M - \lambda I)^{k+1}$ of dimension $k+1$. □

As a matter of fact, Lemma 4.8 is a particular case of a more general fact. More precisely, it is known that for any eigenvalue λ of a matrix M, the sequence $(\dim \text{Ker } (M - \lambda I)^k)_{k \geq 1}$ is concave. The proof of this fact uses similar arguments to those developed in the proof of Lemma 4.8. The following Lemma is a generalization of Lemma 2.8.

Lemma 4.9. *Let $M \in \mathscr{M}_m(\mathbb{C})$ be a companion matrix, that is*

$$M = \begin{pmatrix} \mu_1 & \dots & \dots & \mu_m \\ 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Let λ be a nonzero eigenvalue of M with algebraic multiplicity p. Then for all $k = 1, \dots, p$, there holds

$$\text{Ker } (M - \lambda I)^k = \left\{ (P(m-1)\,\lambda^{m-1}, \dots, P(1)\,\lambda, P(0))^T \,,\, P \in \mathbb{C}_{k-1}[X] \right\}.$$

We warn the reader that Lemma 4.9 is not true in general for block companion matrices.

Proof of Lemma 4.9. The proof is performed by induction on k. The result is clear for $k = 1$ (see Lemma 2.8), and we assume that it holds up to the order $k < p$ (otherwise the proof is already complete). Combining Lemma 2.8 and Lemma 4.8, we already know that Ker $(M - \lambda I)^{k+1}$ has dimension $k+1$. Since λ is nonzero, the linear map

$$P \in \mathbb{C}_k[X] \longmapsto (P(m-1)\,\lambda^{m-1}, \dots, P(1)\,\lambda, P(0))^T \in \mathbb{C}^m$$

is an injection (here we use $k+1 \leq p \leq m$). It therefore only remains to prove that the image of this linear map is included in Ker $(M - \lambda I)^{k+1}$. Since we already know that the image of any polynomial of degree $\leq k-1$ belongs to Ker $(M - \lambda I)^k$, we only need to find one polynomial of degree equal to k and whose image by the latter linear map belongs to Ker $(M - \lambda I)^{k+1}$. We define

$$Q(X) := \prod_{j=0}^{k-1} (X - j)\,,$$

whose degree equals k, and we define $Y := (Q(m-1)\,\lambda^{m-1}, \dots, Q(1)\,\lambda, Q(0))^T$. Using the definition of the companion matrix M, we compute

$$(M - \lambda I)\,Y = \begin{pmatrix} y \\ (Q(m-1) - Q(m-2))\,\lambda^{m-1} \\ \vdots \\ (Q(1) - Q(0))\,\lambda \end{pmatrix}, \quad y := \sum_{\ell=1}^{m} \mu_\ell\, Q(m-\ell)\,\lambda^{m-\ell} - Q(m-1)\,\lambda^m\,.$$

Let us define the polynomial $R(X) := Q(X+1) - Q(X)$, which has degree $k-1$. If we can show that the above complex number y equals $R(m-1)\,\lambda^m$, then we shall have $(M - \lambda I)\,Y \in \text{Ker } (M - \lambda I)^k$ by the induction assumption and the proof will be complete. Let us therefore show $y = R(m-1)\,\lambda^m$. We know that λ is a root of multiplicity $p \geq k+1$ to the characteristic polynomial of M, hence

$$\frac{\mathrm{d}^k}{\mathrm{d}X^k}\Big(X^m - \sum_{\ell=1}^{m} \mu_\ell\, X^{m-\ell} \Big)\Big|_{X=\lambda} = 0\,.$$

Since λ is nonzero, we have

$$\forall\, j \in \mathbb{N}\,, \quad \frac{\mathrm{d}^k}{\mathrm{d}X^k}(X^j)\big|_{X=\lambda} = Q(j)\,\lambda^{j-k}\,,$$

so we get

$$Q(m)\,\lambda^{m-k} - \sum_{\ell=1}^{m} \mu_\ell\, Q(m-\ell)\,\lambda^{m-\ell-k} = 0\,.$$

Combining with the above definition of y, we end up with $y = (Q(m)-Q(m-1))\,\lambda^m = R(m-1)\,\lambda^m$ which is the relation we were aiming at. The proof of Lemma 4.9 is now complete. □

We now turn to the proof of Proposition 4.2.

Proof of Proposition 4.2. We first recall the result of Lemma 2.7 which shows that, under the assumptions of Proposition 4.2, the operators Q_σ are geometrically regular. Theorem 4.1 shows that the matrix $\mathbb{M}$ associated with (111) satisfies the discrete block structure condition, and Theorem 4.2 then shows that $\mathbb{M}$ admits a K-symmetrizer with a vector space $\underline{\mathbb{E}}^s$ of dimension r. Eventually, Theorem 4.3 shows that the stable bundle $\mathbb{E}^s$ of $\mathbb{M}$ extends continuously from $\mathscr{U}$ to $\overline{\mathscr{U}}$. The proof of Proposition 4.2 then splits into two steps.

• Let $\underline{z} \in \overline{\mathscr{U}}$, and let $\underline{\kappa}_1, \dots, \underline{\kappa}_K$ denote the eigenvalues of $\mathbb{M}(\underline{z})$ with corresponding algebraic multiplicities $\alpha_1, \dots, \alpha_K$. For $z \in \mathscr{U}$ close to $\underline{z}$, we know from Lemma 3.7 that the number of stable eigenvalues of $\mathbb{M}(z)$ close to $\underline{\kappa}_k$ is independent of z. We let μ_k denote this number, which can be computed for instance by counting the stable eigenvalues of $\mathbb{M}((1+\varepsilon)\,\underline{z})$, $0 < \varepsilon \ll 1$. Our first goal is to show that $\mathbb{E}^s(\underline{z})$ can be decomposed as

$$\mathbb{E}^s(\underline{z}) = \oplus_{k=1}^K \mathrm{Ker}\ (\mathbb{M}(\underline{z}) - \underline{\kappa}_k\, I)^{\mu_k}\,. \tag{112}$$

Let us first observe that (112) is trivial when $\underline{z} \in \mathscr{U}$ because in that case, the eigenvalues $\underline{\kappa}_k$ either belong to $\mathbb{D}$ (the stable ones) or to $\mathscr{U}$ (the unstable ones). We therefore have $\mu_k = \alpha_k$ if $\underline{\kappa}_k \in \mathbb{D}$, and $\mu_k = 0$ if $\underline{\kappa}_k \in \mathscr{U}$, which clearly implies (112). We thus turn to the more delicate case $\underline{z} \in \mathbb{S}^1$. There is no loss of generality in assuming that the eigenvalues are ordered in such a way that $\underline{\kappa}_1, \dots, \underline{\kappa}_{K_1}$ belong to $\mathbb{D}$ (stable eigenvalues), $\underline{\kappa}_{K_1+1}, \dots, \underline{\kappa}_{K_2}$ belong to $\mathscr{U}$ (unstable eigenvalues), and $\underline{\kappa}_{K_2+1}, \dots, \underline{\kappa}_K$ belong to $\mathbb{S}^1$ (neutral eigenvalues). Of course, we set $\mu_k = \alpha_k$ for $1 \le k \le K_1$, and $\mu_k = 0$ for $K_1 + 1 \le k \le K_2$. Let $\varepsilon > 0$ be so small that the disks centered at $\underline{\kappa}_1, \dots, \underline{\kappa}_K$ and of radius ε are pairwise disjoint. For $n \in \mathbb{N}$ sufficiently large, the matrix $\mathbb{M}((1+2^{-n})\,\underline{z})$ has exactly μ_k stable eigenvalues in the disk centered at $\underline{\kappa}_k$ and of radius ε. We let $\kappa_{k,1}^{(n)}, \dots, \kappa_{k,\mu_k}^{(n)}$ denote these eigenvalues. The eigenvalues $\kappa_{k,j}^{(n)}$ tend to $\underline{\kappa}_k$ as n tends to infinity, so we have

$$\lim_{n\to+\infty} \prod_{j=1}^{\mu_k} \left(\mathbb{M}((1+2^{-n})\,\underline{z}) - \kappa_{k,j}^{(n)}\, I\right) = \left(\mathbb{M}(\underline{z}) - \underline{\kappa}_k\, I\right)^{\mu_k}.$$

Let now $\underline{X} \in \mathbb{E}^s(\underline{z})$, and let $X_n \in \mathbb{E}^s((1+2^{-n})\,\underline{z})$ denote a sequence that converges towards $\underline{X}$. Such a sequence exists since we already know that the whole vector space $\mathbb{E}^s((1+2^{-n})\,\underline{z})$ converges towards $\mathbb{E}^s(\underline{z})$. Using (112) for every n, we have

$$X_n \in \oplus_{k=1}^K \mathrm{Ker} \prod_{j=1}^{\mu_k} \left(\mathbb{M}((1+2^{-n})\,\underline{z}) - \kappa_{k,j}^{(n)}\, I\right) = \mathrm{Ker} \prod_{k=1}^{K} \prod_{j=1}^{\mu_k} \left(\mathbb{M}((1+2^{-n})\,\underline{z}) - \kappa_{k,j}^{(n)}\, I\right).$$

Passing to the limit, we obtain

$$\underline{X} \in \mathrm{Ker} \prod_{k=1}^{K} \left(\mathbb{M}(\underline{z}) - \underline{\kappa}_k\, I\right)^{\mu_k} = \oplus_{k=1}^K \mathrm{Ker}\ (\mathbb{M}(\underline{z}) - \underline{\kappa}_k\, I)^{\mu_k}\,.$$

This relation shows that $\mathbb{E}^s(\underline{z})$ is contained in the vector space on the right hand-side of (112). We also know that $\mathbb{E}^s(\underline{z})$ has dimension r. Furthermore, $\mathbb{M}(\underline{z})$ is a companion matrix so we can apply Lemma 2.8 and Lemma 4.8 which show that each vector space $\mathrm{Ker}\ (\mathbb{M}(\underline{z}) - \underline{\kappa}_k\, I)^{\mu_k}$ has dimension μ_k. Since the sum of all the μ_k's equals r, we have obtained (112) for all $\underline{z} \in \overline{\mathscr{U}}$.

