Fully discrete hyperbolic initial boundary value problems with nonzero initial data

Jean-François Coulombel

Fully discrete hyperbolic initial boundary value problems with nonzero initial data

Jean-François Coulombel*

This article is dedicated to Denis Serre.

December 2, 2014

Abstract

The stability theory for hyperbolic initial boundary value problems relies most of the time on the Laplace transform with respect to the time variable. For technical reasons, this usually restricts the validity of stability estimates to the case of zero initial data. In this article, we consider the class of *non-glancing* finite difference approximations to the hyperbolic operator. We show that the maximal stability estimates that are known for zero initial data and nonzero boundary source term extend to the case of nonzero initial data in ℓ^2. The main novelty of our approach is to cover finite difference schemes with an arbitrary number of time levels. As an easy corollary of our main trace estimate, we recover former stability results in the semigroup sense by Kreiss [Kre68] and Osher [Osh69b].

AMS classification: 65M12, 65M06, 35L50.
Keywords: hyperbolic systems, boundary conditions, difference approximations, stability, semigroup.

Throughout this article, we use the notation

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

We let $\mathscr{M}_{d,p}(\mathbb{K})$ denote the set of $d \times p$ matrices with entries in $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, and we use the notation $\mathscr{M}_d(\mathbb{K})$ when $p = d$. If $M \in \mathscr{M}_d(\mathbb{C})$, $\mathrm{sp}(M)$ denotes the spectrum of M, $\rho(M)$ denotes its spectral radius, and M^* denotes the conjugate transpose of M. We let I denote the identity matrix or the identity operator when it acts on an infinite dimensional space. We use the same notation $x^* y$ for the hermitian product of two vectors $x, y \in \mathbb{C}^d$ and for the euclidean product of two vectors $x, y \in \mathbb{R}^d$. The norm of a vector $x \in \mathbb{C}^d$ is $|x| := (x^* x)^{1/2}$. The corresponding norm on $\mathscr{M}_d(\mathbb{C})$ is denoted $\| \cdot \|$. We let ℓ^2 denote the set of square integrable sequences, without mentioning the indices of the sequences. Sequences may be valued in $\mathbb{C}^k$ for some integer k. In all this article, N is a fixed positive integer.

*CNRS and Université de Nantes, Laboratoire de Mathématiques Jean Leray (UMR CNRS 6629), 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France. Email: jean-francois.coulombel@univ-nantes.fr. Research of the author was supported by ANR project BoND, ANR-13-BS01-0009-01.

1 Introduction

We are interested in finite difference discretizations of hyperbolic initial boundary value problems. The *continuous* 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, for simplicity, we consider the half-line $\mathbb{R}^+$ as the space domain. The matrix $A \in \mathscr{M}_N(\mathbb{R})$ is assumed to be diagonalizable with real eigenvalues, and B is a matrix - not necessarily a square one - 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 condition. Well-posedness for (1) is equivalent to the *algebraic* condition:

$$\operatorname{Ker} B \cap \operatorname{Span}\big(r_1, \dots, r_{N_+}\big) = \{0\}\,,$$

where the vectors $r_1, \dots, r_{N_+}$ span the unstable subspace of A, which corresponds to incoming characteristics. Furthermore, the matrix B should have rank N_+. Provided these conditions are satisfied, the unique solution $u \in \mathscr{C}(\mathbb{R}^+; L^2(\mathbb{R}^+))$ to (1) depends continuously on $f \in L^2(\mathbb{R}^+)$, $g \in L^2(\mathbb{R}^+)$ and $F \in L^2(\mathbb{R}^+ \times \mathbb{R}^+)$. We refer to [BGS07, chapter 4] for a general presentation of the well-posedness theory for (1), as well as for its multidimensional analogue.

The well-posedness theory for finite difference discretizations of (1) is far less clear. Let us first set a few notation. We 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, and we also let p, q, r, s denote some fixed integers. The solution to (1) is approximated by means of 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 $\{x = 0\}$. We consider finite difference approximations of (1) that read:

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} Q_\sigma\, U_j^{n-\sigma} + \Delta t\, F_j^n\,, & j \geq 1\,, \quad n \geq s\,, \\ U_j^{n+1} = \displaystyle\sum_{\sigma=-1}^{s} B_{j,\sigma}\, U_1^{n-\sigma} + g_j^{n+1}\,, & j = 1-r, \dots, 0\,, \quad n \geq s\,, \\ U_j^n = f_j^n\,, & j \geq 1-r\,, \quad n = 0, \dots, s\,, \end{cases} \tag{2}$$

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

$$Q_\sigma := \sum_{\ell=-r}^{p} A_{\ell,\sigma}\, \mathbf{T}^\ell\,, \quad B_{j,\sigma} := \sum_{\ell=0}^{q} B_{\ell,j,\sigma}\, \mathbf{T}^\ell\,. \tag{3}$$

In (3), 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 $\mathbf{T}$ denotes the shift operator on the space grid: $(\mathbf{T}^\ell v)_j := v_{j+\ell}$.

Existence and uniqueness of a solution (U_j^n) to (2) is trivial since the numerical scheme is explicit, so the last requirement for well-posedness is continuous dependence of the solution on the three possible source terms (F_j^n), (g_j^n), (f_j^n). This is a stability problem, and several definitions can be chosen. The following one dates back to the fundamental contribution [GKS72], and is specifically relevant when the boundary conditions are non-homogeneous ($(g_j^n) \not\equiv 0$):

Definition 1 (Strong stability [GKS72]). *The finite difference approximation (2) is said to be "strongly stable" if there exists a constant C_1 such that for all $\gamma > 0$ and all $\Delta t \in]0,1]$, the solution (U_j^n) to (2) with $(f_j^0) = \cdots = (f_j^s) = 0$ satisfies the estimate:*

$$\frac{\gamma}{\gamma\,\Delta t + 1} \sum_{n \geq s+1} \sum_{j \geq 1-r} \Delta t\,\Delta x\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|U_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{p} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|U_j^n|^2$$
$$\leq C_1 \left\{ \frac{\gamma\,\Delta t + 1}{\gamma} \sum_{n \geq s} \sum_{j \geq 1} \Delta t\,\Delta x\,\mathrm{e}^{-2\,\gamma\,(n+1)\,\Delta t}\,|F_j^n|^2 + \sum_{n \geq s+1} \sum_{j=1-r}^{0} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|g_j^n|^2 \right\}. \quad (4)$$

Another more common notion of stability only deals with nonzero initial data in (2), and was considered in the earlier publications [Kre68, Osh69b, Osh69a]:

Definition 2 (Semigroup stability). *The finite difference approximation (2) is said to be "semigroup stable" if there exists a constant C_2 such that for all $\Delta t \in]0,1]$, the solution (U_j^n) to (2) with $(F_j^n) = (g_j^n) = 0$ satisfies the estimate:*

$$\sup_{n \geq 0} \sum_{j \geq 1-r} \Delta x\,|U_j^n|^2 \leq C_2 \sum_{n=0}^{s} \sum_{j \geq 1-r} \Delta x\,|f_j^n|^2\,. \quad (5)$$

Remark 1. *Both Definitions 1 and 2 are independent of the small parameter Δt because of the fixed ratio $\Delta t/\Delta x$. We could therefore assume $\Delta t = 1$, which we sometimes do later on, but have written (4) and (5) with Δt and Δx in order to make the connection with the "continuous" norms.*

Let us observe that semigroup stability for (2) amounts to requiring that the (linear) operator

$$(U^0, \ldots, U^s) \longmapsto (U^1, \ldots, U^{s+1})\,,$$

that is obtained by considering (2) in the case $(F_j^n) = (g_j^n) = 0$, be power bounded on $\ell^2 \times \cdots \times \ell^2$. Let us quote [TE05] at this stage: "The term *GKS-stable* is quite complicated. This is a special definition of stability (...) that involves exponential factors with respect to t and other algebraic terms that remove it significantly from the more familiar stability notion of bounded norms of powers." The goal of this article is to shed new light on the relations between these two notions of stability for (2).

There is clear evidence that semigroup stability does not imply strong stability for (2). One counter-example is given in [Tre84, page 361]. In the PDE multidimensional context, a very simple counter-example can be constructed by considering the symmetric hyperbolic operator

$$\partial_t + \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \partial_{x_1} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \partial_{x_2}$$

with maximally dissipative (but not strictly dissipative) boundary conditions. The maximal dissipativity property yields semigroup stability, see [BGS07, chapter 3], while the violation of the so-called Uniform Kreiss-Lopatinskii Condition precludes any trace estimate in L^2 of the solution in terms of the L^2 norm of the boundary source term.

Yet, a reasonable expectation is that strong stability does imply semigroup stability[1]. In the PDE multidimensional context, this was proved in [Kaj72, Rau72] for both symmetric and strictly hyperbolic

[1] This "uniform power boundedness conjecture" appears in an even stronger (!) version in [KW93].

operators, later extended in [Aud11] to hyperbolic operators with constant multiplicity, and recently in [Mét14] to an even wider class of hyperbolic operators. The symmetric case is more favorable and is easily dealt with by the introduction of auxiliary boundary conditions. Once again, the situation for difference approximations is not as complete. That strong stability implies semigroup stability is somehow hidden in the early works [Kre68, Osh69b, Osh69a] since the assumptions made there actually yield strong stability (even though only semigroup stability was proved then). The first general result on the "uniform power boundedness conjecture" dates back to [Wu95] but is restricted to the case $s = 0$ (numerical schemes with two time levels only) and to scalar problems. The analysis of [Wu95] was generalized in [CG11] to the case of systems in any space dimension, still under the restriction $s = 0$ and assuming that the discretized hyperbolic operator does not increase the ℓ^2 norm on all $\mathbb{Z}$ ($\mathbb{Z}^d$ in several space dimensions).

The present article is a first attempt to tackle the "uniform power boundedness conjecture" for schemes with more than two time levels, that is, when $s \geq 1$. Our main result, which is Theorem 1 below, gives a trace estimate for the solution to (2) in the case of nonzero initial data. We are not able yet to give a positive answer to the conjecture in a general framework, but we recover the results of [Kre68, Osh69b, Osh69a] as an easy corollary of Theorem 1. Unlike [Wu95, CG11], our argument does not use the auxiliary Dirichlet boundary condition but relies on an easy summation by parts argument, as what one does for toy problems such as the upwind or Lax-Friedrichs schemes. Unfortunately, this summation by parts argument is restricted so far to the case $s = 0$, but we do hope that our trace estimate for nonzero initial data does imply semigroup stability even for $s \geq 1$. This might require adapting the PDE arguments to the framework of difference approximations and is postponed to a future work.

2 Assumptions and main result

We adopt the framework of [Cou09, Cou11]. Let us first introduce the so-called amplification matrix:

$$\forall\, \kappa \in \mathbb{C} \setminus \{0\}\,,\ \mathscr{A}(\kappa) := \begin{pmatrix} \widehat{Q_0}(\kappa) & \dots & \dots & \widehat{Q_s}(\kappa) \\ I & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & I & 0 \end{pmatrix} \in \mathscr{M}_{N(s+1)}(\mathbb{C})\,, \quad \widehat{Q_\sigma}(\kappa) := \sum_{\ell=-r}^{p} \kappa^\ell\, A_{\ell,\sigma}\,. \tag{6}$$

A necessary condition for both strong and semigroup stability of (2) is that the discretization of the Cauchy problem be ℓ^2 stable. We thus make our first assumption.

Assumption 1 (Stability for the discrete Cauchy problem)**.** *There exists a constant $C > 0$ such that the amplification matrix $\mathscr{A}$ in (6) satisfies:*

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

In particular, the von Neumann condition $\rho(\mathscr{A}(\mathrm{e}^{i\,\eta})) \leq 1$ holds.