• The resolvent equation for (111) reads (59) with

$$\forall\, z \in \mathbb{C} \setminus \{0\}\,,\ \mathbb{B}(z) = \mathbb{B} := \begin{pmatrix} 0 & \dots & 0 & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ 0 & \dots & 0 & 0 & & 1 \end{pmatrix} \in \mathscr{M}_{r,p+r}(\mathbb{C})\,. \tag{113}$$

We recall that we consider the case of scalar problems so N equals 1 here. Applying Proposition 4.1, we need to show that the kernel of the constant matrix $\mathbb{B}$ does not contain any element of $\mathbb{E}^s(\underline{z})$

for all $\underline{z} \in \overline{\mathscr{U}}$. Consequently, let $\underline{z} \in \overline{\mathscr{U}}$. By the noncharacteristic discrete boundary assumption, we know that the companion matrix $\mathbb{M}(\underline{z})$ does not have 0 as an eigenvalue. We recall (112) and use Lemma 4.9 to compute a basis of $\mathbb{E}^s(\underline{z})$. Up to reordering the eigenvalues, we can assume that $\mu_k > 0$ for all $k = 1, \dots, K$ and do not consider the other eigenvalues of $\mathbb{M}(\underline{z})$ anylonger. For each k, we define the polynomials

$$P_{k,1}(X) := 1\,, \quad P_{k,2}(X) := \frac{1}{\underline{\kappa}_k}\, X\,, \quad \dots\,, \quad P_{k,\mu_k}(X) := \frac{1}{\underline{\kappa}_k^{\mu_k-1}} \prod_{j=0}^{\mu_k-2} (X-j)\,. \tag{114}$$

It is clear that the polynomials $P_{k,\ell}$, $1 \le \ell \le \mu_k$, span $\mathbb{C}_{\mu_k-1}[X]$ and Lemma 4.9 shows that the vectors

$$E_{k,1} := \begin{pmatrix} P_{k,1}(p+r-1)\,\underline{\kappa}_k^{p+r-1} \\ \vdots \\ P_{k,1}(1)\,\underline{\kappa}_k \\ P_{k,1}(0) \end{pmatrix}, \quad \dots\,, \quad E_{k,\mu_k} := \begin{pmatrix} P_{k,\mu_k}(p+r-1)\,\underline{\kappa}_k^{p+r-1} \\ \vdots \\ P_{k,\mu_k}(1)\,\underline{\kappa}_k \\ P_{k,\mu_k}(0) \end{pmatrix},$$

span Ker $(\mathbb{M}(\underline{z}) - \underline{\kappa}_k\, I)^{\mu_k}$. Using the decomposition (112), we wish to show that the vectors $\mathbb{B}\, E_{k,j}$, $1 \le k \le K$, $1 \le j \le \mu_k$, are linearly independent. This is indeed equivalent to showing that the kernel of $\mathbb{B}$ does not intersect $\mathbb{E}^s(\underline{z})$. Applying the matrix $\mathbb{B}$ in (113) to a vector of $\mathbb{C}^{p+r}$ amounts to keeping only the last r coordinates of the vector. Therefore, showing that the kernel of $\mathbb{B}$ does not intersect $\mathbb{E}^s(\underline{z})$ amounts to proving that the matrix

$$\begin{pmatrix} P_{1,1}(r-1)\,\underline{\kappa}_1^{r-1} & \dots & P_{1,1}(1)\,\underline{\kappa}_1 & P_{1,1}(0) \\ \vdots & & & \vdots \\ P_{1,\mu_1}(r-1)\,\underline{\kappa}_1^{r-1} & \dots & P_{1,\mu_1}(1)\,\underline{\kappa}_1 & P_{1,\mu_1}(0) \\ \vdots & & & \vdots \\ P_{K,1}(r-1)\,\underline{\kappa}_K^{r-1} & \dots & P_{K,1}(1)\,\underline{\kappa}_K & P_{K,1}(0) \\ \vdots & & & \vdots \\ P_{K,\mu_K}(r-1)\,\underline{\kappa}_K^{r-1} & \dots & P_{K,\mu_K}(1)\,\underline{\kappa}_K & P_{K,\mu_K}(0) \end{pmatrix} \tag{115}$$

is invertible. (In (115), the first μ_1 rows correspond to $(\mathbb{B}\, E_{1,1})^T, \dots, (\mathbb{B}\, E_{1,\mu_1})^T$ and so on.) Before going on, let us observe that when all the μ_k's equal 1, then K equals r and the latter matrix coincides with the Vandermonde matrix

$$\begin{pmatrix} \underline{\kappa}_1^{r-1} & \dots & \underline{\kappa}_1 & 1 \\ \vdots & & \vdots & \vdots \\ \underline{\kappa}_r^{r-1} & \dots & \underline{\kappa}_r & 1 \end{pmatrix},$$

which is known to be invertible (the $\underline{\kappa}_k$'s are pairwise ditinct). Let us go back to the general case and assume that the vector $(c_{r-1}, \dots, c_0)^T$ belongs to the kernel of the matrix in (115). We define the polynomial

$$\mathbf{P}(X) := c_0 + \cdots + c_{r-1}\, X^{r-1}\,.$$

For all $j = 1, \dots, \mu_1$, there holds

$$\sum_{\ell=0}^{r-1} c_\ell\, P_{1,j}(\ell)\, \underline{\kappa}_1^\ell = 0\,.$$

From the definition (114) of the polynomials $P_{1,j}$, we have

$$P_{1,j}(\ell) = \begin{cases} \underline{\kappa}_1^{1-j}\, \dfrac{\ell!}{(\ell+1-j)!}\,, & \text{if } j \le \ell+1, \\ 0\,, & \text{otherwise.} \end{cases}$$

We therefore obtain

$$\forall\, j = 1, \dots, \mu_1\,, \quad \sum_{\ell=j-1}^{r-1} c_\ell\, \frac{\ell!}{(\ell+1-j)!}\, \underline{\kappa}_1^{\ell+1-j} = 0\,,$$

or equivalently

$$\forall\, j = 1, \dots, \mu_1\,, \quad \mathbf{P}^{(j-1)}(\underline{\kappa}_1) = 0\,.$$

The same analysis can be done for all the $\underline{\kappa}_k$'s, and we find that $\mathbf{P}$ can be factorized by

$$\prod_{k=1}^{K} (X - \underline{\kappa}_k)^{\mu_k}\,.$$

Since the sum of the μ_k's equals r, and the degree of $\mathbf{P}$ does not exceed $r-1$, we can conlude that $\mathbf{P}$ equals 0, or equivalently that the kernel of the matrix in (115) is trivial. We have thus shown that the Uniform Kreiss-Lopatinskii Condition is satisfied by the Dirichlet boundary conditions and Theorem 3.5 shows that the numerical scheme (111) is strongly stable. □

5. Fully discrete initial boundary value problems: stability with general initial data

The goal of this section is to understand how one can incorporate nonzero initial data in the numerical scheme (32). Of course, one can always consider initial conditions $(f^0),\dots,(f^s)$ in (32), and the numerical scheme is still well-defined. The main problem is to understand how one can control the numerical solution (U_j^n) in $\ell_n^\infty(\ell_j^2)$. In particular, if one can show a bound of the form $\|U^n\|_{\ell_j^2} \le C^n \|U^0\|_{\ell_j^2}$, $C>1$, this would correspond for the continuous problem to a bound of the form $\|u(t)\|_{L^2(\mathbb{R})} \le C^{t/\Delta t} \|u|_{t=0}\|_{L^2(\mathbb{R})}$, which would be useless in the limit $\Delta t \to 0$. Basically, we are looking for an energy estimate of the solution in $\ell_n^\infty(\ell_j^2)$ that is compatible, in the limit $\Delta t \to 0$, with an energy estimate for the continuous problem.

5.1. A simple but unsufficient argument.

As we have seen in Section 2, it is very easy to incorporate initial conditions for the Cauchy problem and to obtain $\ell_n^\infty(\ell_j^2)$ bounds thanks to Fourier transform. Using the linearity of (32), we can thus try to decompose the solution (U_j^n) as the sum $U_j^n = V_j^n + W_j^n$, where (V_j^n) is a solution to a Cauchy problem and (W_j^n) is a solution to a problem of the form (32) with zero initial data. This strategy gives the following result.