We then make the following geometric regularity assumption on the difference operators Q_σ in (2):

Assumption 2 (Geometrically regular operator)**.** *The amplification matrix $\mathscr{A}$ defined by (6) 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 $\underline{\alpha}$, then there exist some functions $\zeta_1(\kappa), \dots, \zeta_{\underline{\alpha}}(\kappa)$ that are holomorphic in a neighborhood $\mathscr{W}$ of $\underline{\kappa}$ in $\mathbb{C}$, that satisfy*

$$\zeta_1(\underline{\kappa}) = \dots = \zeta_{\underline{\alpha}}(\underline{\kappa}) = \underline{z}\,, \quad \det\big(z\,I - \mathscr{A}(\kappa)\big) = \vartheta(\kappa, z) \prod_{k=1}^{\underline{\alpha}} \big(z - \zeta_k(\kappa)\big)\,,$$

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 furthermore, there exist some vectors $e_1(\kappa), \dots, e_{\underline{\alpha}}(\kappa) \in \mathbb{C}^{N(s+1)}$ that depend holomorphically on $\kappa \in \mathscr{W}$, that are linearly independent for all $\kappa \in \mathscr{W}$, and that satisfy

$$\forall\, \kappa \in \mathscr{W}\,, \quad \forall\, k = 1, \dots, \underline{\alpha}\,, \quad \mathscr{A}(\kappa)\, e_k(\kappa) = \zeta_k(\kappa)\, e_k(\kappa)\,.$$

Let us recall that in the scalar case ($N = 1$), Assumption 2 is actually a consequence of Assumption 1, see [Cou13, Lemma 7]. For technical reasons to be specified later in Section 3, we make a final assumption on the amplification matrix $\mathscr{A}$:

Assumption 3 (Non-glancing discretization)**.** *The amplification matrix $\mathscr{A}$ defined by (6) 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 $\underline{\alpha}$, then the eigenvalues $\zeta_1(\kappa), \dots, \zeta_{\underline{\alpha}}(\kappa)$ of $\mathscr{A}(\kappa)$ that are close to $\underline{z}$ when κ is close to $\underline{\kappa}$ satisfy:*

$$\forall\, k = 1, \dots, \underline{\alpha}\,, \quad \zeta_k'(\underline{\kappa}) \neq 0\,.$$

Many standard finite difference approximations satisfy Assumptions 1, 2 and 3, as for instance the upwind, Lax-Friedrichs and Lax-Wendroff schemes under a suitable CFL condition. The leap-frog approximation satisfies Assumptions 1 and 2 but violates Assumption 3. The case $\zeta_k'(\underline{\kappa}) = 0$ gives rise to *glancing* wave packets with a vanishing group velocity, see [Tre82, Tre84]. Here we assume that no such wave packet occurs.

For geometrically regular operators, the main results of [Cou09, Cou11] show that strong stability is equivalent to an *algebraic condition*, known as the Uniform Kreiss-Lopatinskii Condition. Let us summarize the main steps in the analysis since some notation and results will be used later on. The main tool in the characterization of strong stability is the Laplace transform with respect to the time variable, which leads to the resolvent equation

$$\begin{cases} W_j - \displaystyle\sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, W_j = F_j\,, & j \geq 1\,, \\ W_j - \displaystyle\sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{j,\sigma}\, W_1 = g_j\,, & j = 1-r, \dots, 0\,, \end{cases} \tag{7}$$

with $z \in \mathscr{U}$. The induction relation (7) can be written in a more compact form by using an augmented vector. We introduce the matrices:

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

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

$$\forall\, \ell = 0, \dots, q\,, \quad \forall\, j = 1-r, \dots, 0\,, \quad \forall\, z \in \mathbb{C} \setminus \{0\}\,, \quad \mathbb{B}_{\ell,j}(z) := \sum_{\sigma=-1}^{s} z^{-\sigma-1}\, B_{\ell,j,\sigma}\,. \tag{8}$$

Our final assumption is rather standard and already appears in [Kre68].

Assumption 4 (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 $\varepsilon_0 \in]0, 1]$.*

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 (7) are coordinates of the augmented vector[2] $\mathscr{W}_1 := (W_p, \dots, W_{1-r}) \in \mathbb{C}^{N(p+r)}$. Using Assumption 4, 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 $\overline{\mathscr{U}}$:

$$\forall\, z \in \mathscr{V}\,, \quad \mathbb{M}(z) := \begin{pmatrix} -\mathbb{A}_p(z)^{-1}\,\mathbb{A}_{p-1}(z) & \dots & \dots & -\mathbb{A}_p(z)^{-1}\,\mathbb{A}_{-r}(z) \\ I & 0 & \dots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & 0 & I & 0 \end{pmatrix} \in \mathscr{M}_{N(p+r)}(\mathbb{C})\,. \tag{9}$$

We also define the matrix that encodes the boundary conditions for (7), namely

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

with the $\mathbb{B}_{\ell,j}$'s defined in (8). With such definitions, it is a simple exercise to rewrite the resolvent equation (7) 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

$$\begin{cases} \mathscr{W}_{j+1} = \mathbb{M}(z)\,\mathscr{W}_j + \mathscr{F}_j\,, \quad j \geq 1\,, \\ \mathbb{B}(z)\,\mathscr{W}_1 = \mathscr{G}\,, \end{cases} \tag{10}$$

where the new source terms $(\mathscr{F}_j), \mathscr{G}$ in (10) are given by:

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

There is a similar equivalent form of (7) in the case $q \geq p$, and we refer the reader to [Cou13, page 145] for the details. The main results of [GKS72] and later [Cou09, Cou11] characterize strong stability of (2) in terms of an algebraic condition that involves the matrices $\mathbb{M}(z)$ and $\mathbb{B}(z)$ in (10). This characterization of strong stability relies on a precise description of the stable and unstable spaces of the matrix $\mathbb{M}(z)$, including when z becomes arbitrarily close to the unit circle. Some ingredients of the analysis are recalled and used in Section 3.

Our main result is an estimate for the solution to (2) with nonzero initial data. This estimate is entirely similar to (4) as far as the left hand-side of the inequality is concerned. Namely, we extend the known estimate for zero initial data to nonzero initial data by simply adding the ℓ^2 norm of the initial data on the right hand-side of the inequality.

Theorem 1. *Let Assumptions 1, 2, 3 and 4 be satisfied. If the scheme (2) is strongly stable, then for all integer $P \in \mathbb{N}$, there exists a constant $C_P > 0$ such that for all $\gamma > 0$ and all $\Delta t \in]0, 1]$, the solution (U_j^n) to (2) with $(F_j^n) = (g_j^n) = 0$ satisfies the estimate:*

$$\frac{\gamma}{\gamma\,\Delta t + 1} \sum_{n \geq 0} \sum_{j \geq 1-r} \Delta t\,\Delta x\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|U_j^n|^2 + \sum_{n \geq 0} \sum_{j=1-r}^{P} \Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|U_j^n|^2 \leq C_P \sum_{n=0}^{s} \sum_{j \geq 1-r} \Delta x\,|f_j^n|^2\,. \tag{11}$$

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

The analogue of the estimate (11) is a key tool in [Kaj72] for proving the semigroup boundedness in the PDE multidimensional context. This requires however rather strong algebraic properties in order to justify some integration by parts argument (in a possibly non-symmetric context).

Let us now explain the links between Theorem 1 and previous results in the literature, and more specifically with the analysis in [Osh69b] (which is already an extension of [Kre68]). As explained earlier, Assumption 1 is necessary for any kind of stability result. It corresponds to condition (1) in the main Theorem of [Osh69b] (see [Osh69b, XIX]). Assumption 2 is automatically satisfied in [Osh69b] because the equations are scalar and the scheme involves only two time levels (recall that for $N = 1$, Assumption 2 actually follows from Assumption 1). Assumption 2 seems to be rather natural in one space dimension, whatever the values of N and s, see the discussion in [Cou13, Section 2.2]. Assumption 3 is hidden in condition (2) of the main Theorem of [Osh69b], but allows for slightly more general situations. Eventually, strong stability corresponds to condition (4) in the main Theorem of [Osh69b]. So at this stage, one might reasonably ask whether Theorem 1 does imply the main result of [Osh69b], that is, semigroup stability of (2) when $s = 0$. This is the purpose of the following Corollary.

Corollary 1. *In addition to Assumptions 1, 2, 3 and 4, let us assume*[3] $s = 0$ *and:*

$$\sum_{\ell=-r}^{p} A_{\ell,0} = I\,, \quad \|Q_0\|_{\ell^2(\mathbb{Z})\to\ell^2(\mathbb{Z})} = 1\,.$$

If the scheme (2) is strongly stable, then it is also semigroup stable.

We emphasize that the decomposition technique used in [Osh69b] does not seem to easily extend to the case $s \geq 1$, and this is the main reason why we advocate an alternative approach that is based on the trace estimate (11) and a suitable integration by parts formula (see Section 4 for the proof of Corollary 1). Comparing with the derivation of semigroup estimates for (2) in [Wu95, CG11], the present approach is closer to the one that has been used in the PDE context, see e.g. [Kaj72, Rau72, Aud11], and is also closer to the one that is used on toy problems such as the Lax-Friedrichs or upwind schemes, see [GKO95, chapter 11].

The proof of Theorem 1 is given in Section 3 and follows some arguments that appear in the surprisingly unnoticed[4] contribution [Sar65], see also [Sar77, section 5]. Our goal is to adapt such arguments to difference approximations and to make precise the new arguments involved in this extension. More precisely, the non-glancing Assumption is used in the proof of Theorem 1 to show a trace estimate for the solution to the fully discrete Cauchy problem on $\mathbb{Z}$. Thanks to this trace estimate, we can incorporate the initial data for (2) in the solution to a Cauchy problem, which reduces the study of (2) to zero initial data and nonzero boundary source term. There is a wide literature on trace operators for hyperbolic Cauchy problems, see for instance the "well-known", though unpublished, reference [MT] and works cited therein. We do not aim at a thorough description of the trace operator here, but rather focus on its ℓ^2-boundedness. As explained in Appendix A, ℓ^2-boundedness of the trace operator for the discrete Cauchy problem will be seen to be equivalent[5] to the non-glancing condition in Assumption 3.

[3] All these extra assumptions are also present in [Osh69b].

[4] Actually, one of the main results of [Sar65] shows that the uniform Lopatinskii condition is a sufficient condition for strong well-posedness of strictly hyperbolic initial boundary value problems, but the proof in [Sar65] is restricted to constant coefficients linear systems, while the technique developed in [Kre70] extends to variable coefficients and therefore to nonlinear problems by fixed point iteration. Another main result in [Sar65] gives stability estimates for solutions to initial boundary value problems with nonzero initial data, and this seems to be the first result of this kind for non-symmetric systems.

[5] The equivalent result for PDE problems seems to be part of folklore, though we have not found a detailed proof based on elementary arguments.

3 Proof of Theorem 1

From now on, we consider the scheme (2) and assume that it is strongly stable in the sense of Definition 1. When the interior and boundary source terms vanish, the scheme reads

$$\begin{cases} U_j^{n+1} = \sum_{\sigma=0}^{s} Q_\sigma \, U_j^{n-\sigma} \,, & j \geq 1 \,, \quad n \geq s \,, \\ U_j^{n+1} = \sum_{\sigma=-1}^{s} B_{j,\sigma} \, U_1^{n-\sigma} \,, & j = 1-r, \dots, 0 \,, \quad n \geq s \,, \\ U_j^n = f_j^n \,, & j \geq 1-r \,, \quad n = 0, \dots, s \,, \end{cases} \tag{12}$$

with initial data $f^0, \dots, f^s \in \ell^2$.

All constants appearing in the estimates below are independent of the Laplace parameter $\gamma > 0$, when present.

3.1 Reduction to a Cauchy problem

We decompose the solution (U_j^n) to (12) as $U_j^n = V_j^n + W_j^n$, where (V_j^n) satisfies a pure Cauchy problem that incorporates the initial data of (12), and (W_j^n) satisfies a system of the form (2) with zero initial data and nonzero boundary source term. More precisely, (V_j^n) denotes the solution to

$$\begin{cases} V_j^{n+1} = \sum_{\sigma=0}^{s} Q_\sigma \, V_j^{n-\sigma} \,, & j \in \mathbb{Z} \,, \quad n \geq s \,, \\ V_j^n = f_j^n \,, & j \geq 1-r \,, \quad n = 0, \dots, s \,, \\ V_j^n = 0 \,, & j \leq -r \,, \quad n = 0, \dots, s \,, \end{cases} \tag{13}$$

and (W_j^n) denotes the solution to

$$\begin{cases} W_j^{n+1} = \sum_{\sigma=0}^{s} Q_\sigma \, W_j^{n-\sigma} \,, & j \geq 1 \,, \quad n \geq s \,, \\ W_j^{n+1} = \sum_{\sigma=-1}^{s} B_{j,\sigma} \, W_1^{n-\sigma} + g_j^{n+1} \,, & j = 1-r, \dots, 0 \,, \quad n \geq s \,, \\ W_j^n = 0 \,, & j \geq 1-r \,, \quad n = 0, \dots, s \,, \end{cases} \tag{14}$$

where the source term (g_j^n) in (14) is defined by

$$\forall \, j = 1-r, \dots, 0 \,, \quad \forall \, n \geq s+1 \,, \quad g_j^n := -V_j^n + \sum_{\sigma=-1}^{s} B_{j,\sigma} \, V_1^{n-1-\sigma} \,. \tag{15}$$

The following result shows that Theorem 1 only relies on a trace estimate for the solution to (13).