Proposition 5.1. *Let us assume that the numerical scheme (14) is stable for the Cauchy problem (in the sense of Definition 2.2) and that (32) is strongly stable in the sense of Definition 3.1. Then there exists a constant $C>0$ such that for all $\gamma \ge 1$ and for all $\Delta t \in]0,1]$, the solution to (32) satisfies*

$$\begin{aligned}\sup_{n\ge 0} \mathrm{e}^{-2\gamma n\Delta t} \sum_{j\ge 1-r} \Delta x\, |U_j^n|^2 + \frac{\gamma}{\gamma\Delta t+1} \sum_{n\ge 0}\sum_{j\ge 1-r} \Delta t\,\Delta x\, \mathrm{e}^{-2\gamma n\Delta t}\, |U_j^n|^2 \\ + \sum_{n\ge 0}\sum_{j=1-r}^{p} \Delta t\, \mathrm{e}^{-2\gamma n\Delta t}\, |U_j^n|^2 \le \frac{C}{\Delta t^2} \left\{ \sum_{j\ge 1-r} \Delta x \left(|f_j^0|^2 + \cdots + |f_j^s|^2 \right) \right. \\ \left. + \frac{\gamma\Delta t+1}{\gamma} \sum_{n\ge s}\sum_{j\ge 1} \Delta t\,\Delta x\, \mathrm{e}^{-2\gamma(n+1)\Delta t}\, |F_j^n|^2 + \sum_{n\ge s+1}\sum_{j=1-r}^{0} \Delta t\, \mathrm{e}^{-2\gamma n\Delta t}\, |g_j^n|^2 \right\}. \end{aligned} \tag{116}$$

Proof of Proposition 5.1. • We first extend the initial conditions $(f^0),\dots,(f^s)$ and the interior source term (F_j^n) by zero for $j\le -r$. We also decompose the solution (U_j^n) to (32) as $U_j^n = V_j^n + W_j^n$, where (V_j^n) is a solution to

$$\begin{cases} V_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} Q_\sigma\, V_j^{n-\sigma} + \Delta t\, F_j^n\,, & j\in\mathbb{Z}\,, \quad n\ge s\,, \\ V_j^n = f_j^n\,, & j\in\mathbb{Z}\,, \quad n=0,\dots,s\,, \end{cases} \tag{117}$$

and (W_j^n) is a solution to

$$\begin{cases} W_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} Q_\sigma\, W_j^{n-\sigma}\,, & j\ge 1\,, \quad n\ge s\,, \\ W_j^{n+1} = \displaystyle\sum_{\sigma=-1}^{s} B_{j,\sigma}\, W_1^{n-\sigma} + \widetilde{g}_j^{n+1}\,, & j=1-r,\dots,0\,, \quad n\ge s\,, \\ W_j^n = 0\,, & j\ge 1-r\,, \quad n=0,\dots,s\,, \end{cases} \tag{118}$$

with

$$\forall\, j=1-r,\dots,0\,, \quad \forall\, n\ge s\,, \quad \widetilde{g}_j^{n+1} := g_j^{n+1} - V_j^{n+1} + \sum_{\sigma=-1}^{s} B_{j,\sigma}\, V_1^{n-\sigma}\,. \tag{119}$$

This strategy will allow us to use the strong stability assumption for (32) on the sequence (W_j^n) since the initial conditions for (118) vanish.

• Our first goal is to estimate (V_j^n). We start from (117) and apply a partial Fourier transform with respect to the space variable (as in the proof of Proposition 2.2). With the amplification matrix $\mathscr{A}$ defined in (16), we obtain

$$\forall\, n \geq s\,, \quad \forall\, \xi \in \mathbb{R}\,, \quad \begin{pmatrix} \widehat{V^{n+1}}(\xi) \\ \vdots \\ \widehat{V^{n+1-s}}(\xi) \end{pmatrix} = \mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi}) \begin{pmatrix} \widehat{V^n}(\xi) \\ \vdots \\ \widehat{V^{n-s}}(\xi) \end{pmatrix} + \Delta t \begin{pmatrix} \widehat{F^n}(\xi) \\ \vdots \\ 0 \end{pmatrix}.$$

This relation yields

$$\forall\, n \geq s\,, \forall\, \xi \in \mathbb{R}\,, \quad \begin{pmatrix} \widehat{V^n}(\xi) \\ \vdots \\ \widehat{V^{n-s}}(\xi) \end{pmatrix} = \mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi})^{n-s} \begin{pmatrix} \widehat{V^s}(\xi) \\ \vdots \\ \widehat{V^0}(\xi) \end{pmatrix} + \Delta t \sum_{m=s}^{n-1} \mathscr{A}(\mathrm{e}^{i\,\Delta x\,\xi})^{n-1-m} \begin{pmatrix} \widehat{F^m}(\xi) \\ \vdots \\ 0 \end{pmatrix}.$$

Using the uniform bound for the amplification matrix $\mathscr{A}$ (here we use the stability assumption for the Cauchy problem), we obtain, for a given numerical constant C_0,

$$\forall\, n \geq s\,, \forall\, \xi \in \mathbb{R}\,, \quad \left|\widehat{V^n}(\xi)\right| + \cdots + \left|\widehat{V^{n-s}}(\xi)\right| \leq C_0 \left(\left|\widehat{f^s}(\xi)\right| + \cdots + \left|\widehat{f^0}(\xi)\right|\right) + C_0\,\Delta t \sum_{m=s}^{n-1} \left|\widehat{F^m}(\xi)\right|. \tag{120}$$

It only remains to "integrate" (120) with respect to n. For the sake of clarity, we state this kind of Gronwall inequality separately (the proof is a simple application of the $\ell^1 \star \ell^2$ convolution inequality and we leave it as an exercise for the interested reader).

Lemma 5.1. *Let s be an integer, and let C_1, C_2 be some nonnegative constants. Let $(a_n)_{n\geq s}$ and $(b_n)_{n\geq s}$ denote some sequences of nonnegative numbers that satisfy*

$$\forall\, n \geq s\,, \quad a_n \leq C_1\, a_s + C_2 \sum_{m=s}^{n-1} b_m\,.$$

Then for all $\gamma > 0$ and all $\Delta t \in]0,1]$, there holds

$$\sup_{n\geq s} \mathrm{e}^{-2\,\gamma\,n\,\Delta t}\, a_n^2 + \frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq s+1} \Delta t\, \mathrm{e}^{-2\,\gamma\,n\,\Delta t}\, a_n^2 \leq 2\, C_1^2\, \mathrm{e}^{-2\,\gamma\,s\,\Delta t}\, a_s^2 + 2\, \frac{C_2^2}{\Delta t^2}\, \frac{1+\gamma\,\Delta t}{\gamma} \sum_{n\geq s} \Delta t\, \mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\, b_n^2\,.$$

We apply Lemma 5.1 to (120), and obtain

$$\begin{aligned} &\sup_{n\geq s} \mathrm{e}^{-2\,\gamma\,n\,\Delta t} \left|\widehat{V^n}(\xi)\right|^2 + \frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq s+1} \Delta t\, \mathrm{e}^{-2\,\gamma\,n\,\Delta t} \left|\widehat{V^n}(\xi)\right|^2 \\ &\quad \leq C \left(\mathrm{e}^{-2\,\gamma\,s\,\Delta t} \left(\left|\widehat{f^0}(\xi)\right|^2 + \cdots + \left|\widehat{f^s}(\xi)\right|^2\right) + \frac{1+\gamma\,\Delta t}{\gamma} \sum_{n\geq s} \Delta t\, \mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t} \left|\widehat{F^n}(\xi)\right|^2 \right) \\ &\quad \leq C \left(\left|\widehat{f^0}(\xi)\right|^2 + \cdots + \left|\widehat{f^s}(\xi)\right|^2 + \frac{1+\gamma\,\Delta t}{\gamma} \sum_{n\geq s} \Delta t\, \mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t} \left|\widehat{F^n}(\xi)\right|^2 \right), \end{aligned}$$

with an appropriate numerical constant C. We integrate the latter inequality with respect to ξ, use Plancherel's and Fubini's Theorems, and obtain our first main estimate for the sequence (V_j^n):

$$\begin{aligned} &\sup_{n\geq s} \mathrm{e}^{-2\,\gamma\,n\,\Delta t} \sum_{j\in\mathbb{Z}} \Delta x\, |V_j^n|^2 + \frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq s+1} \sum_{j\in\mathbb{Z}} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\,n\,\Delta t}\, |V_j^n|^2 \\ &\quad \leq C \left(\sum_{j\geq 1-r} \Delta x \left(|f_j^0|^2 + \cdots + |f_j^s|^2\right) + \frac{1+\gamma\,\Delta t}{\gamma} \sum_{n\geq s} \sum_{j\geq 1-r} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\, |F_j^n|^2 \right). \end{aligned} \tag{121}$$

Observe that in the right hand-side of (121), the sums with respect to j only start at $j = 1-r$ since the initial conditions and the interior source term vanish for $j \leq -r$.

To make the estimates below easier to read, we define the quantity

$$\text{Source} := \sum_{j\geq 1-r} \Delta x \left(|f_j^0|^2 + \cdots + |f_j^s|^2 \right) + \frac{1+\gamma\,\Delta t}{\gamma} \sum_{n\geq s} \sum_{j\geq 1-r} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\, |F_j^n|^2 + \sum_{n\geq s+1} \sum_{j=1-r}^{0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |g_j^n|^2 ,$$

which gives a measure of the source terms with some appropriate weights. With this definition, the inequality (121) reads

$$\sup_{n\geq s} \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j\in\mathbb{Z}} \Delta x\, |V_j^n|^2 + \frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq s+1} \sum_{j\in\mathbb{Z}} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |V_j^n|^2 \leq C\,\text{Source}\,.$$

If we add some terms on the left hand-side that are obviously smaller than the right hand-side, we get

$$\sup_{n\geq 0} \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j\in\mathbb{Z}} \Delta x\, |V_j^n|^2 + \frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq 0} \sum_{j\in\mathbb{Z}} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |V_j^n|^2 \leq C\,\text{Source}\,.$$

We then easily deduce (here we use $\gamma \geq 1$):

$$\sum_{n\geq 0} \sum_{j=1-r}^{\max(p,q+1)} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |V_j^n|^2 \leq C\,\frac{1+\gamma\,\Delta t}{\gamma\,\Delta x}\,\text{Source} \leq C\,\frac{1+\Delta t}{\Delta x}\,\text{Source} \leq \frac{2\,C\,\lambda}{\Delta t}\,\text{Source}\,.$$

Combining with (121), we have already derived the inequality

$$\sup_{n\geq 0} \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j\geq 1-r} \Delta x\, |V_j^n|^2 + \frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq 0} \sum_{j\geq 1-r} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |V_j^n|^2 + \sum_{n\geq 0} \sum_{j=1-r}^{\max(p,q+1)} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |V_j^n|^2 \leq \frac{C}{\Delta t}\,\text{Source}\,. \tag{122}$$

with a new numerical constant that is still denoted C. The inequality (122) respresents "half" of (116). More precisely, it is now sufficient to prove a similar estimate to (122) for the sequence (W_j^n) and the combination of both estimates will give (116).

• We recall the definition (119) of the source term $\widetilde{g}_j^n$, $n \geq s+1$. The operators $B_{j,\sigma}$ are defined in (33). In particular, there exists a numerical constant C such that

$$\forall\, j = 1-r, \ldots, 0\,, \quad \forall\, n \geq s+1\,, \quad \left|\widetilde{g}_j^n\right| \leq |g_j^n| + |V_j^n| + C \sum_{\sigma=0}^{s+1} \sum_{\ell=1}^{q+1} |V_\ell^{n-\sigma}|\,.$$

We then obtain

$$\sum_{n\geq s+1} \sum_{j=1-r}^{0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, \left|\widetilde{g}_j^n\right|^2 \leq \sum_{n\geq s+1} \sum_{j=1-r}^{0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |g_j^n|^2 + \sum_{n\geq 0} \sum_{j=1-r}^{q+1} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |V_j^n|^2 \leq \frac{C}{\Delta t}\,\text{Source}\,,$$

where we have used (122) in the end to estimate the traces of (V_j^n) on $j = 1, \ldots, q+1$. We now use the fact that (118) is strongly stable and get

$$\frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq s+1} \sum_{j\geq 1-r} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |W_j^n|^2 + \sum_{n\geq s+1} \sum_{j=1-r}^{p} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |W_j^n|^2 \leq \frac{C}{\Delta t}\,\text{Source}\,.$$

Adding zero to the left hand-side (the initial conditions in (118) vanish), we obtain

$$\frac{\gamma}{1+\gamma\,\Delta t} \sum_{n\geq 0} \sum_{j\geq 1-r} \Delta t\,\Delta x\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |W_j^n|^2 + \sum_{n\geq 0} \sum_{j=1-r}^{p} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, |W_j^n|^2 \leq \frac{C}{\Delta t}\,\text{Source}\,. \tag{123}$$

Using (123), we derive the $\ell^\infty_n(\ell^2_j)$ estimate (we use the same type of inequalities as above):

$$\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j\geq 1-r}\Delta x\,|W_j^n|^2\leq\frac{1+\gamma\,\Delta t}{\gamma\,\Delta t}\,\frac{C}{\Delta t}\,\text{Source}\leq\frac{2\,C}{\Delta t^2}\,\text{Source}\,.$$

Combining with (123), we end up with

$$\sup_{n\geq 0}\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j\geq 1-r}\Delta x\,|W_j^n|^2+\frac{\gamma}{1+\gamma\,\Delta t}\sum_{n\geq 0}\sum_{j\geq 1-r}\Delta t\,\Delta x\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|W_j^n|^2$$
$$+\sum_{n\geq 0}\sum_{j=1-r}^{p}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|W_j^n|^2\leq\frac{C}{\Delta t^2}\,\text{Source}\,.\quad(124)$$

Summing (124) and (122), we complete the proof of Proposition 5.1. □

Of course, the result of Proposition 5.1 is not satisfactory because it does not give any information in the limit $\Delta t\to 0$. Nevertheless, the proof of Proposition 5.1 gave us the opportunity to introduce some major tools in the derivation of so-called semigroup estimates (meaning estimates in $\ell^\infty_n(\ell^2_j)$ for the solution). The first main tool is to introduce an auxiliary problem that takes care of the initial condition. By linearity of the problem, we are reduced to the case of zero initial data for (32). There are two important steps in the estimates of the solution, and at each of these steps we have lost one (large) factor Δt^{-1} in the proof of Proposition 5.1. The first crucial point is to obtain trace estimates for the solution to the auxiliary problem. These trace estimates should be obtained for a solution to a numerical scheme for which the initial conditions do not vanish (consequently it does not seem possible to exploit the results of Section 3 to derive these estimates). There is no clear reason why the solution to the Cauchy problem should satisfy a trace estimate uniformly in Δt, so our strategy in Proposition 5.1 looks a little hopeless. The second crucial point is to obtain semigroup estimates for the solution to (32) with zero initial data. Without any additional information, this step yields a factor Δt^{-1}, so a new strategy is needed.

As far as the choice of the auxiliary problem is concerned, we can try to follow Rauch's method [18]. The most simple strategy is to find some kind of "strictly dissipative" numerical boundary conditions. This strategy is the main guideline of [26] and was also used in [6] to extend the result of [26] to multidimensional problems.

5.2. **Wu's argument.** From now on, we consider numerical schemes with only one time step, meaning that $s=0$ in (32). Furthermore, in this paragraph, and this paragraph only, we consider scalar problems, meaning that $N=1$. The numerical scheme thus reads

$$\begin{cases}U_j^{n+1}=Q\,U_j^n\,, & j\geq 1\,,\quad n\geq 0\,,\\ U_j^{n+1}=\sum_{\ell=0}^{q}b_{\ell,j,-1}\,U_{1+\ell}^{n+1}+b_{\ell,j,0}\,U_{1+\ell}^n\,, & j=1-r,\dots,0\,,\quad n\geq 0\,,\\ U_j^0=f_j\,, & j\geq 1-r\,,\end{cases}\quad(125)$$

where the operator Q is given by

$$Q=\sum_{\ell=-r}^{p}a_\ell\,\mathbf{T}^\ell\,,\quad(a_{-r},\dots,a_p)\in\mathbb{R}^{p+r+1}\,,$$

and the $b_{\ell,j,-1},b_{\ell,j,0}$ are real numbers. The integer r and p in Q are fixed by the conditions $a_{-r}\neq 0$, $a_p\neq 0$. The unknown (U_j^n) in (125) is a sequence of real numbers. Let us first observe that the amplification matrix $\mathscr{A}$ associated with Q is a complex number, see (11). Consequently, if the numerical scheme for the Cauchy problem is stable in the sense of Definition 2.1, then one necessarily has (this is the von Neumann condition)

$$\forall\,\eta\in\mathbb{R}\,,\quad\left|\mathscr{A}(\mathrm{e}^{i\,\eta})\right|\leq 1\,,$$

and this implies

$$\forall\,v\in\ell^2(\mathbb{Z})\,,\quad\|Q\,v\|_{\ell^2(\mathbb{Z})}\leq\|v\|_{\ell^2(\mathbb{Z})}\,.\quad(126)$$

In other words, we are in the trivial case of stability for the Cauchy problem.

The following Lemma is proved in [26] and states that there exists at least one choice of numerical boundary conditions for which one can perform energy estimates "by hand" and incorporate nonzero initial data.

Lemma 5.2 ([26]). *Let either $r \geq 1$ or let $r = 0$ and $a_{-r} \neq 1$. Let us further assume that the operator Q in (125) satisfies (126). Then there exists a choice of real numbers $b_{1,aux}, \dots, b_{p+1,aux}$ such that the solution to*

$$\begin{cases} V_j^{n+1} = Q\, V_j^n\,, & j \geq 1\,, \quad n \geq 0\,, \\ V_j^{n+1} = 0\,, & j = 2-r, \dots, 0\,, \quad n \geq 0\,, \\ V_{1-r}^{n+1} = \sum_{\ell=0}^p b_{1+\ell,aux}\, V_{1+\ell}^{n+1}\,, & j = 1-r, \dots, 0\,, \quad n \geq 0\,, \\ V_j^0 = f_j\,, & j \geq 1-r\,, \end{cases} \tag{127}$$

satisfies

$$\sup_{n\geq 0} \sum_{j\geq 1-r} \Delta x\, |V_j^n|^2 + \sum_{n\geq 0} \Delta t \sum_{j=1-r}^{1} |V_j^n|^2 \leq C \sum_{j\geq 1-r} \Delta x\, |f_j|^2\,. \tag{128}$$

for all $\Delta t \in]0,1]$ with a constant C that does not depend on the initial condition (f_j) in (127), nor on Δt.