Lemma 1. *Let Assumption 1 be satisfied. Assume furthermore that for all $P \in \mathbb{N}$, there exists a constant $C_P > 0$, that does not depend on the initial data in (13), such that the solution (V_j^n) to (13) satisfies*

$$\sum_{n \geq 0} \sum_{j=1-r}^{P} |V_j^n|^2 \leq C_P \sum_{n=0}^{s} \sum_{j \geq 1-r} |f_j^n|^2 \,. \tag{16}$$

Then the conclusion of Theorem 1 holds.

Proof. Assumption 1 shows that the discrete Cauchy problem is stable in ℓ^2, that is to say, there exists a numerical constant C such that

$$\sup_{n\geq 0}\sum_{j\in\mathbb{Z}}\Delta x\,|V_j^n|^2\leq C\sum_{n=0}^{s}\sum_{j\geq 1-r}\Delta x\,|f_j^n|^2\,.$$

Introducing the parameter $\gamma>0$, and summing with respect to $n\in\mathbb{N}$, we get

$$\frac{\gamma}{\gamma+1}\sum_{n\geq 0}\sum_{j\in\mathbb{Z}}\Delta x\,\mathrm{e}^{-2\,\gamma\,n}\,|V_j^n|^2\leq C\,\frac{\gamma}{(1-\mathrm{e}^{-2\,\gamma})\,(\gamma+1)}\sum_{n=0}^{s}\sum_{j\geq 1-r}\Delta x\,|f_j^n|^2\leq C\sum_{n=0}^{s}\sum_{j\geq 1-r}\Delta x\,|f_j^n|^2\,.$$

The substitution $\gamma\to\gamma\,\Delta t$ and the trace estimate (16) already yield:

$$\frac{\gamma}{\gamma\,\Delta t+1}\sum_{n\geq 0}\sum_{j\geq 1-r}\Delta t\,\Delta x\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|V_j^n|^2+\sum_{n\geq 0}\sum_{j=1-r}^{P}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|V_j^n|^2\leq C_P\sum_{n=0}^{s}\sum_{j\geq 1-r}\Delta x\,|f_j^n|^2\,.\quad(17)$$

The trace estimate (16) for (V_j^n) gives a bound for the boundary source term (g_j^n) in (15). Indeed, we have

$$|g_j^n|\leq C\sum_{\sigma=-1}^{s}\sum_{\ell=1-r}^{1+q}|V_\ell^{n-1-\sigma}|\,,$$

with a constant C that does not depend on j, n, nor on the sequence (V_j^n). Introducing the parameter $\gamma>0$, we thus obtain

$$\sum_{n\geq s+1}\sum_{j=1-r}^{0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|g_j^n|^2\leq C\sum_{n\geq 0}\sum_{j=1-r}^{1+q}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|V_j^n|^2\leq C\sum_{n=0}^{s}\sum_{j\geq 1-r}\Delta x\,|f_j^n|^2\,,$$

where we have used (16) again (with $P=1+q$). Since the scheme (2) is strongly stable and (14) starts with zero initial conditions, we can use the strong stability estimate and obtain

$$\begin{aligned}\frac{\gamma}{\gamma\,\Delta t+1}\sum_{n\geq 0}\sum_{j\geq 1-r}\Delta t\,\Delta x\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|W_j^n|^2+\sum_{n\geq 0}\sum_{j=1-r}^{p}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|W_j^n|^2&\\ \leq C\sum_{n\geq s+1}\sum_{j=1-r}^{0}\Delta t\,\mathrm{e}^{-2\,\gamma\,n\,\Delta t}\,|g_j^n|^2\leq C\sum_{n=0}^{s}\sum_{j\geq 1-r}\Delta x\,|f_j^n|^2\,.&\quad(18)\end{aligned}$$

The combination of both estimates (17) and (18) gives the conclusion of Theorem 1. □

Our goal now is to show that the trace estimate (16) is valid for the solution to the Cauchy problem (13). This is summarized in the following result.

Proposition 1. *Let Assumptions 1, 2, 3 and 4 be satisfied. Then for all $P\in\mathbb{N}$, there exists a constant $C_P>0$ such that for all $\gamma>0$, the solution $(V_j^n)_{j\in\mathbb{Z},n\in\mathbb{N}}$ to (13) satisfies*

$$\sum_{n\geq 0}\sum_{j=1-r}^{P}\mathrm{e}^{-2\,\gamma\,n}\,|V_j^n|^2\leq C_P\sum_{n=0}^{s}\sum_{j\geq 1-r}|f_j^n|^2\,.$$

Proposition 1 clearly implies the validity of (16) by passing to the limit $\gamma\to 0$, and therefore the validity of Theorem 1. We thus now focus on the proof of Proposition 1, for which we first recall some fundamental properties of the matrix $\mathbb{M}(z)$ in (9).

3.2 A brief reminder on the normal modes analysis

The main result of [Cou09] can be stated as follows.

Theorem 2 (Block reduction of $\mathbb{M}$). *Let Assumptions 1, 2, 3 and 4 be satisfied. Then for all $z \in \mathscr{U}$, the matrix $\mathbb{M}(z)$ in (9) has $N\,r$ eigenvalues, counted with their multiplicity, in $\mathbb{D} \setminus \{0\}$, and $N\,p$ eigenvalues, counted with their multiplicity, in $\mathscr{U}$. We let $\mathbb{E}^s(z)$, resp. $\mathbb{E}^u(z)$, denote the $N\,r$-dimensional, resp. $N\,p$-dimensional, generalized eigenspace associated with those eigenvalues that lie in $\mathbb{D} \setminus \{0\}$, resp. $\mathscr{U}$.*

Furthermore, 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:

$$\forall\, z \in \mathscr{O}\,, \quad T(z)^{-1}\,\mathbb{M}(z)\,T(z) = \begin{pmatrix} \mathbb{M}_1(z) & & 0 \\ & \ddots & \\ 0 & & \mathbb{M}_L(z) \end{pmatrix},$$

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

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

We refer to the blocks $\mathbb{M}_\ell$ as being of the first, second or third type.

Observe that Assumption 4 precludes the occurrence of blocks of the fourth type in the terminology of [Cou09], because such blocks only arise when glancing modes are present. In our framework, we shall only deal with elliptic blocks (first or second type) or scalar blocks. The latter correspond to eigenvalues of $\mathbb{M}(z)$ that depend holomorphically on z.

3.3 Proof of the trace estimate for the Cauchy problem

3.3.1 The resolvent equation

As already seen in the proof of Lemma 1, the solution (V_j^n) to the Cauchy problem (13) satisfies

$$\frac{\gamma}{\gamma+1} \sum_{n\geq 0} \sum_{j\in\mathbb{Z}} \Delta x\, \mathrm{e}^{-2\,\gamma\,n}\, |V_j^n|^2 \leq C \sum_{n=0}^{s} \sum_{j\geq 1-r} \Delta x\, |f_j^n|^2\,, \tag{19}$$

for all $\gamma > 0$. The estimate (19) shows that, for all $j \in \mathbb{Z}$, we can define the Laplace transform of the step function

$$V_j(t) := \begin{cases} 0 & \text{if } t < 0\,, \\ V_j^n & \text{if } t \in [n, n+1[\,,\ n \in \mathbb{N}\,. \end{cases}$$

The Laplace transform $\widehat{V}_j$ is holomorphic in the right half-plane $\{\mathrm{Re}\,\tau > 0\}$ for all $j \in \mathbb{Z}$, and Plancherel Theorem gives

$$\forall\, \gamma > 0\,, \quad \sum_{j\in\mathbb{Z}} \int_{\mathbb{R}} |\widehat{V}_j(\gamma + i\,\theta)|^2\, \mathrm{d}\theta < +\infty\,.$$

In particular, for all $\gamma > 0$, the sequence $\big(\widehat{V}_j(\gamma + i\,\theta)\big)_{j\in\mathbb{Z}}$ belongs to ℓ^2 for almost every $\theta \in \mathbb{R}$.

Applying the Laplace transform to (13) yields the resolvent equation on $\mathbb{Z}$:

$$\forall\, j \in \mathbb{Z}\,, \quad \widehat{V}_j(\tau) - \sum_{\sigma=0}^{s} z^{-\sigma-1}\, Q_\sigma\, \widehat{V}_j(\tau) = F_j(\tau)\,, \tag{20}$$

where the source term F_j is defined by

$$\forall\, j \in \mathbb{Z}\,, \quad F_j(\tau) := \frac{1-z^{-1}}{\tau} \left\{ \sum_{n=0}^{s} z^{-n}\, f_j^n - \sum_{\ell=-r}^{p} \sum_{\sigma=0}^{s-1} \sum_{n=0}^{s-\sigma-1} z^{-n-\sigma-1}\, A_{\ell,\sigma}\, f_{j+\ell}^n \right\}, \tag{21}$$

and it is understood, as always in what follows, that τ is a complex number of positive real part γ, and $z := \mathrm{e}^\tau \in \mathscr{U}$. In (21), we use the convention $f_j^n = 0$ if $j \leq -r$. Using the matrix $\mathbb{M}(z)$ that has been defined in (9), we can rewrite (20) as

$$\forall\, j \in \mathbb{Z}\,,\, \mathscr{W}_{j+1}(\tau) = \mathbb{M}(z)\, \mathscr{W}_j(\tau) + \mathscr{F}_j(\tau)\,, \quad \mathscr{W}_j(\tau) := \begin{pmatrix} \widehat{V}_{j+p-1}(\tau) \\ \vdots \\ \widehat{V}_{j-r}(\tau) \end{pmatrix}, \mathscr{F}_j(\tau) := \begin{pmatrix} \mathbb{A}_p(z)^{-1}\, F_j(\tau) \\ \\ 0 \end{pmatrix}. \tag{22}$$

Our goal now is to estimate the term $\mathscr{W}_{1-p-r}(\tau)$ of the solution $(\mathscr{W}_j)$ to (22), and then to estimate finitely many $\mathscr{W}_\nu(\tau)$, $\nu \geq 1-p-r$.

3.3.2 Estimates for γ small

In what follows, we always use the notation $\tau = \gamma + i\,\theta$, and we recall the notation $z := \mathrm{e}^\tau$. The source term $\mathscr{F}_j$ in (22) is given in terms of F_j, whose expression is given in (21). The initial data $(f_j^0), \ldots, (f_j^s)$ in (13) vanish for $j \leq -r$, and so therefore do F_j and $\mathscr{F}_j$ for $j \leq -p-r$ (and even for $j \leq -r$ if $s = 0$). This means that for all $j \leq -p-r$, the sequence $(\mathscr{W}_j)$ satisfies

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

and we know moreover that for all $\gamma > 0$, the sequence $(\mathscr{W}_j(\gamma + i\,\theta))_{j\in\mathbb{Z}}$ belongs to ℓ^2 for almost every $\theta \in \mathbb{R}$. Applying Theorem 2, this means that the vector $\mathscr{W}_{1-p-r}(\tau)$ belongs to $\mathbb{E}^u(z)$ for almost every $\theta \in \mathbb{R}$.

Let us introduce the projectors $\Pi^s(z), \Pi^u(z)$ associated with the decomposition

$$\mathbb{C}^{N\,(p+r)} = \mathbb{E}^s(z) \oplus \mathbb{E}^u(z)\,.$$

Using the exponential decay of $\mathbb{M}(z)^{-k}\, \Pi^u(z)$ as k tends to infinity, the induction relation (22) gives for almost every $\theta \in \mathbb{R}$:

$$\mathscr{W}_{1-p-r}(\tau) = \Pi^u(z)\, \mathscr{W}_{1-p-r}(\tau) = -\sum_{j\geq 0} \mathbb{M}(z)^{-1-j}\, \Pi^u(z)\, \mathscr{F}_{1-p-r+j}(\tau)\,. \tag{23}$$

We now focus on formula (23) and its consequences for small values of γ. More precisely, we consider a point $\underline{z}$ of the unit circle $\mathbb{S}^1$ and apply Theorem 2. Let us introduce neighborhoods of the form as depicted in Figure 1:

$$\forall\, \varepsilon > 0\,, \quad \mathscr{V}_{\underline{z},\varepsilon} := \Big\{ \underline{z}\, \mathrm{e}^{\alpha+i\,\beta}\,,\, \alpha, \beta \in]-\varepsilon, \varepsilon[\Big\}\,.$$

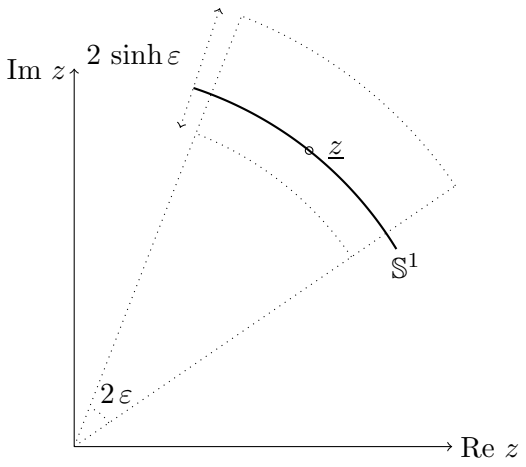

Figure 1: The neighborhood $\mathscr{V}_{\underline{z},\varepsilon}$.