We refer to [26, page 84], see also [9, page 583], for the proof of Lemma 5.2. The estimate (128) is very strong because there is even no exponential weight in the terms on the left hand-side. Of course, one trivial consequence of (128) is the following estimate that looks more than what we were used to:

$$\sup_{n\geq 0} \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j\geq 1-r} \Delta x\, |V_j^n|^2 + \sum_{n\geq 0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j=1-r}^{1} |V_j^n|^2 \leq C \sum_{j\geq 1-r} \Delta x\, |f_j|^2\,.$$

One important thing to notice in Lemma 5.2 compared with Proposition 5.1 is that now we have a very good control of the trace of the solution to the auxiliary problem. Lemma 5.2 is the building block for proving the following Theorem that answers the problem of semigroup estimates for scalar equations and one time step schemes.

Theorem 5.1 (Semigroup stability for scalar problems [26]). *Let either $r \geq 1$ or let $r = 0$ and $a_{-r} \neq 1$. Let us consider the numerical scheme (125) with an operator Q that satisfies (126). Let us further assume that the scheme (125) is strongly stable in the sense of Definition 3.1. Then there exists a constant $C > 0$ such that for all $\gamma > 0$ and all $\Delta t \in]0,1]$, the solution to (125) satisfies*

$$\sup_{n\geq 0} \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j\geq 1-r} \Delta x\, |U_j^n|^2 \leq C \sum_{j\geq 1-r} \Delta x\, |f_j|^2\,.$$

The proof of Theorem 5.1 is based on a decomposition $U = V + W$ that is similar to the one used in the proof of Proposition 5.1. Lemma 5.2 gives the semigroup estimate for the auxiliary problem as well as some trace estimates. Unfortunately, Lemma 5.2 does not give a trace estimate for any fixed index j; it only gives a control of the traces from $j = 1-r$ up to $j = 1$. To control the traces for any index j, the argument in [26] relies on the Goldberg-Tadmor Lemma and this is the main point of the proof where it is crucial to deal with scalar equations. Deriving a semigroup estimate for W follows from the same argument as for V since we already know that the traces of W are controlled (this is the strong stability assumption). Since one step in the proof of Theorem 5.1 heavily relies on Proposition 4.2 (the Goldberg-Tadmor Lemma), it is not clear that the result extends to multidimensional systems because such systems usually do not reduce to decoupled scalar equations.

5.3. **A more general framework for semigroup stability.** Our goal in this paragraph is to propose an analogous method to that of Wu but that can be extended to multidimensional problems. In particular, a crucial issue is to avoid using the fact that the equation is scalar, or to avoid using Proposition 4.2. One should perform similar calculations to those in [26] but always in a vectorial framework. The main point to keep in mind is that (126) is a property that can hold even for non-scalar problems and this will be our starting point for the analysis of this paragraph. The results

that we present here are all taken from [6]. As in the preceeding paragraph, we restrict to one time step schemes:

$$\begin{cases} U_j^{n+1} = Q\,U_j^n + \Delta t\,F_j^n\,, & j \geq 1\,, \quad n \geq 0\,, \\ U_j^{n+1} = \sum_{\ell=0}^q B_{\ell,j,-1}\,U_{1+\ell}^{n+1} + B_{\ell,j,0}\,U_{1+\ell}^n + g_j^{n+1}\,, & j = 1-r,\dots,0\,, \quad n \geq 0\,, \\ U_j^0 = f_j\,, & j \geq 1-r\,, \end{cases} \tag{129}$$

where the operator Q is given by

$$Q = \sum_{\ell=-r}^{p} A_\ell\,\mathbf{T}^\ell\,, \quad A_{-r},\dots,A_p \in \mathscr{M}_N(\mathbb{R})\,,$$

and the unknown (U_j^n) in (129) is a sequence of vectors in $\mathbb{R}^N$. Similarly, the matrices $B_{\ell,j,-1}, B_{\ell,j,0}$ in (129) belong to $\mathscr{M}_N(\mathbb{R})$. We then make the following assumption.

Assumption 5.1 (Trivial stability for the Cauchy problem)**.** *The operator Q in (129) satisfies $\|Q\,v\|_{\ell^2(\mathbb{Z})} \leq \|v\|_{\ell^2(\mathbb{Z})}$ for all $v \in \ell^2(\mathbb{Z})$.*

For simplicty, we shall use the following notation for the ℓ^2 norms: $\Delta x > 0$ being the space step, then for all integers $m_1 \leq m_2$, we set

$$\|V\|_{m_1,m_2}^2 := \Delta x \sum_{j=m_1}^{m_2} |V_j|^2$$

to denote the ℓ^2-norm on the interval $[m_1, m_2]$ (m_1 may equal $-\infty$ and m_2 may equal $+\infty$). The corresponding scalar product is denoted by $(\cdot,\cdot)_{m_1,m_2}$. Our main result gives semigroup estimates as well as interior and trace estimates for the solution to (129) with arbitrary initial data in ℓ^2.

Theorem 5.2 ([6])**.** *Let Assumptions 3.1 and 5.1 be satisfied, and assume that the scheme (129) is strongly stable in the sense of Definition 3.1. Then there exists a constant C such that for all $\gamma > 0$ and all $\Delta t \in]0,1]$, the solution to (129) satisfies the estimate*

$$\begin{aligned} &\sup_{n\geq 0} \mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|U^n\|_{1-r,+\infty}^2 + \frac{\gamma}{\gamma\,\Delta t + 1} \sum_{n\geq 0} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|U^n\|_{1-r,+\infty}^2 + \sum_{n\geq 0} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t} \sum_{j=1-r}^{p} |U_j^n|^2 \\ &\leq C \left\{ \|f\|_{1-r,+\infty}^2 + \frac{\gamma\,\Delta t + 1}{\gamma} \sum_{n\geq 0} \Delta t\,\mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\,\|F^n\|_{1,+\infty}^2 + \sum_{n\geq 1} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t} \sum_{j=1-r}^{0} |g_j^n|^2 \right\}. \end{aligned} \tag{130}$$

As in [26], the proof of Theorem 5.2 relies on the introduction of an auxiliary problem where, compared with (129), we modify the numerical boundary conditions. Our auxiliary problem is not the same as in [26]. As a matter of fact, we directly show by means of the energy method that the Dirichlet boundary conditions are what we call *strictly dissipative*. This is an improved version of Proposition 4.2 since we are able to prove the strong stability estimate and also a semigroup estimate for the solution to the numerical scheme with Dirichlet boundary conditions and arbitrary initial data (recall that Proposition 4.2 first assumes that the equation is scalar and only gives a strong stability estimate for zero initial data). Moreover, since we are able to obtain a direct proof of the Goldberg-Tadmor Lemma (with an even stronger result), we do not need to rely anylonger on Proposition 4.2 and we can thus go further than the scalar case. More precisely, it is shown in [6] that the approach developed for proving Theorem 5.2 works in exactly the same way for multidimensional problems (not necessarily salar ones). As far as we know, this result even gives the first examples of strongly stable schemes for genuine multidimensional problems (meaning problems that do not reduce to scalar equations). What remains of this paragraph is devoted to the proof of Theorem 5.2. We first focus on the case of Dirichlet boundary conditions, and we shall then see how this preliminary result can be used to prove Theorem 5.2.

We therefore begin with the proof of the following refined version of Goldberg-Tadmor's Lemma. Considering the numerical scheme

$$\begin{cases} V_j^{n+1} = Q\, V_j^n + \Delta t\, F_j^n\,, & j \geq 1\,, \quad n \geq 0\,, \\ V_j^{n+1} = g_j^{n+1}\,, & j = 1-r, \dots, 0\,, \quad n \geq 0\,, \\ V_j^0 = f_j\,, & j \geq 1-r\,, \end{cases} \tag{131}$$

we are going to show

Theorem 5.3 ([6]). *Let Assumptions 3.1 and 5.1 be satisfied. Then there exists a constant C such that for all $\gamma > 0$ and all $\Delta t \in]0,1]$, the solution to (131) satisfies the estimate*

$$\begin{aligned} &\sup_{n\geq 0} \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, \|V^n\|_{1-r,+\infty}^2 + \frac{\gamma}{\gamma\,\Delta t + 1} \sum_{n\geq 0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t}\, \|V^n\|_{1-r,+\infty}^2 \\ &+ \sum_{n\geq 0} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j=1-r}^{\max(p,q+1)} |V_j^n|^2 \leq C \left\{ \|f\|_{1-r,+\infty}^2 + \frac{\gamma\,\Delta t + 1}{\gamma} \sum_{n\geq 0} \Delta t\, \mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\, \|F^n\|_{1,+\infty}^2 \right. \\ &\qquad \left. + \sum_{n\geq 1} \Delta t\, \mathrm{e}^{-2\,\gamma\, n\,\Delta t} \sum_{j=1-r}^{0} |g_j^n|^2 \right\}. \end{aligned} \tag{132}$$

In particular, the discretization (131) is strongly stable in the sense of Definition 3.1.