According to Theorem 2, there exists some $\varepsilon > 0$ such that on $\mathscr{V}_{\underline{z},\varepsilon}$, there is a holomorphic change of basis $T(z)$ that block-diagonalizes $\mathbb{M}(z)$, with blocks satisfying one of the properties stated in Theorem 2. There is no loss of generality in assuming that blocks $\mathbb{M}_\ell(z)$ of the third type, which correspond to eigenvalues of $\mathbb{M}(z)$, can further be written as

$$\mathbb{M}_\ell(z) = \mathrm{e}^{\xi_\ell(z)}\,, \quad \xi_\ell(\underline{z}) \in i\,\mathbb{R}\,, \quad \underline{z}\,\xi_\ell'(\underline{z}) \in \mathbb{R} \setminus \{0\}\,, \tag{24}$$

where ξ_ℓ is holomorphic on $\mathscr{V}_{\underline{z},\varepsilon}$ and

$$\forall\, z \in \mathscr{V}_{\underline{z},\varepsilon}\,, \quad |\mathrm{Re}\,(z\,\xi_\ell'(z))| \geq \frac{1}{2}\,|\xi_\ell'(\underline{z})| > 0\,.$$

In particular, $|\xi_\ell'|$ is uniformly bounded from below by a positive constant on $\mathscr{V}_{\underline{z},\varepsilon}$. We can further assume that $T(z)$ and its inverse are uniformly bounded on $\mathscr{V}_{\underline{z},\varepsilon}$.

Remark 2. *Since $\mathbb{M}_\ell(z)$ is an eigenvalue of $\mathbb{M}(z)$, there holds $\xi_\ell(z) \notin i\,\mathbb{R}$ for $z \in \mathscr{V}_{\underline{z},\varepsilon} \cap \mathscr{U}$. More precisely, the $\xi_\ell(z)$'s of positive real part correspond to eigenvalues of $\mathbb{M}(z)$ in $\mathscr{U}$ (the unstable ones), and those of negative real part correspond to eigenvalues in $\mathbb{D}$ (the stable ones).*

Our goal is to derive a bound of the form

$$\int_{\mathscr{I}} |\mathscr{W}_{1-p-r}(\tau)|^2\,\mathrm{d}\theta \leq C \sum_{n=0}^{s} \sum_{j \geq 1-r} |f_j^n|^2\,, \tag{25}$$

uniformly with respect to $\gamma \in]0,\varepsilon[$, where $\mathscr{I}$ denotes the set[6]:

$$\mathscr{I} := \{\theta \in \mathbb{R} \,/\, \mathrm{e}^\tau \in \mathscr{V}_{\underline{z},\varepsilon}\} = \cup_{k\in\mathbb{Z}}\,]\underline{\theta} + 2\,k\,\pi - \varepsilon, \underline{\theta} + 2\,k\,\pi + \varepsilon[\,, \quad \underline{z} = \mathrm{e}^{i\,\underline{\theta}}\,. \tag{26}$$

[6]Observe that the form of the neighborhood $\mathscr{V}_{\underline{z},\varepsilon}$ implies that $\mathscr{I}$ is independent of γ, which is the reason for introducing such neighborhoods.

Let $\gamma \in]0,\varepsilon[$ be fixed. For almost every $\theta \in \mathscr{I}$, the vector $\mathscr{W}_{1-p-r}(\tau)$ is given by (23), and we can diagonalize $\mathbb{M}(z)$ with the matrix $T(z)$. In order to cover all possible cases[7], we assume that the block diagonalization of $\mathbb{M}(z)$ reads

$$T(z)^{-1}\,\mathbb{M}(z)\,T(z) = \mathrm{diag}\,(\mathbb{M}_\sharp(z), \mathbb{M}_\flat(z), \mathbb{M}_1^+(z), \dots, \mathbb{M}_{L^+}^+(z), \mathbb{M}_1^-(z), \dots, \mathbb{M}_{L^-}^-(z))\,,$$

where $\mathbb{M}_\sharp(z)$ is a block of the first type, $\mathbb{M}_\flat(z)$ is a block of the second type, and all other blocks are (scalars) of the third type with

$$\begin{aligned}
&\forall\, \ell = 1, \dots, L^+\,, \quad \mathbb{M}_\ell^+(\underline{z}) \in \mathbb{S}^1\,, \quad \overline{\mathbb{M}_\ell^+(\underline{z})}\,\underline{z}\,(\mathbb{M}_\ell^+)'(\underline{z}) \in \mathbb{R}_+^*\,,\\
&\forall\, \ell = 1, \dots, L^-\,, \quad \mathbb{M}_\ell^-(\underline{z}) \in \mathbb{S}^1\,, \quad \overline{\mathbb{M}_\ell^-(\underline{z})}\,\underline{z}\,(\mathbb{M}_\ell^-)'(\underline{z}) \in \mathbb{R}_-^*\,.
\end{aligned}$$

Then the generalized eigenspace $\mathbb{E}^u(z)$ is spanned by those column vectors of $T(z)$ which correspond to the blocks $\mathbb{M}_\sharp, \mathbb{M}_1^+, \dots, \mathbb{M}_{L^+}^+$, while the generalized eigenspace $\mathbb{E}^s(z)$ is spanned by those column vectors of $T(z)$ which correspond to the blocks $\mathbb{M}_\flat, \mathbb{M}_1^-, \dots, \mathbb{M}_{L^-}^-$, see, e.g., [Tre84, Lemma 3.3] or [Cou09]. An easy corollary of this "decoupling" property is that both projectors Π^s, Π^u extend holomorphically to $\mathscr{V}_{\underline{z},\varepsilon}$ and are bounded. We can even decompose $\Pi^u(z)$ as

$$\Pi^u(z) = \Pi_\sharp(z) + \sum_{\ell=1}^{L^+} \Pi_\ell^+(z)\,,$$

with self-explanatory notation. For almost every $\theta \in \mathscr{I}$, the formula (23) then reads

$$\Pi_\sharp(z)\,\mathscr{W}_{1-p-r}(\tau) = -\sum_{j\geq 0} \mathbb{M}(z)^{-1-j}\,\Pi_\sharp(z)\,\mathscr{F}_{1-p-r+j}(\tau)\,, \tag{27}$$

$$\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau) = -\sum_{j\geq 0} \mathrm{e}^{-(1+j)\,\xi_\ell^+(z)}\,\Pi_\ell^+(z)\,\mathscr{F}_{1-p-r+j}(\tau)\,. \tag{28}$$

The norm of $\mathbb{M}(z)^{-1-j}\,\Pi_\sharp(z)$ decays exponentially with j, uniformly with respect to $z \in \mathscr{V}_{\underline{z},\varepsilon}$, because $\mathbb{M}_\sharp$ is a block of the first type. Hence (27) implies, with a constant C that is uniform with respect to γ and $\theta \in \mathscr{I}$:

$$|\Pi_\sharp(z)\,\mathscr{W}_{1-p-r}(\tau)|^2 \leq C \sum_{j\geq 0} |\mathscr{F}_{1-p-r+j}(\tau)|^2\,.$$

We then use the definitions (22) and (21) to derive

$$|\Pi_\sharp(z)\,\mathscr{W}_{1-p-r}(\tau)|^2 \leq C\,\frac{|1-z^{-1}|^2}{|\tau|^2} \sum_{n=0}^{s} \sum_{j\geq 1-r} |f_j^n|^2\,.$$

We end up with the estimate of the elliptic part of $\mathscr{W}_{1-p-r}$:

$$\begin{aligned}
\int_{\mathscr{I}} |\Pi_\sharp(z)\,\mathscr{W}_{1-p-r}(\tau)|^2\,\mathrm{d}\theta &\leq C \int_{\mathbb{R}} \frac{|1-\mathrm{e}^{-\gamma-i\,\theta}|^2}{\gamma^2+\theta^2}\,\mathrm{d}\theta \sum_{n=0}^{s} \sum_{j\geq 1-r} |f_j^n|^2\\
&\leq C\,\frac{1-\mathrm{e}^{-2\,\gamma}}{\gamma} \sum_{n=0}^{s} \sum_{j\geq 1-r} |f_j^n|^2 \leq C \sum_{n=0}^{s} \sum_{j\geq 1-r} |f_j^n|^2\,.
\end{aligned} \tag{29}$$

[7] If one type of block is not present in the reduction close to $\underline{z}$, the proof of (25) simplifies accordingly.

We now turn to the hyperbolic components $\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau)$, whose analysis relies on arguments that are similar to those in [Sar65]. Since $\Pi_\ell^+(z)$ projects on a one-dimensional vector space, we can rewrite (28) as

$$\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau) = -\sum_{j\geq 0} \mathrm{e}^{-(1+j)\,\xi_\ell^+(z)}\,(L_\ell(z)\,\mathscr{F}_{1-p-r+j}(\tau))\,T_\ell(z)\,,$$

where L_ℓ is a row vector that depends holomorphically on z, and $T_\ell(z)$ is a column vector of $T(z)$. Using the expression (22) of $\mathscr{F}_{1-p-r+j}(\tau)$, we find that, up to multiplying by harmless bounded functions of z, $\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau)$ reads as a linear combination of the $s+1$ functions

$$\frac{1-z^{-1}}{\tau}\sum_{j\geq 0}\mathrm{e}^{-j\,\xi_\ell^+(z)}\,f^n_{1-r+j}\,,\quad n=0,\dots,s\,,$$

which coincide, up to multiplying by harmless bounded functions of z, with:

$$\frac{1-z^{-1}}{\tau}\,\mathbb{F}^n(\xi_\ell^+(z))\,,\quad n=0,\dots,s\,,$$

where $\mathbb{F}^n$ denotes the Laplace transform of the initial condition

$$f^n(x) := \begin{cases} f^n_{1-r+j}\,, & x\in[j,j+1[\,,\ j\in\mathbb{N}\,,\\ 0\,, & \text{otherwise.}\end{cases}$$

Recall that $\xi_\ell^+(z)$ has positive real part for $\gamma>0$, so the Laplace transform $\mathbb{F}^n$ is well-defined at $\xi_\ell^+(z)$. At this stage, the decomposition of $\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau)$ implies the uniform bound

$$\int_{\mathscr{I}} |\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau)|^2\,\mathrm{d}\theta \leq C\sum_{n=0}^{s}\int_{\mathscr{I}}\frac{|1-\mathrm{e}^{-\gamma-i\,\theta}|^2}{\gamma^2+\theta^2}\,|\mathbb{F}^n(\xi_\ell^+(z))|^2\,\mathrm{d}\theta\,. \tag{30}$$

We first simplify (30) by observing that θ enters the integrand on the right hand-side only through $\mathrm{e}^{i\,\theta}$ but at one place, which is the $1/(\gamma^2+\theta^2)$ factor. The form (26) of $\mathscr{I}$ and some straightforward changes of variable turn (30) into

$$\int_{\mathscr{I}} |\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau)|^2\,\mathrm{d}\theta \leq C\sum_{n=0}^{s}\int_{\underline{\theta}-\varepsilon}^{\underline{\theta}+\varepsilon}|\mathbb{F}^n(\xi_\ell^+(z))|^2\,\mathrm{d}\theta\,,$$

with a constant C that is still uniform with respect to γ. Because $|\xi_\ell'|$ is uniformly bounded away from zero on $\mathscr{V}_{\underline{z},\varepsilon}$, we obtain

$$\int_{\mathscr{I}} |\Pi_\ell^+(z)\,\mathscr{W}_{1-p-r}(\tau)|^2\,\mathrm{d}\theta \leq C\sum_{n=0}^{s}\int_{\underline{\theta}-\varepsilon}^{\underline{\theta}+\varepsilon}|\mathbb{F}^n(\xi_\ell^+(z))|^2\,|i\,z\,(\xi_\ell^+)'(z)|\,\mathrm{d}\theta = C\sum_{n=0}^{s}\int_{\mathscr{C}_{\ell,\gamma}}|\mathbb{F}^n(z)|^2\,|\mathrm{d}z|\,, \tag{31}$$

where $\mathscr{C}_{\ell,\gamma}$ denotes the (analytic) curve

$$\mathscr{C}_{\ell,\gamma} := \left\{\xi_\ell^+\left(\underline{z}\,\mathrm{e}^{\gamma+i\,\theta}\right)\,,\ \theta\in\,]-\varepsilon,\varepsilon[\,\right\}. \tag{32}$$

The argument now relies on Carlson's Lemma [Car43], which gives a bound for curvilinear integrals of Laplace transforms in terms of the L^2 norm of the original function. More precisely, there holds

$$\int_{\mathscr{C}_{\ell,\gamma}}|\mathbb{F}^n(z)|^2\,|\mathrm{d}z| \leq \frac{1}{\pi}\int_{i\,\mathbb{R}}|\mathbb{F}^n(w)|^2\,A_\ell(\gamma,w)\,|\mathrm{d}w|\,,$$

where $A_\ell(\gamma, w)$ denotes the total variation of the argument of $z - w$ as z runs through the curve $\mathscr{C}_{\ell,\gamma}$. In particular, if we can prove a uniform bound of the type

$$\sup_{\gamma \in]0,\varepsilon[} \sup_{w \in i\,\mathbb{R}} A_\ell(\gamma, w) < +\infty\,,$$

then we shall obtain from (31) and Carlson's Lemma the uniform bound

$$\int_{\mathscr{I}} |\Pi_\ell^+(z)\, \mathscr{W}_{1-p-r}(\tau)|^2 \,\mathrm{d}\theta \leq C \sum_{n=0}^{s} \sum_{j \geq 1-r} |f_j^n|^2\,, \tag{33}$$

and the combination of (29) and (33) will yield (25).

3.3.3 Bounding the total variation of the argument

The goal of this paragraph is to prove the following technical Lemma on families of analytic curves such as the $\mathscr{C}_{\ell,\gamma}$'s in (32). We give a complete proof of this fact since the details in [Sar65] are omitted and we consider an even more general situation than the corresponding one in [Sar65].

Lemma 2. *Let $\varepsilon > 0$, and let f be holomorphic on $]-\varepsilon, \varepsilon[^2 \subset \mathbb{C}$ with:*

- *$f(0) = 0$, $f'(0) \in \mathbb{R}_+^*$,*
- *for all $(\gamma, \theta) \in]0, \varepsilon[\times]-\varepsilon, \varepsilon[$, $f(\gamma + i\,\theta)$ has positive real part.*

For $w \in \mathbb{R}$ and $(\gamma, \theta) \in]0, \varepsilon[\times]-\varepsilon, \varepsilon[$, let $v(\gamma, \theta, w) \in]-\pi/2, \pi/2[$ denote the argument of $f(\gamma + i\,\theta) - i\,w$. Then, up to shrinking ε, there exists a constant $C > 0$ such that

$$\sup_{\gamma \in]0,\varepsilon[} \sup_{w \in \mathbb{R}} \int_{-\varepsilon}^{\varepsilon} |\partial_\theta v(\gamma, \theta, w)| \,\mathrm{d}\theta \leq C\,. \tag{34}$$

Proof. There are two cases (see a similar argument in [Cou11, Proposition 4.5]). Since f is holomorphic, then either $f(i\,\theta)$ is purely imaginary for all $\theta \in]-\varepsilon, \varepsilon[$, or there exists a smallest $k \in \mathbb{N}^*$ and a constant $c > 0$ such that

$$\mathrm{Re}\ f(i\,\theta) \geq c\,\theta^{2\,k}\,. \tag{35}$$

The proof of (34) is different in each of these two cases. (The analysis in [Sar65] only deals with the first case.)

Observe that we can always change ε for $\varepsilon/2$, so that we can assume that f together with any of its derivatives is bounded on the square $]-\varepsilon, \varepsilon[^2$.

• Case 1 : we assume that f is such that $f(i\,\theta)$ is purely imaginary for all $\theta \in]-\varepsilon, \varepsilon[$, which amounts to assuming $i^{n-1}\, f^{(n)}(0) \in \mathbb{R}$ for all $n \in \mathbb{N}$. Since v denotes the argument of $f(\gamma + i\,\theta) - i\,w$, there holds

$$\partial_\theta v(\gamma, \theta, w) = \mathrm{Re}\left(\frac{f'(\gamma + i\,\theta)}{f(\gamma + i\,\theta) - i\,w}\right)\,. \tag{36}$$

The function f is holomorphic and vanishes at 0, so there exists a constant $C_0 > 0$, which does not depend on ε, such that, up to choosing ε small enough, there holds

$$\sup_{(\gamma,\theta) \in\,]-\varepsilon,\varepsilon[^2} |f(\gamma + i\,\theta)| \leq C_0\,\varepsilon\,. \tag{37}$$

The constant C_0 is now fixed. If $|w| > 2\,C_0\,\varepsilon$ and $\gamma > 0$, then (36) yields

$$|\partial_\theta v(\gamma,\theta,w)| \leq \frac{1}{C_0\,\varepsilon} \sup_{(\gamma,\theta)\in\,]-\varepsilon,\varepsilon[^2} |f'(\gamma+i\,\theta)|\,,$$

and therefore

$$\sup_{\gamma\in\,]0,\varepsilon[} \sup_{|w|\geq 2C_0\,\varepsilon} \int_{-\varepsilon}^{\varepsilon} |\partial_\theta v(\gamma,\theta,w)|\,\mathrm{d}\theta \leq 2\,C_0 \sup_{(\gamma,\theta)\in\,]-\varepsilon,\varepsilon[^2} |f'(\gamma+i\,\theta)|\,.$$

It therefore only remains to study the case $|w| \leq 2\,C_0\,\varepsilon$, for which we are going to show that $\partial_\theta v$ is positive. The formula (36) shows that $\partial_\theta v$ has the same sign as

$$\mathrm{Re}\,\Big(f'(\gamma+i\,\theta)\,\big(\overline{f(\gamma+i\,\theta)}+i\,w\big)\Big)\,,$$

and from the assumption on f, we find that $\partial_\theta v$ has the same sign as

$$\mathrm{Re}\,\Big(f'(\gamma+i\,\theta)\,\big(\overline{f(\gamma+i\,\theta)}+i\,w\big) - f'(i\,\theta)\,\big(\overline{f(i\,\theta)}+i\,w\big)\Big)\,.$$

We rewrite the latter quantity as

$$\mathrm{Re}\,\Big(f'(i\,\theta)\,\big(\overline{f(\gamma+i\,\theta)}-\overline{f(i\,\theta)}\big)\Big) - w\,\mathrm{Im}(f'(\gamma+i\,\theta)-f'(i\,\theta)) + \mathrm{Re}\,\Big((f'(\gamma+i\,\theta)-f'(i\,\theta))\,\overline{f(\gamma+i\,\theta)}\Big)\,,$$

which, for ε sufficiently small, is bounded from below by (here we use $|w| \leq 2\,C_0\,\varepsilon$):

$$\frac{f'(0)^2}{2}\,\gamma - 3\,C_0\,\varepsilon\,\gamma \sup_{(\gamma,\theta)\in\,]-\varepsilon,\varepsilon[^2} |f''(\gamma+i\,\theta)|\,.$$

In particular, for $\varepsilon > 0$ sufficiently small, there holds $\partial_\theta v(\gamma,\theta,w) > 0$ for all $(\gamma,\theta) \in\,]0,\varepsilon[\times]-\varepsilon,\varepsilon[$ and $|w| \leq 2\,C_0\,\varepsilon$. This property yields

$$\sup_{\gamma\in\,]0,\varepsilon[} \sup_{|w|\leq 2C_0\,\varepsilon} \int_{-\varepsilon}^{\varepsilon} |\partial_\theta v(\gamma,\theta,w)|\,\mathrm{d}\theta \leq \pi\,,$$

and (34) holds.

• Case 2 : we now assume that f satisfies (35) for some minimal integer $k \in \mathbb{N}^*$, which amounts to assuming

$$\forall\, n = 0,\dots,2\,k-1\,,\quad i^{n-1}\,f^{(n)}(0) \in \mathbb{R}\,,\quad \text{and}\quad \mathrm{Re}\,((-1)^k\,f^{(2\,k)})(0)) > 0\,.$$

We can still assume that f satisfies (37) for some constant $C_0 > 0$, and therefore the same argument as in Case 1 gives a uniform bound for the total variation of v when $|w| \geq 2\,C_0\,\varepsilon$, for in that case, $i\,w$ lies at a uniformly positive distance $C_0\,\varepsilon$ from the curves

$$\mathscr{C}_\gamma := \big\{ f(\gamma+i\theta)\,,\ \theta \in\,]-\varepsilon,\varepsilon[\,\big\}\,.$$

Let us therefore consider from now on the case $|w| \leq 2\,C_0\,\varepsilon$. We can assume that f' does not vanish on $]-\varepsilon,\varepsilon[^2$, so the curve $\mathscr{C}_\gamma$ only consists of regular points. Hence its curvature equals, up to multiplying by a positive quantity:

$$K(\gamma,\theta) := \mathrm{Re}\,f'(\gamma+i\,\theta)\,\mathrm{Im}\,f''(\gamma+i\,\theta) - \mathrm{Re}\,f''(\gamma+i\,\theta)\,\mathrm{Im}\,f'(\gamma+i\,\theta) = \mathrm{Re}\,\big(f'(\gamma+i\,\theta)\,\overline{f''(\gamma+i\,\theta)}\big)\,.$$

Performing a Taylor expansion of f' and f'', we compute

$$K(0,\theta) = -\frac{f'(0)\,\mathrm{Re}\,((-1)^k\,f^{(2\,k)}(0))}{(2\,k-2)!}\,\theta^{2\,k-2} + O(\theta^{2\,k-1})\,.$$

Choosing ε small enough, this means that there exists positive constants c and C, that do not depend on ε, such that the curvature K satisfies

$$K(\gamma,\theta) \leq -c\,\theta^{2\,k-2} + C\,\gamma\,.$$

If $k = 1$, the curvature K is uniformly negative, and we can conclude that the family of curves $\mathscr{C}_\gamma$, $0 < \gamma < \varepsilon$, consists of arcs of convex closed curves in the right half-place $\{\mathrm{Re}\,\zeta > 0\}$. For $k = 1$, this shows that the total variation

$$\int_{-\varepsilon}^{\varepsilon} |\partial_\theta v(\gamma,\theta,w)|\,\mathrm{d}\theta\,,$$

is not larger than $2\,\pi$ and the bound (34) follows. In the case $k \geq 2$, we still have $K \leq 0$ as long as $|\theta| \geq (\gamma/C)^{1/(2\,k-2)}$ for some suitable constant C, which means that the two arcs

$$\Big\{ f(\gamma+i\theta)\,,\ \theta \in]-\varepsilon, \max(-\varepsilon, -(\gamma/C)^{1/(2\,k-2)})] \Big\}\,, \quad \Big\{ f(\gamma+i\theta)\,,\ \theta \in\ [\min(\varepsilon,(\gamma/C)^{1/(2\,k-2)}),\varepsilon[\ \Big\}\,,$$

are convex[8]. In particular, there holds

$$\int_{]-\varepsilon,\varepsilon[\setminus]-(\gamma/C)^{1/(2\,k-2)},(\gamma/C)^{1/(2\,k-2)}[} |\partial_\theta v(\gamma,\theta,w)|\,\mathrm{d}\theta \leq 4\,\pi\,. \tag{38}$$

We now consider the regime where θ is small, meaning $|\theta| \leq (\gamma/C)^{1/(2\,k-2)}$ with the same constant C as the one for which (38) holds. We are going to show that in this regime, the derivative $\partial_\theta v$ is positive. Using (36), this derivative has the same sign as

$$\mathrm{Re}\,\Big(f'(\gamma+i\,\theta)\,\big(\overline{f(\gamma+i\,\theta)} + i\,w\big)\Big)\,,$$

which, similarly to what we did in Case 1, we rewrite as[9]

$$\mathrm{Re}\,\Big(f'(i\,\theta)\,\big(\overline{f(\gamma+i\,\theta)} - \overline{f(i\,\theta)}\big)\Big) - w\,\mathrm{Im}(f'(\gamma+i\,\theta) - f'(i\,\theta)) + \mathrm{Re}\,\Big((f'(\gamma+i\,\theta) - f'(i\,\theta))\,\overline{f(\gamma+i\,\theta)}\Big)$$
$$+ \mathrm{Re}\,\Big(f'(i\,\theta)\,\big(\overline{f(\gamma+i\,\theta)} + i\,w\big)\Big)\,.$$

Using the same lower bounds as in Case 1, the latter quantity is lower bounded, for ε sufficiently small, by (here we use $|w| \leq 2\,C_0\,\varepsilon$):

$$\frac{f'(0)^2}{4}\,\gamma + \mathrm{Re}\,\Big(f'(i\,\theta)\,\big(\overline{f(\gamma+i\,\theta)} + i\,w\big)\Big)\,.$$

Performing a Taylor expansion for f and f', we have derived the following lower bound:

$$|f(\gamma+i\,\theta) - i\,w|^2\,\partial_\theta v \geq \frac{f'(0)^2}{4}\,\gamma - C\,\theta^{2\,k} - C\,\varepsilon\,|\theta|^{2\,k-1} \geq c\,\gamma - C'\,\varepsilon\,|\theta|^{2\,k-1}\,,$$

[8] By convex, we mean that these curves are arcs of closed convex curves.