Proof of Theorem 5.3. • For simplicity, we shall give the proof of Theorem 5.3 in the special case where $F_j^n = 0$ in (131). The argument is simpler in this case, and we refer to [6] for a complete treatment of the case with an interior source term. We decompose the operator Q as

$$Q := I + \widetilde{Q}\,.$$

Assumption 5.1 is then equivalent to the inequality

$$\forall\, V \in \ell^2\,, \quad 2\,\big(V, \widetilde{Q}\, V\big)_{-\infty,+\infty} + \|\widetilde{Q}\, V\|_{-\infty,+\infty}^2 \leq 0\,. \tag{133}$$

We first use the relation $V_j^{n+1} = (I + \widetilde{Q})\, V_j^n$ for $j \geq 1$ (recall that we assume $F_j^n = 0$ here), and derive

$$\|V^{n+1}\|_{1,+\infty}^2 - \|V^n\|_{1,+\infty}^2 = 2\,\big(V^n, \widetilde{Q}\, V^n\big)_{1,+\infty} + \|\widetilde{Q}\, V^n\|_{1,+\infty}^2\,. \tag{134}$$

For a fixed integer n, we introduce the sequence $(W_j)_{j\in\mathbb{Z}}$ such that $W_j = V_j^n$ for $j \geq 1-r$ and $W_j = 0$ for $j \leq -r$. Due to the structure of the operator $\widetilde{Q}$ (a linear combination of the shifts $\mathbf{T}^{-r}, \dots, \mathbf{T}^p$), we have $\widetilde{Q}\, W_j = 0$ if $j \leq -r-p$, and $\widetilde{Q}\, W_j = \widetilde{Q}\, V_j^n$ if $j \geq 1$. Using (133), we thus get

$$\begin{aligned} 0 &\geq 2\,\big(W, \widetilde{Q}\, W\big)_{-\infty,+\infty} + \|\widetilde{Q}\, W\|_{-\infty,+\infty}^2 \\ &= 2\,\big(W, \widetilde{Q}\, W\big)_{1-r,0} + 2\,\big(W, \widetilde{Q}\, W\big)_{1,+\infty} + \|\widetilde{Q}\, W\|_{1-r-p,-r}^2 + \|\widetilde{Q}\, W\|_{1-r,0}^2 + \|\widetilde{Q}\, W\|_{1,+\infty}^2 \\ &= 2\,\big(V^n, \widetilde{Q}\, W\big)_{1-r,0} + 2\,\big(V^n, \widetilde{Q}\, V^n\big)_{1,+\infty} + \|\widetilde{Q}\, W\|_{1-r-p,-r}^2 + \|\widetilde{Q}\, W\|_{1-r,0}^2 + \|\widetilde{Q}\, V^n\|_{1,+\infty}^2 \\ &= 2\,\big(V^n, \widetilde{Q}\, V^n\big)_{1,+\infty} + \|\widetilde{Q}\, V^n\|_{1,+\infty}^2 + \|V^n + \widetilde{Q}\, W\|_{1-r,0}^2 + \|\widetilde{Q}\, W\|_{1-r-p,-r}^2 - \|V^n\|_{1-r,0}^2\,. \end{aligned} \tag{135}$$

We insert (135) into (134) and obtain

$$\|V^{n+1}\|_{1,+\infty}^2 - \|V^n\|_{1,+\infty}^2 + \|\widetilde{Q}\, W\|_{1-r-p,-r}^2 + \|V^n + \widetilde{Q}\, W\|_{1-r,0}^2 \leq \|V^n\|_{1-r,0}^2\,. \tag{136}$$

At this point, two situations may occur depending on the integer p. Let us first consider the case $p \geq 1$. Then, by Assumption 3.1 (with $s = 0$), A_p is an invertible matrix and the following result holds.

Lemma 5.3. *Let $p \geq 1$ and let A_p be invertible. Then there exists a constant $c > 0$ that does not depend on Δt nor on V^n such that the following estimate holds:*

$$\|\widetilde{Q}\, W\|_{1-r-p,-r}^2 + \|V^n + \widetilde{Q}\, W\|_{1-r,0}^2 \geq c\, \|V^n\|_{1-r,p}^2\,.$$

Proof of Lemma 5.3. Proving Lemma 5.3 is equivalent to proving that the quadratic form (that is independent on n)

$$(V^n_{1-r},\dots,V^n_p) \longmapsto \sum_{j=1-r-p}^{-r} |\widetilde{Q}\,W_j|^2 + \sum_{j=1-r}^{0} |V^n_j + \widetilde{Q}\,W_j|^2 \tag{137}$$

is positive definite. Recall that W denotes the extension of V^n by zero for $j \le -r$. The quadratic form (137) is clearly nonnegative. Let us therefore consider some vector $(V^n_{1-r},\dots,V^n_p)$ that satisfies

$$\forall\, j = 1-r-p,\dots,-r, \quad \widetilde{Q}\,W_j = 0\,, \qquad \forall\, j = 1-r,\dots,0, \quad V^n_j + \widetilde{Q}\,W_j = 0\,. \tag{138}$$

We first show by induction on j that $V^n_j = 0$ for all $j = 1-r,\dots,p-r$. Let us recall that $p \ge 1$, so we can write $\widetilde{Q} = Q - I$ in the form

$$\widetilde{Q} = A_p\,\mathbf{T}^p + \sum_{\ell=-r}^{p-1} \widetilde{A}_\ell\,\mathbf{T}^\ell\,.$$

In particular, we have $\widetilde{Q}\,W_{1-r-p} = A_p\,V^n_{1-r}$, and $V^n_{1-r} = 0$ because A_p is invertible. For $j = 1-r-p,\dots,-r$, $\widetilde{Q}\,W_j$ equals $A_p\,V^n_{j+p}$ plus a linear combination of the V^n_ℓ, $\ell < j+p$. Since the first term V^n_{1-r} is zero, we can proceed by induction and get $V^n_{1-r} = \cdots = V^n_{p-r} = 0$.

We now use the second set of equalities in (138). In particular, we have $V^n_{1-r} + \widetilde{Q}\,W_{1-r} = \widetilde{Q}\,W_{1-r} = A_p\,V^n_{1-r+p}$. Therefore, $V^n_{1-r+p} = 0$, and the rest of the proof follows from another induction argument. We have thus shown that (138) implies $(V^n_{1-r},\dots,V^n_p) = 0$. Hence the quadratic form (137) is positive definite. The proof of Lemma 5.3 is complete. □

We now complete the estimate of the sequence (V^n_j). Going back to (136) and using Lemma 5.3, we have

$$\|V^{n+1}\|^2_{1,+\infty} - \|V^n\|^2_{1,+\infty} + c\,\Delta x \sum_{j=1-r}^{p} |V^n_j|^2 \le \Delta x \sum_{j=1-r}^{0} |V^n_j|^2\,. \tag{139}$$

The end of the proof consists in "integrating" (139) over $\mathbb{N}$. We let $\gamma > 0$ and, for the sake of clarity, we introduce the notation

$$\mathscr{V}_n := \mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|V^n\|^2_{1,+\infty}\,, \quad \mathscr{B}_n := \mathrm{e}^{-2\,\gamma\,n\,\Delta t} \sum_{j=1-r}^{p} |V^n_j|^2\,, \quad \mathscr{G}_n := \mathrm{e}^{-2\,\gamma\,n\,\Delta t} \sum_{j=1-r}^{0} |V^n_j|^2\,.$$

We multiply (139) by $\exp(-2\,\gamma\,n\,\Delta t)$ to obtain

$$\mathrm{e}^{2\,\gamma\,\Delta t}\,\mathscr{V}_{n+1} - \mathscr{V}_n + \frac{c}{\lambda}\,\Delta t\,\mathscr{B}_n \le \frac{1}{\lambda}\,\Delta t\,\mathscr{G}_n\,.$$

Summing this inequality from 0 to N yields

$$\mathrm{e}^{2\,\gamma\,\Delta t}\,\mathscr{V}_{N+1} + \frac{\mathrm{e}^{2\,\gamma\,\Delta t}-1}{\Delta t} \sum_{1}^{N} \Delta t\,\mathscr{V}_n + \frac{c}{\lambda} \sum_{0}^{N} \Delta t\,\mathscr{B}_n \le \mathscr{V}_0 + \frac{1}{\lambda} \sum_{0}^{N} \Delta t\,\mathscr{G}_n \le \mathscr{V}_0 + \frac{1}{\lambda} \sum_{n\ge 0} \Delta t\,\mathscr{G}_n\,.$$

Letting N tend to $+\infty$, we have proved

$$\mathrm{e}^{2\,\gamma\,\Delta t} \sup_{n\ge 1} \mathscr{V}_n + \gamma \sum_{n\ge 1} \Delta t\,\mathscr{V}_n + \sum_{n\ge 0} \Delta t\,\mathscr{B}_n \le C \left(\mathscr{V}_0 + \Delta x\,\mathscr{G}_0 + \sum_{n\ge 1} \Delta t\,\mathscr{G}_n \right)\,. \tag{140}$$

The right-hand side of (140) is directly estimated by the right-hand side of (132), see the definition above for $\mathscr{G}_n$ and use (131) (recall that there is no interior source term here). The constant C in (140) is independent of γ and Δt.