[9] The property $\mathrm{Re}\,(f'(i\,\theta)\,(\overline{f(i\,\theta)} + i\,w)) = 0$ does not hold any longer, and this is the reason why some new terms arise comparing to Case 1.

for suitable constants $c, C' > 0$. In the regime $\theta^{2k-2} \leq \gamma/C$, C fixed as in (38), there holds $c\gamma - C'\varepsilon|\theta|^{2k-1} \geq c\gamma/2$ for ε small enough, and we have thus shown that $\partial_\theta v$ is positive. This gives the bound

$$\int_{]\min(-\varepsilon,-(\gamma/C)^{1/(2k-2)}),\max(\varepsilon,(\gamma/C)^{1/(2k-2)})[} |\partial_\theta v(\gamma,\theta,w)|\,\mathrm{d}\theta \leq 2\pi\,,$$

which we combine with (38) to derive

$$\sup_{\gamma\in]0,\varepsilon[}\ \sup_{|w|\leq 2C_0\varepsilon} \int_{-\varepsilon}^{\varepsilon} |\partial_\theta v(\gamma,\theta,w)|\,\mathrm{d}\theta \leq 6\pi\,.$$

This completes the proof of (34) in Case 2. □

The above proof of Lemma 2 crucially uses the holomorphy of f, which corresponds, in the block reduction of $\mathbb{M}$, to the fact that there is no glancing frequency. When glancing frequencies occur, some eigenvalues of $\mathbb{M}$ display algebraic singularities, see [GKS72], that are combined with some possible dissipative behavior. A complete classification was made in [Cou11]. The proof of the uniform BV bound (34) is much more intricate when f has an algebraic singularity at 0, and we have not managed to complete it so far in a general framework.

Let us now explain how Lemma 2 yields (33). We consider a family of curves $\mathscr{C}_{\ell,\gamma}$ in (32). We can rewrite $\mathscr{C}_{\ell,\gamma}$ as

$$\mathscr{C}_{\ell,\gamma} := \underbrace{\xi_\ell^+(\underline{z})}_{\in i\,\mathbb{R}} + \Big\{ f(\gamma + i\,\theta)\,,\ \theta \in]-\varepsilon,\varepsilon[\,\Big\}\,,$$

with

$$f(\gamma + i\,\theta) := \xi_\ell^+\big(\underline{z}\,\mathrm{e}^{\gamma+i\,\theta}\big) - \xi_\ell^+(\underline{z})\,.$$

The function f satisfies all the assumptions of Lemma 2, therefore, up to shrinking ε, we can assume that the argument of $z - i\,w$, as z runs through the curve $\mathscr{C}_{\ell,\gamma}$ and $w \in \mathbb{R}$, satisfies the uniform bound (34). Applying Carlson's Lemma, we have thus obtained (33).

3.3.4 Conclusion

We still consider a fixed $\underline{z} \in \mathbb{S}^1$. Then for some sufficiently small $\varepsilon > 0$, we have shown that, uniformly with respect to the parameter $\gamma \in]0,\varepsilon[$, the estimate (25) holds. For $\ell \geq 1-p-r$, we use the induction relation (22), and easily derive the uniform bound

$$|\mathscr{W}_{\ell+1}(\tau)|^2 \leq C_\ell\,|\mathscr{W}_{1-p-r}(\tau)|^2 + C_\ell \sum_{j=1-p-r}^{\ell} |F_j(\tau)|^2\,.$$

Using (25) and the definition (21), we obtain

$$\forall\,\ell \geq 1-p-r\,,\quad \int_{\mathscr{I}} |\mathscr{W}_\ell(\tau)|^2\,\mathrm{d}\theta \leq C_\ell \sum_{n=0}^{s} \sum_{j\geq 1-r} |f_j^n|^2\,,$$

uniformly with respect to $\gamma \in]0,\varepsilon[$. We have therefore proved that for all $P \in \mathbb{N}$, there exists a constant $C_P > 0$ such that

$$\sum_{j=1-r}^{P} \int_{\mathscr{I}} |\widehat{V}_j(\tau)|^2\,\mathrm{d}\theta \leq C_P \sum_{n=0}^{s} \sum_{j\geq 1-r} |f_j^n|^2\,.$$

We now use the compactness of $\mathbb{S}^1$ and cover it by finitely many neighborhoods $\mathscr{V}_{z_1,\varepsilon_1},\dots,\mathscr{V}_{z_K,\varepsilon_K}$ such that, for each $k=1,\dots,K$ and $P\in\mathbb{N}$, there exists a constant $C_{k,P}$ for which there holds

$$\forall\,\gamma\in\,]0,\varepsilon_k[\,,\quad \sum_{j=1-r}^{P}\int_{\mathscr{I}_k}|\widehat{V}_j(\tau)|^2\,\mathrm{d}\theta\le C_{k,P}\sum_{n=0}^{s}\sum_{j\ge 1-r}|f_j^n|^2\,, \tag{39}$$

with the obvious notation

$$\mathscr{I}_k:=\{\theta\in\mathbb{R}\,/\,\mathrm{e}^{\tau}\in\mathscr{V}_{z_k,\varepsilon_k}\}\,.$$

The sets $\mathscr{I}_1,\dots,\mathscr{I}_K$ cover $\mathbb{R}$, so adding the estimates (39) gives

$$\sum_{j=1-r}^{P}\int_{\mathbb{R}}|\widehat{V}_j(\tau)|^2\,\mathrm{d}\theta\le C_P\sum_{n=0}^{s}\sum_{j\ge 1-r}|f_j^n|^2\,,$$

for $0<\gamma<\min\varepsilon_k$, and some suitable constant $C_P>0$. Applying Plancherel Theorem, we get

$$\sum_{n\in\mathbb{N}}\sum_{j=1-r}^{P}\mathrm{e}^{-2\,\gamma\,n}\,|V_j^n|^2\le C_P\sum_{n=0}^{s}\sum_{j\ge 1-r}|f_j^n|^2\,,$$

for $\gamma\in\,]0,\min\varepsilon_k[$. This proves Proposition 1 for sufficiently small values of γ.

The case where γ is not close to zero[10] is much easier for in that case, we already have the estimate (19), and an obvious lower bound then gives

$$\frac{\min\varepsilon_k}{1+\min\varepsilon_k}\sum_{n\in\mathbb{N}}\sum_{j=1-r}^{P}\mathrm{e}^{-2\,\gamma\,n}\,|V_j^n|^2\le C\sum_{n=0}^{s}\sum_{j\ge 1-r}|f_j^n|^2\,,$$

for $\gamma\ge\min\varepsilon_k$. The proof of Proposition 1, and ultimately of Theorem 1, is thus complete.

4 Proof of Corollary 1. The uniform power boundedness conjecture for schemes with two time levels

4.1 The discrete Leibniz formula and integration by parts

In this paragraph, we recall the discrete version of Leibniz formula and its consequence for integrating by parts. We recall that given $\nu\in\mathbb{Z}$ and a sequence $v=(v_j)_{j\ge\nu}$, $\mathbf{T}v$ denotes the sequence defined by $(\mathbf{T}v)_j:=v_{j+1}$ for all $j\ge\nu-1$, and $\mathbf{T}^{-1}v$ denotes the sequence defined by $(\mathbf{T}^{-1}v)_j:=v_{j-1}$ for all $j\ge\nu+1$. (Of course, v may also be indexed by all $\mathbb{Z}$.) Powers of $\mathbf{T}$ and $\mathbf{T}^{-1}$ are defined similarly. We let $\mathbf{D}$ denote the operator $\mathbf{T}-I$, where I is the identity. The operator $\mathbf{D}$ represents a discrete derivative[11].

The following result is a discrete version of the Leibniz rule.

[10] It should be understood that we use the scaling $\Delta t/\Delta x=\mathrm{Cst}$ and therefore only deal with one single parameter γ but γ is in fact a placeholder for $\gamma\,\Delta t$, so the regime "γ small" can be thought of as that of the continuous limit $\Delta t\to 0$ with a fixed γ. It is then rather obvious that in that case, the trace estimate of Proposition 1 can not be proved by just isolating the trace terms in (19), which corresponds to passing from an $L^2_{t,x}$ estimate to an L^2 estimate at $x=0$.

[11] Our notation $\mathbf{D}$ corresponds to D_+ in [GKO95], but we omit the $+$ sign since we shall never use the other discrete derivative $D_-=I-\mathbf{T}^{-1}$, nor the centered derivative $D_0=(\mathbf{T}-\mathbf{T}^{-1})/2$.

Lemma 3 (Discrete Leibniz formula)**.** *Let u, v be two sequences with values in $\mathbb{C}^N$ and indexed either by $j \geq \nu$ for some $\nu \in \mathbb{Z}$, or by all $\mathbb{Z}$. Then for all $k \in \mathbb{N}$, there holds*

$$\mathbf{D}^k\,(u^*\,v) = \sum_{\substack{j_1,j_2=0,\\ j_1+j_2\geq k}}^{k} \frac{k!}{(k-j_1)!\,(k-j_2)!\,(j_1+j_2-k)!}\,(\mathbf{D}^{j_1}u)^*\,\mathbf{D}^{j_2}v\,.$$

Proof. One starts with the formula

$$\mathbf{D}^k\,(u^*\,v) = \sum_{j=0}^{k} \frac{k!}{(k-j)!\,j!}\,(\mathbf{D}^j u)^*\,\mathbf{T}^j\mathbf{D}^{k-j}v\,,$$

which is obtained by a straightforward induction argument, and then use the binomial identity

$$\forall\, j \in \mathbb{N}\,, \quad \mathbf{T}^j = \sum_{\ell=0}^{j} \frac{j!}{(j-\ell)!\,\ell!}\,\mathbf{D}^\ell\,.$$

□

The first consequence of Lemma 3 is the following integration by parts formula, which mimics the analogous one for the product $u^*\,A\,u^{(k)}$, when u is a k-times differentiable function and A a hermitian matrix. Corollary 2 below is a generalization of [GKO95, Lemma 11.1.1].