It remains to treat the case $p = 0$ for which Lemma 5.3 does not hold anymore. In this case, we go back to (136) and simply ignore the nonnegative "boundary terms" on the left hand-side

$$\|V^{n+1}\|^2_{1,+\infty} - \|V^n\|^2_{1,+\infty} \le \|V^n\|^2_{1-r,0}\,.$$

We then proceed as above (with the same notation) to derive the weighted-in-time estimate

$$\mathrm{e}^{2\,\gamma\,\Delta t}\,\sup_{n\geq 1}\mathscr{V}_n+\gamma\,\sum_{n\geq 1}\Delta t\,\mathscr{V}_n\leq C\left(\mathscr{V}_0+\Delta x\,\mathscr{G}_0+\sum_{n\geq 1}\Delta t\,\mathscr{G}_n\right).$$

We have thus derived the inequality

$$\mathrm{e}^{2\,\gamma\,\Delta t}\,\sup_{n\geq 1}\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|V^n\|_{1,+\infty}^2+\gamma\,\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|V^n\|_{1,+\infty}^2 \\ \leq C\left\{\|f\|_{1-r,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|g_j^n|^2\right\}.$$

Adding some terms on the left hand-side that are obivously estimated by the right hand-side, we obtain (recall $p=0$)

$$\sup_{n\geq 0}\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|V^n\|_{1-r,+\infty}^2+\frac{\gamma}{\gamma\,\Delta t+1}\,\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|V^n\|_{1-r,+\infty}^2+\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{p}|V_j^n|^2 \\ \leq C\left\{\|f\|_{1-r,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|g_j^n|^2\right\}. \quad (141)$$

• The estimate (141) completes the proof of Theorem 5.3 when $q<p$. We thus assume from now on $q\geq p$. In that case, we need some additional trace estimates, namely we need to control

$$\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=p+1}^{q+1}|V_j^n|^2.$$

This is done by using the "shift trick" introduced in [26]. More precisely, when $p\geq 1$, we define the sequence $W_j^n:=V_{j+1}^n$ for $n\geq 0$ and $j\geq 1-r$, which solves the system (recall that F_j^n equals 0 in (131) for the case that we consider here):

$$\begin{cases} W_j^{n+1}=Q\,W_j^n\,, & j\geq 1\,,\quad n\geq 0\,,\\ W_j^{n+1}=g_{j+1}^{n+1}\,, & j=1-r,\dots,-1\,,\quad n\geq 0\,,\\ W_0^{n+1}=V_1^{n+1}\,, & n\geq 0\,,\\ W_j^0=f_{j+1}\,, & j\geq 1-r\,. \end{cases}$$

Applying (141) to W, we obtain

$$\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|W_p^n|^2\leq C\left\{\|f\|_{2-r,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=2-r}^{0}|g_j^n|^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|V_1^n|^2\right\}.$$

Using again (141) to estimate the last term of the right hand-wide (this is possible because $p\geq 1$) yields

$$\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|V_{p+1}^n|^2\leq C\left\{\|f\|_{1-r,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|g_j^n|^2\right\}.$$

We have therefore derived a trace estimate for $(V_{p+1}^n)_{n\geq 0}$. A straightforward induction argument gives

$$\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=p+1}^{q+1}|V_j^n|^2\leq C\left\{\|f\|_{1-r,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|g_j^n|^2\right\}. \quad (142)$$

The combination of (141) and (142) proves the main stability estimate (132) for $p\geq 1$.

• To complete the proof of Theorem 5.3, we only need to show how to pass from (141) to (132) in the case $p=0$. Since (141) does not give any trace estimate for (V_j^n), $j\geq 1$, the shift argument of [26] cannot be used anylonger. From Assumption 3.1, we know that the spectral radius of A_0 is

strictly less than 1. Hence, there exist a positive definite symmetric matrix H and a positive number ε_0 such that if we consider the new Euclidean norm on $\mathbb{R}^D$

$$\forall\, X \in \mathbb{R}^D\,, \quad |X|_H := \sqrt{X^*\, H\, X}\,,$$

then we have

$$\forall\, X \in \mathbb{R}^D\,, \quad |A_0\, X|_H \leq \sqrt{1-2\,\varepsilon_0}\,|X|_H\,.$$

From the relation

$$V_1^{n+1} = A_0\, V_1^n + \sum_{\ell=-r}^{-1} A_\ell\, V_{1+\ell}^n = A_0\, V_1^n + \underbrace{\sum_{j=1-r}^{0} A_{j-1}\, g_j^n}_{=:X^n}\,,$$

where we use the notation $g_j^0 := f_j$ for $j = 1-r, \ldots, 0$, we get

$$\begin{aligned}|V_1^{n+1}|_H^2 &= |A_0\, V_1^n|_H^2 + 2\,(A_0\, V_1^n)^*\, H\, X^n + |X^n|_H^2\\ &\leq (1-2\,\varepsilon_0)\,|V_1^n|_H^2 + 2\,(A_0\, V_1^n)^*\, H\, X^n + |X^n|_H^2 \leq (1-\varepsilon_0)\,|V_1^n|_H^2 + (1+\varepsilon_0^{-1})\,|X^n|_H^2\,.\end{aligned}$$

By definition of X^n, this turns into

$$|V_1^{n+1}|_H^2 - |V_1^n|_H^2 + \varepsilon_0\,|V_1^n|_H^2 \leq C \sum_{j=1-r}^{0} |g_j^n|^2\,.$$

Using the same summation process as earlier, we obtain

$$\begin{aligned}&\left\{(1-\mathrm{e}^{-2\,\gamma\,\Delta t}) + \varepsilon_0\,\mathrm{e}^{-2\,\gamma\,\Delta t}\right\} \sum_{n\geq 0} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|V_1^n|_H^2\\ &\qquad\qquad \leq C\left\{\|f\|_{1-r,+\infty}^2 + \sum_{n\geq 1} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t} \sum_{j=1-r}^{0} |g_j^n|^2\right\}.\end{aligned}$$

The norm $|\cdot|_H$ and the standard Euclidean norm are equivalent, so that

$$\sum_{n\geq 0} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|V_1^n|^2 \leq C\left\{\|f\|_{1-r,+\infty}^2 + \sum_{n\geq 1} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t} \sum_{j=1-r}^{0} |g_j^n|^2\right\},$$

with a constant C that does not depend on γ nor on Δt. The proof of (132) follows from an induction argument where we apply the above method to recover the estimate for the trace $(V_j^n)_{n\geq 0}$, $j = 2, \ldots, q+1$. The proof of Theorem 5.3 is now complete. □

It only remains to show how Theorem 5.3, which is already interesting on its own, also implies Theorem 5.2.

Proof of Theorem 5.2. • We rewrite the solution to (129) as $U_j^n = V_j^n + W_j^n$, where (V_j^n) satisfies

$$\begin{cases} V_j^{n+1} = Q\, V_j^n + \Delta t\, F_j^n\,, & j \geq 1\,, \quad n \geq 0\,,\\ V_j^{n+1} = g_j^{n+1}\,, & j = 1-r, \ldots, 0\,, \quad n \geq 0\,,\\ V_j^0 = f_j\,, & j \geq 1-r\,,\end{cases}$$

and (W_j^n) satisfies

$$\begin{cases} W_j^{n+1} = Q\, W_j^n\,, & j \geq 1\,, \quad n \geq 0\,,\\ W_j^{n+1} = \sum_{\ell=0}^{q} B_{\ell,j,-1}\, W_{1+\ell}^{n+1} + B_{\ell,j,0}\, W_{1+\ell}^n + \widetilde{g}_j^{n+1}\,, & j = 1-r, \ldots, 0\,, \quad n \geq 0\,,\\ W_j^0 = 0\,, & j \geq 1-r\,.\end{cases} \tag{143}$$

The source term $\widetilde{g}$ in (143) is defined by

$$\forall\, j = 1-r, \ldots, 0\,, \quad \forall\, n \geq 1\,, \quad \widetilde{g}_j^n := \sum_{\ell=0}^{q} B_{\ell,j,-1}\, V_{1+\ell}^n + B_{\ell,j,0}\, V_{1+\ell}^{n-1}\,. \tag{144}$$

The estimate for (V_j^n) is given by Theorem 5.3. In addition, since the discretization (129) is strongly stable in the sense of Definition 3.1 and the initial data in (143) is zero, (W_j^n) satisfies

$$\frac{\gamma}{\gamma\,\Delta t+1}\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|W^n\|_{1-r,+\infty}^2+\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|W_j^n|^2 \\ \leq C\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|\widetilde{g}_j^n|^2\,.$$

The defining equation (144) together with (132) allow us to control the term involving $\widetilde{g}_j^n$ by

$$\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|\widetilde{g}_j^n|^2 \\ \leq C\left\{\|f\|_{1-r,+\infty}^2+\frac{\gamma\,\Delta t+1}{\gamma}\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\,\|F^n\|_{1,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|g_j^n|^2\right\}. \tag{145}$$

Hence, we obtain

$$\frac{\gamma}{\gamma\,\Delta t+1}\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|W^n\|_{1-r,+\infty}^2+\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|W_j^n|^2 \\ \leq C\left\{\|f\|_{1-r,+\infty}^2+\frac{\gamma\,\Delta t+1}{\gamma}\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\,\|F^n\|_{1,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|g_j^n|^2\right\}. \tag{146}$$

The combination of (146) for (W_j^n) and of (132) for (V_j^n) proves a first part of Theorem 5.2. To complete the proof, it only remains to control the $\ell_n^\infty(\ell_j^2)$ norm of (W_j^n).

• We start from (143) and apply the strategy of the proof of Theorem 5.3. Since the derivation of the inequality (136) only relies on Assumption 5.1 and not on the numerical boundary conditions, we have (just ignore the nonnegative boundary terms on the left hand-side of (136))

$$\|W^{n+1}\|_{1,+\infty}^2-\|W^n\|_{1,+\infty}^2\leq\|W^n\|_{1-r,0}^2\,.$$

We multiply this inequality by $\exp(-2\,\gamma\,n\,\Delta t)$ and use the summation process as in the proof of Theorem 5.3. Since the initial data for (143) vanish, this yields

$$\sup_{n\geq 0}\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|W^n\|_{1,+\infty}^2\leq C\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|W_j^n|^2\,.$$

We now use the strong stability of (143) and the above estimate for the source term $(\widetilde{g}_j^n)$ to derive

$$\sup_{n\geq 0}\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,\|W^n\|_{1,+\infty}^2 \\ \leq C\left\{\|f\|_{1-r,+\infty}^2+\frac{\gamma\,\Delta t+1}{\gamma}\sum_{n\geq 0}\Delta t\,\mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\,\|F^n\|_{1,+\infty}^2+\sum_{n\geq 1}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\sum_{j=1-r}^{0}|g_j^n|^2\right\}.$$

Summing the latter inequality with (146) and the estimate (132) for (V_j^n), we complete the proof of the estimate (130). □

5.4. **The Lax-Friedrichs scheme.** The above analysis applies to the Lax-Friedrichs scheme (20) provided that we can check Assumption 5.1. More precisely, let us consider the numerical scheme (20) with a real symmetric matrix A. The amplification matrix $\mathscr{A}_{LF}$ satisfies the von Neumann condition if $\lambda\,\rho(A) \leq 1$, see (21). Moreover, when A is symmetric, the amplification matrix $\mathscr{A}_{LF}$ is a normal matrix. Hence its norm equals its spectral radius and we can conlude that Assumption 5.1 is satisfied. We can now state our main result for the Lax-Friedrichs scheme with general boundary conditions:

$$\begin{cases} U_j^{n+1} = \dfrac{U_{j-1}^n + U_{j+1}^n}{2} - \dfrac{\lambda\,A}{2}\,(U_{j+1}^n - U_{j-1}^n) + \Delta t\,F_j^n\,, & j \geq 1\,, \quad n \geq 0\,, \\ U_0^{n+1} = \sum_{\ell=0}^q B_{\ell,-1}\,U_{1+\ell}^{n+1} + B_{\ell,0}\,U_{1+\ell}^n + g^{n+1}\,, & n \geq 0\,, \\ U_j^0 = f_j\,, & j \geq 0\,. \end{cases} \tag{147}$$

Theorem 5.4. *Let A be a real symmetric matrix and let $\lambda > 0$ satisfy $\lambda\,\rho(A) < 1$. If the numerical scheme (147) is strongly stable in the sense of Definition 3.1, then there exists a constant $C > 0$ such that for all $\gamma > 0$ and all $\Delta t \in]0,1]$, the solution to (147) satisfies the estimate (130).*

If all eigenvalues of A are negative, we have seen that the Neumann boundary condition $U_0^{n+1} = U_1^{n+1} + g^{n+1}$ yields a strongly stable scheme, so Theorem 5.4 applies. Of course, the result is not very spectacular for such simple numerical schemes, but for schemes that involve many grid points (as in the case of Runge-Kutta schemes introduced in Section 2), it can become very complicated to verify an estimate like (130). As observed in numerous places in these notes, our future goal is to extend all the results presented here to multidimensional problems and we hope that our future results may bring more significant progress in this direction.

6. A partial conclusion

In these notes, we have tried to make a general and complete presentation of the derivation of stability estimates for fully discretized hyperbolic initial boundary value problems. The theory involves quite many arguments that we briefly summarize.

(i) The stability theory for the discretized Cauchy problem gives rise to the well-known *von Neumann condition*. The latter is a necessary condition for stability. In the class of *geometrically regular operators*, it turns out to be also a sufficient condition for stability.

(ii) The stability theory for discretized initial boundary value problems deals first with problems with zero initial data. In that case, an appropriate notion of stability was introduced in [10] and is referred to as *strong stability*. Using the Laplace transform, strong stability is first shown to be equivalent to an estimate for the resolvent equation. This preliminary reduction shows that the so-called *Godunov-Ryabenkii condition* is necessary for strong stability to hold. A refined and more quantitative version of the Godunov-Ryabenkii condition arises for strongly stable schemes and was referred to in these notes as the *Uniform Kreiss-Lopatinskii condition*.

(iii) The difficult part of the theory is to show that this condition is not only necessary but also sufficient for strong stability. The main technical points for doing so is to reduce the symbol $\mathbb{M}$ of the resolvent equation to the discrete block structure and then to construct a Kreiss symmetrizer. Reducing the symbol to the discrete block structure is possible in the framework of geometrically regular operators, while the construction of a Kreiss symmetrizer also requires the the fulfillment of the UKLC.

(iv) Once the case of zero initial data is clarified, the remaining part of the theory consists in incorporating arbitrary initial data and proving semigroup estimates. This does not seem possible without any further assumption on the numerical schemes that we consider. In these notes, we have presented a general argument that works for many one time step schemes.

In the case of zero initial data, the stability theory presented here seems to be complete since we do not know of any stable discretization for the Cauchy problem that violates the geometric regularity condition. The situation becomes far less clear in several space dimensions. In that case, even simple examples show that geometric regularity can be lost and further arguments need to be developed. Dissipative schemes were considered in [17] and we hope to push the analysis beyond this class in a near future. From a practical point of view, it would also be very interesting to develop powerful computational tools to check the UKLC in some situations where it cannot be done analytically.

Incorporating nonzero initial data for one time step schemes works the same in one or several space dimensions with the argument presented here. Hence the main open problem is to consider numerical schemes with several time steps.

References

[1] H. Baumgärtel. *Analytic perturbation theory for matrices and operators*. Birkhäuser Verlag, 1985.

[2] S. Benzoni-Gavage, D. Serre. *Multidimensional hyperbolic partial differential equations*. Oxford University Press, 2007. First-order systems and applications.

[3] J. Chazarain, A. Piriou. *Introduction to the theory of linear partial differential equations*. North-Holland, 1982.

[4] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. *SIAM J. Numer. Anal.*, 47(4):2844–2871, 2009.

[5] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems II. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, X(1):37–98, 2011.

[6] J.-F. Coulombel, A. Gloria. Semigroup stability of finite difference schemes for multidimensional hyperbolic initial boundary value problems. *Math. Comp.*, 80(273):165–203, 2011.

[7] M. Goldberg, E. Tadmor. Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comp.*, 36(154):603–626, 1981.

[8] B. Gustafsson. The convergence rate for difference approximations to mixed initial boundary value problems. *Math. Comp.*, 29(130):396–406, 1975.

[9] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods*. John Wiley & Sons, 1995.

[10] B. Gustafsson, H.-O. Kreiss, and A. Sundström. Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comp.*, 26(119):649–686, 1972.

[11] L. Hörmander. *An introduction to complex analysis in several variables*. North-Holland, 1990.

[12] H.-O. Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comp.*, 22:703–714, 1968.

[13] H.-O. Kreiss. Initial boundary value problems for hyperbolic systems. *Comm. Pure Appl. Math.*, 23:277–298, 1970.

[14] G. Métivier. The block structure condition for symmetric hyperbolic problems. *Bull. London Math. Soc.*, 32:689–702, 2000.

[15] G. Métivier, K. Zumbrun. Symmetrizers and continuity of stable subspaces for parabolic-hyperbolic boundary value problems. *Discrete Contin. Dyn. Syst.*, 11(1):205–220, 2004.

[16] G. Métivier, K. Zumbrun. Hyperbolic boundary value problems for symmetric systems with variable multiplicities. *J. Differential Equations*, 211(1):61–134, 2005.

[17] D. Michelson. Stability theory of difference approximations for multidimensional initial-boundary value problems. *Math. Comp.*, 40(161):1–45, 1983.

[18] J. Rauch. $\mathscr{L}^2$ is a continuable initial condition for Kreiss' mixed problems. *Comm. Pure Appl. Math.*, 25:265–285, 1972.

[19] W. Rudin. *Real and complex analysis*. McGraw-Hill, 1987.

[20] M. Schatzman. *Numerical analysis*. Oxford University Press, 2002.

[21] D. Serre. *Matrices*. Graduate Texts in Mathematics. Springer-Verlag, 2002. Theory and applications.

[22] J. C. Strikwerda, B. A. Wade. A survey of the Kreiss matrix theorem for power bounded families of matrices and its extensions. In *Linear operators (Warsaw, 1994)*, volume 38 of *Banach Center Publ.*, pages 339–360. Polish Acad. Sci., 1997.

[23] E. Tadmor. The equivalence of L_2-stability, the resolvent condition, and strict H-stability. *Linear Algebra Appl.*, 41:151–159, 1981.

[24] E. Tadmor. Complex symmetric matrices with strongly stable iterates. *Linear Algebra Appl.*, 78:65–77, 1986.

[25] L. N. Trefethen. Group velocity in finite difference schemes. *SIAM Rev.*, 24(2):113–136, 1982.

[26] L. Wu. The semigroup stability of the difference approximations for initial-boundary value problems. *Math. Comp.*, 64(209):71–88, 1995.

E-mail address: `jean-francois.coulombel@math.univ-lille1.fr`