Corollary 2. *Let $A \in \mathscr{M}_N(\mathbb{C})$ be hermitian and nonzero, and let $k \in \mathbb{N}^*$. Then there exists a unique hermitian form $q_{A,k}$ on $\mathbb{C}^{N\,k}$, and a unique collection of real numbers $\alpha_{1,k}, \dots, \alpha_{k,k}$ that only depend on k and not on A, such that for all sequence u with values in $\mathbb{C}^N$, there holds*

$$2\,\mathrm{Re}\,(u^*\,A\,\mathbf{D}^k u) = \mathbf{D}\left(q_{A,k}(u, \dots, \mathbf{D}^{k-1}u)\right) + \sum_{j=1}^{k} \alpha_{j,k}\,(\mathbf{D}^j u)^*\,A\,\mathbf{D}^j u\,. \tag{40}$$

Proof. Let us first prove the existence of the decomposition (40), which is done by induction. For $k = 1$, one just uses Lemma 3 and the fact that A is hermitian to obtain

$$2\,\mathrm{Re}\,(u^*\,A\,\mathbf{D}u) = \frac{1}{2}\,\mathbf{D}\left(u^*\,A\,u\right) - \frac{1}{2}\,(\mathbf{D}u)^*\,A\,\mathbf{D}u\,.$$

Let us therefore assume that the existence of the decomposition (40) holds up to some integer k. We use Lemma 3 and the fact that A is hermitian to obtain

$$\begin{aligned} 2\,\mathrm{Re}\,(u^*\,A\,\mathbf{D}^{k+1}u) = \mathbf{D}\left(\mathbf{D}^k(u^*\,A\,u)\right) &+ \sum_{\substack{j=1,\\ 2\,j\geq k+1}}^{k+1} \star\,(\mathbf{D}^j u)^*\,A\,\mathbf{D}^j u \\ &+ \sum_{\substack{j_1,j_2=1,\\ j_1+j_2\geq k+1, j_1<j_2}}^{k+1} \star\,\mathrm{Re}\,((\mathbf{D}^{j_1}u)^*\,A\,\mathbf{D}^{j_2}u)\,, \end{aligned}$$

where the $\star$ symbols represent harmless (real) numerical coefficients. Lemma 3 shows that the term $\mathbf{D}^k(u^*\,A\,u)$ can be written as a hermitian form of $(u,\dots,\mathbf{D}^k u)$. The terms $\mathrm{Re}\,((\mathbf{D}^{j_1}u)^*\,A\,\mathbf{D}^{j_2}u)$, $j_1<j_2$, are simplified by using the induction assumption:

$$\mathrm{Re}\,((\mathbf{D}^{j_1}u)^*\,A\,\mathbf{D}^{j_2}u)=\frac{1}{2}\,\mathbf{D}\left(q_{A,j_2-j_1}(\mathbf{D}^{j_1}u,\dots,\mathbf{D}^{j_2-1}u)\right)+\frac{1}{2}\sum_{j=1}^{j_2-j_1}\alpha_{j,j_2-j_1}\,(\mathbf{D}^{j_1+j}u)^*\,A\,\mathbf{D}^{j_1+j}u\,.$$

The existence of the decomposition (40) up to the integer $k+1$ follows.

Let us now prove that the decomposition (40) is unique. If two such decompositions exist, this means that we can find a hermitian form q (which may depend on A), and a collection of real numbers $\alpha_{1,k},\dots,\alpha_{k,k}$ (which do not depend on A) such that for all sequences u with values in $\mathbb{C}^N$, there holds

$$\mathbf{D}\left(q(u,\dots,\mathbf{D}^{k-1}u)\right)+\sum_{j=1}^{k}\alpha_{j,k}\,(\mathbf{D}^j u)^*\,A\,\mathbf{D}^j u=0\,. \tag{41}$$

Given arbitrary vectors $x_0,\dots,x_k\in\mathbb{C}^N$, we can find a sequence u with values in $\mathbb{C}^N$ and indexed by $\mathbb{N}$, such that

$$\forall\, j=0,\dots,k\,,\quad (\mathbf{D}^j u)_0=x_j\,.$$

Equation (41) evaluated at the index $\ell=0$ gives

$$q(u_1,\dots,(\mathbf{D}^{k-1}u)_1)-q(x_0,\dots,x_{k-1})+\sum_{j=1}^{k}\alpha_{j,k}\,x_j^*\,A\,x_j=0\,,$$

that is to say

$$q(x_1+x_0,\dots,x_k+x_{k-1})-q(x_0,\dots,x_{k-1})+\sum_{j=1}^{k}\alpha_{j,k}\,x_j^*\,A\,x_j=0\,.$$

Let Q denote the hermitian matrix associated with q, which therefore satisfies

$$X_1^*\,Q\,X_1+2\,\mathrm{Re}\,(X_0^*\,Q\,X_1)+\sum_{j=1}^{k}\alpha_{j,k}\,x_j^*\,A\,x_j=0\,,\quad X_0:=\begin{pmatrix}x_0\\ \vdots\\ x_{k-1}\end{pmatrix},\ X_1:=\begin{pmatrix}x_1\\ \vdots\\ x_k\end{pmatrix}. \tag{42}$$

The vector x_0 enters Equation (42) only through the term $\mathrm{Re}\,(X_0^*\,Q\,X_1)$. Since $x_0,\dots,x_k$ are arbitrary in $\mathbb{C}^N$, the block decomposition of Q:

$$Q=\begin{pmatrix}Q_{0,0} & \cdots & Q_{0,k-1}\\ \vdots & & \vdots\\ Q_{k-1,0} & \cdots & Q_{k-1,k-1}\end{pmatrix},$$

necessarily satisfies $Q_{0,0}=\cdots=Q_{0,k-1}=0$. Since Q is hermitian, this implies of course $Q_{0,0}=\cdots=Q_{k-1,0}=0$. In other words, the hermitian form q only depends on its $k-1$ last arguments, which reduces (42), with obvious notation, to

$$Y_1^*\,\widetilde{Q}\,Y_1+2\,\mathrm{Re}\,(Y_0^*\,\widetilde{Q}\,Y_1)+\sum_{j=1}^{k}\alpha_{j,k}\,x_j^*\,A\,x_j=0\,,\quad Y_0:=\begin{pmatrix}x_1\\ \vdots\\ x_{k-1}\end{pmatrix},\ Y_1:=\begin{pmatrix}x_2\\ \vdots\\ x_k\end{pmatrix}.$$

Choosing $Y_1 = 0$ in the latter relation gives $\alpha_{1,k}\, x_1^*\, A\, x_1 = 0$, which means that $\alpha_{1,k}$ equals zero (here we use the fact that A is nonzero), and uniqueness of the decomposition (40) follows by induction on k. □

Let us observe that in Corollary 2, if A is a real symmetric matrix, then the corresponding $q_{A,k}$ is a real quadratic form on $\mathbb{R}^{N\,k}$. The proof of Corollary 1 requires an extension of Corollary 2 to the case where A is real and skew-symmetric, which we state now.

Corollary 3. *Let $A \in \mathscr{M}_N(\mathbb{R})$ be skew-symmetric and nonzero, and let $k \in \mathbb{N}$, $k \geq 2$. Then there exists a unique quadratic form $q_{A,k}$ on $\mathbb{R}^{N\,k}$, and a unique collection of real numbers $\beta_{1,k}, \dots, \beta_{k-1,k}$ that only depend on k and not on A, such that for all sequence u with values in $\mathbb{R}^N$, there holds*

$$u^*\, A\, \mathbf{D}^k u = \mathbf{D}\left(q_{A,k}(u, \dots, \mathbf{D}^{k-1}u)\right) + \sum_{j=1}^{k-1} \beta_{j,k}\, (\mathbf{D}^j u)^*\, A\, \mathbf{D}^{j+1} u\,. \tag{43}$$

Proof. The proof follows closely that of Corollary 3. We briefly indicate the induction argument for the existence of the decomposition (43). For $k = 2$, we use Lemma 3 and the fact that A is skew-symmetric to obtain

$$u^*\, A\, \mathbf{D}^2 u = \mathbf{D}\left(u^*\, A\, \mathbf{D}u\right) - (\mathbf{D}u)^*\, A\, \mathbf{D}^2 u\,.$$

Since u is real valued, the term $u^*\, A\, \mathbf{D}u$ coincides with $q(u, \mathbf{D}u)$, where the matrix of the quadratic form q is

$$\frac{1}{2}\begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix}.$$

If the existence of the decomposition (43) holds up to some integer k, then Lemma 3 gives

$$u^*\, A\, \mathbf{D}^{k+1} u = \mathbf{D}\left(u^*\, A\, \mathbf{D}^k u\right) - (\mathbf{D}u)^*\, A\, \mathbf{D}^k u - (\mathbf{D}u)^*\, A\, \mathbf{D}^{k+1} u\,.$$

We apply the induction assumption for decomposing the term $(\mathbf{D}u)^*\, A\, \mathbf{D}^{k+1}u$. There are two cases for the remaining term $(\mathbf{D}u)^*\, A\, \mathbf{D}^k u$. Either $k = 2$, and this term is already in an irreducible form, or $k \geq 3$, and we can apply the induction assumption, which eventually yields the decomposition (43) up to $k + 1$.

Uniqueness of the decomposition (43) relies on more or less the same arguments as those used in the proof of Corollary 2. More precisely, assuming that two decompositions (43) exist, we can find a quadratic form q on $\mathbb{R}^{N\,k}$, with a corresponding real symmetric matrix Q, and a collection of real numbers $\beta_{1,k}, \dots, \beta_{k-1,k}$ that satisfy[12]

$$X_1^*\, Q\, X_1 + 2\, X_0^*\, Q\, X_1 + \sum_{j=1}^{k-1} \beta_{j,k}\, x_j^*\, A\, x_{j+1} = 0\,, \quad X_0 := \begin{pmatrix} x_0 \\ \vdots \\ x_{k-1} \end{pmatrix}, \; X_1 := \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}.$$

Identifying the x_0 term shows, as in the proof of Corollary 2, that the block decomposition of Q reads

$$Q = \begin{pmatrix} 0 & \dots & \dots & 0 \\ \vdots & & & \\ \vdots & & \widetilde{Q} & \\ 0 & & & \end{pmatrix},$$

[12] Here the vectors $x_0, \dots, x_k$ are real.

which means that the following relation holds for all vectors $x_1, \dots, x_k \in \mathbb{R}^N$:

$$Y_1^* \, \widetilde{Q} \, Y_1 + 2 \, Y_0^* \, \widetilde{Q} \, Y_1 + \sum_{j=1}^{k-1} \beta_{j,k} \, x_j^* \, A \, x_{j+1} = 0 \,, \quad Y_0 := \begin{pmatrix} x_1 \\ \vdots \\ x_{k-1} \end{pmatrix}, \; Y_1 := \begin{pmatrix} x_2 \\ \vdots \\ x_k \end{pmatrix} .$$

Here the proof differs slightly from that of Corollary 2 since there is no quadratic term with respect to x_1. Instead, we identify the quadratic terms with respect to (x_1, x_2), which amounts to taking first a partial derivative with respect to x_2 and then a partial derivative with respect to x_1. This yields

$$2 \, \widetilde{Q}_{1,1} + \beta_{1,k} \, A = 0 \,,$$

where $\widetilde{Q}_{1,1}$ denotes the upper left block of $\widetilde{Q}$ in its block decomposition. Observe now that $\widetilde{Q}$ is symmetric, and therefore so is $\widetilde{Q}_{1,1}$, while A is skew-symmetric and $\beta_{1,k}$ is real. Hence $\beta_{1,k}$ is zero and uniqueness of the decomposition (43) follows by induction. □

4.2 Consequences for Cauchy problems

In this paragraph, we explain some consequences of Corollaries 2 and 3 for showing stability of finite difference discretizations of Cauchy problems. We consider a numerical discretization with two time levels, that is:

$$\begin{cases} U_j^{n+1} = Q \, U_j^n \,, & j \in \mathbb{Z} \,, \quad n \geq 0 \,, \\ U_j^0 = f_j \,, & j \in \mathbb{Z} \,, \end{cases} \tag{44}$$

with $(f_j)_{j \in \mathbb{Z}} \in \ell^2$, and

$$Q = \sum_{\ell=-r}^{p} A_\ell \, \mathbf{T}^\ell \,, \quad \sum_{\ell=-r}^{p} A_\ell = I \,.$$

The latter consistency assumption allows us to express the finite difference operator Q as a sum of discrete derivatives. Namely, we write

$$\mathbf{T}^r \, (Q - I) = \sum_{\substack{\ell=-r, \\ \ell \neq 0}}^{p} A_\ell \, (\mathbf{T}^{r+\ell} - \mathbf{T}^r) \,,$$

and then decompose each $\mathbf{T}^{r+\ell} - \mathbf{T}^r$ as a linear combination of $\mathbf{D}, \dots, \mathbf{D}^{p+r}$ (which amounts to decomposing the polynomial $X^{r+\ell} - X^r$ on the family $(X-1, \dots, (X-1)^{p+r})$ which forms a basis of the space of real polynomials that vanish at 1 and whose degree is not larger than $p+r$). Summing up, the operator Q can be written equivalently as

$$Q = I + \mathbf{T}^{-r} \sum_{\ell=1}^{p+r} \widetilde{A}_\ell \, \mathbf{D}^\ell \,, \tag{45}$$

for suitable matrices $\widetilde{A}_1, \dots, \widetilde{A}_{p+r}$ whose expression can be obtained from $A_{-r}, \dots, A_p$. It is rather clear that all matrices $\widetilde{A}_1, \dots, \widetilde{A}_{p+r}$ are real, and they are symmetric if $A_{-r}, \dots, A_p$ are symmetric (which we shall not assume, but this might simplify some of the calculations below in some given situation).

The following Lemma is a direct consequence of Corollaries 2 and 3.

Lemma 4. *There exists a quadratic form q on $\mathbb{R}^{N\,(p+r)}$, some real symmetric matrices $S_1,\dots,S_{p+r}$ and some real skew-symmetric matrices $\widetilde{S}_1,\dots,\widetilde{S}_{p+r-1}$ such that for all sequence U with values in $\mathbb{R}^N$, there holds*

$$2\,U^*\,(Q-I)\,U+|(Q-I)\,U|^2=\mathbf{T}^{-r}\,\mathbf{D}\left(q(U,\dots,\mathbf{D}^{p+r-1}U)\right)+\mathbf{T}^{-r}\sum_{\ell=1}^{p+r}(\mathbf{D}^\ell U)^*\,S_\ell\,\mathbf{D}^\ell U \\ +\mathbf{T}^{-r}\sum_{\ell=1}^{p+r-1}(\mathbf{D}^\ell U)^*\,\widetilde{S}_\ell\,\mathbf{D}^{\ell+1}U\,. \tag{46}$$

If the sequence U is indexed by $j\ge 1-r$, then (46) is valid for all indeces $j\ge 1$, while if the sequence U is indexed by $\mathbb{Z}$, then (46) is valid on all $\mathbb{Z}$.

In particular, the solution (U_j^n) to (44) satisfies

$$\forall\, n\in\mathbb{N}\,,\quad \sum_{j\in\mathbb{Z}}|U_j^{n+1}|^2-\sum_{j\in\mathbb{Z}}|U_j^n|^2=\sum_{j\in\mathbb{Z}}\sum_{\ell=1}^{p+r}(\mathbf{D}^\ell U_j^n)^*\,S_\ell\,\mathbf{D}^\ell U_j^n+\sum_{j\in\mathbb{Z}}\sum_{\ell=1}^{p+r-1}(\mathbf{D}^\ell U_j^n)^*\,\widetilde{S}_\ell\,\mathbf{D}^{\ell+1}U_j^n\,. \tag{47}$$

The decomposition (46) is unique provided that Q is not the identity operator.

Proof. The existence of the decomposition (46) is indeed an easy consequence of Corollaries 2 and 3. Due to (45), the term $U^*\,(Q-I)\,U$ is a sum of terms of the form

$$U^*\,(\mathbf{T}^{-r}\,\widetilde{A}_\ell\,\mathbf{D}^\ell U)=\mathbf{T}^{-r}\left((\mathbf{T}^r U)^*\,\widetilde{A}_\ell\,\mathbf{D}^\ell U\right),$$

which can be written as a (real) linear combination of terms of the form $(\mathbf{D}^{\ell_1}U)^*\,\widetilde{A}_{\ell_2}\,\mathbf{D}^{\ell_2}U$ by simply expanding $\mathbf{T}^r$ as a linear combination of $I,\dots,\mathbf{D}^r$ (which is nothing but the binomial identity). We then split $\widetilde{A}_{\ell_2}$ as the sum of its symmetric and skew-symmetric parts and apply Corollaries 2 and 3 (if $\ell_1=\ell_2$, nothing needs to be done). The term $|(Q-I)\,U|^2$ can also be written under the form on the right hand-side of (46) since it is a sum of terms of the form

$$\mathbf{T}^{-r}\left(\left(\widetilde{A}_{\ell_1}\,\mathbf{D}^{\ell_1}U\right)^*\,\widetilde{A}_{\ell_2}\,\mathbf{D}^{\ell_2}U\right)=\mathbf{T}^{-r}\left((\mathbf{D}^{\ell_1}U)^*\left(\widetilde{A}_{\ell_1}^*\,\widetilde{A}_{\ell_2}\right)\mathbf{D}^{\ell_2}U\right),$$

and it only remains to split $\widetilde{A}_{\ell_1}^*\,\widetilde{A}_{\ell_2}$ as the sum of its symmetric and anti-symmetric parts and to apply Corollaries 2 and 3 (if $\ell_1=\ell_2$, nothing needs to be done).

The energy balance (47) follows by observing that the sum on $\mathbb{Z}$ of the discrete derivative $\mathbf{D}\,q$ vanishes. The remaining terms incorporate the (possible) dissipative behavior of the discretization. □

As a concrete example, let us explain Lemma 4 for three points schemes and scalar equations. In that case, $N=1$ so that there is no skew-symmetric matrix except 0, and the scheme reads

$$U_j^{n+1}=a_{-1}\,U_{j-1}^n+a_0\,U_j^n+a_1\,U_{j+1}^n\,,$$

with a triple of real numbers a_{-1},a_0,a_1 that satisfies $a_{-1}+a_0+a_1=1$. In that case, the integration by parts procedure leads to the relation

$$2\,U_j^n\,(Q-I)\,U_j^n+|(Q-I)\,U_j^n|^2=\mathbf{T}^{-1}\,\mathbf{D}\left((a_1-a_{-1})\,|U_j^n|^2+2\,a_1\,U_j^n\,\mathbf{D}\,U_j^n+a_1\,(a_1-a_{-1})\,|\mathbf{D}\,U_j^n|^2\right) \\ +\mathbf{T}^{-1}\left(d_1\,|\mathbf{D}\,U_j^n|^2+d_2\,|\mathbf{D}^2\,U_j^n|^2\right),$$

with

$$d_1 := -\frac{a_1 + a_{-1}}{2} + (a_1 - a_{-1})^2\,, \quad d_2 := a_1\, a_{-1}\,.$$

In that case, stability for the Cauchy problem, that is fulfillment of Assumption 1, is equivalent to the property

$$\max\left(d_1, d_1 + 4\, d_2\right) \leq 0\,,$$

or, in other words,

$$\max\left((a_1 - a_{-1})^2, (a_1 + a_{-1})^2\right) \leq \frac{1 - a_0}{2}\,.$$

For the upwind, Lax-Friedrichs and Lax-Wendroff schemes, one gets the standard CFL condition $\lambda\,|a| \leq 1$, with a the velocity of the transport equation one is willing to approximate.

4.3 Proof of Corollary 1

We consider the numerical scheme (2) with $s = 0$, zero interior source term and zero boundary source term. Writing Q instead of Q_0 for simplicity, the scheme reads

$$\begin{cases} U_j^{n+1} = Q\, U_j^n\,, & j \geq 1\,, \quad n \geq 0\,, \\ U_j^{n+1} = B_{j,-1}\, U_1^{n+1} + B_{j,0}\, U_1^n\,, & j = 1 - r, \dots, 0\,, \quad n \geq 0\,, \\ U_j^0 = f_j\,, & j \geq 1 - r\,, \end{cases} \tag{48}$$

with $(f_j)_{j \geq 1-r} \in \ell^2$, and

$$Q = \sum_{\ell=-r}^{p} A_\ell\, \mathbf{T}^\ell\,, \quad \sum_{\ell=-r}^{p} A_\ell = I\,.$$

We use the decomposition (46) of Q. The solution[13] (U_j^n) to (48) satisfies

$$\begin{aligned} \forall\, j \geq 1\,, \quad |U_j^{n+1}|^2 - |U_j^n|^2 = 2\,(U_j^n)^*\,(Q - I)\,U_j^n + |(Q - I)\,U_j^n|^2 = \mathbf{T}^{-r}\,\mathbf{D}\left(q(U_j^n, \dots, \mathbf{D}^{p+r-1} U_j^n)\right) \\ + \mathbf{T}^{-r} \sum_{\ell=1}^{p+r} (\mathbf{D}^\ell U_j^n)^*\, S_\ell\, \mathbf{D}^\ell U_j^n + \mathbf{T}^{-r} \sum_{\ell=1}^{p+r-1} (\mathbf{D}^\ell U_j^n)^*\, \widetilde{S}_\ell\, \mathbf{D}^{\ell+1} U_j^n\,. \end{aligned}$$

Summing with respect to $j \geq 1$, we get

$$\begin{aligned} \sum_{j \geq 1} |U_j^{n+1}|^2 - \sum_{j \geq 1} |U_j^n|^2 = &- q(U_{1-r}^n, \dots, \mathbf{D}^{p+r-1} U_{1-r}^n) \\ &+ \sum_{j \geq 1-r} \sum_{\ell=1}^{p+r} (\mathbf{D}^\ell U_j^n)^*\, S_\ell\, \mathbf{D}^\ell U_j^n + \sum_{j \geq 1-r} \sum_{\ell=1}^{p+r-1} (\mathbf{D}^\ell U_j^n)^*\, \widetilde{S}_\ell\, \mathbf{D}^{\ell+1} U_j^n\,, \end{aligned} \tag{49}$$

where, comparing with Lemma 4, the novelty is the "boundary" term $q(U_{1-r}^n, \dots, \mathbf{D}^{p+r-1} U_{1-r}^n)$.

Our goal now is to estimate the terms which appear in the second line of (49). Following an argument already used in [Wu95, CG11], we extend the sequence $(U_j^n)_{j \geq 1-r}$ by zero for $j \leq -r$, and still denote it

[13] It is assumed here that the initial condition consists of real vectors, so that the solution to (48) is real. The extension to complex sequences is straightforward because the scheme is linear with real coefficients.

(U_j^n). This extended sequence belongs to $\ell^2(\mathbb{Z})$, and we can therefore use the assumption of Corollary 1 on the action of Q on $\ell^2(\mathbb{Z})$. We obtain

$$\sum_{j\in\mathbb{Z}} 2\,(U_j^n)^*\,(Q-I)\,U_j^n + |(Q-I)\,U_j^n|^2 \le 0\,,$$

which, using the decomposition (45) and the fact that U_j^n vanishes for $j \le -r$, gives

$$\sum_{j\ge 1-r}\sum_{\ell=1}^{p+r}(\mathbf{D}^\ell U_j^n)^*\,S_\ell\,\mathbf{D}^\ell U_j^n + \sum_{j\ge 1-r}\sum_{\ell=1}^{p+r-1}(\mathbf{D}^\ell U_j^n)^*\,\widetilde{S}_\ell\,\mathbf{D}^{\ell+1}U_j^n$$
$$\le -\sum_{j=1-p-2r}^{-r}\sum_{\ell=1}^{p+r}(\mathbf{D}^\ell U_j^n)^*\,S_\ell\,\mathbf{D}^\ell U_j^n - \sum_{j=1-p-2r}^{-r}\sum_{\ell=1}^{p+r-1}(\mathbf{D}^\ell U_j^n)^*\,\widetilde{S}_\ell\,\mathbf{D}^{\ell+1}U_j^n\,. \quad (50)$$

The combination of (49) and (50) shows that there exists a quadratic form $q_\flat$ on $\mathbb{R}^{N(p+r)}$, which only depends on Q, such that any solution to (48) satisfies

$$\sum_{j\ge 1}|U_j^{n+1}|^2 - \sum_{j\ge 1}|U_j^n|^2 \le q_\flat(U_{1-r}^n,\dots,U_p^n)\,.$$

In particular, there exists a numerical constant C, that only depends on the operator Q and not on the solution (U_j^n) to (48), such that

$$\sum_{j\ge 1}|U_j^{n+1}|^2 - \sum_{j\ge 1}|U_j^n|^2 \le C\sum_{j=1-r}^{p}|U_j^n|^2\,.$$

Summing with respect to n, and using the fact that $\Delta t/\Delta x$ is constant, we end up with

$$\sup_{n\in\mathbb{N}}\sum_{j\ge 1}\Delta x\,|U_j^n|^2 \le \sum_{j\ge 1}\Delta x\,|f_j|^2 + C\sum_{n\ge 0}\sum_{j=1-r}^{p}\Delta t\,|U_j^n|^2\,. \quad (51)$$

Let us now observe that for all $n\in\mathbb{N}$, we have

$$\sum_{j=1-r}^{0}\Delta x\,|U_j^n|^2 \le \frac{1}{\lambda}\sum_{j=1-r}^{0}\Delta t\,|U_j^n|^2 \le \frac{1}{\lambda}\sum_{\nu\in\mathbb{N}}\sum_{j=1-r}^{0}\Delta t\,|U_j^\nu|^2\,,$$

so that the left hand-side of (51) can be slightly increased in order to obtain

$$\sup_{n\in\mathbb{N}}\sum_{j\ge 1-r}\Delta x\,|U_j^n|^2 \le \sum_{j\ge 1}\Delta x\,|f_j|^2 + C\sum_{n\ge 0}\sum_{j=1-r}^{p}\Delta t\,|U_j^n|^2\,.$$

We then use Theorem 1 to control the trace of (U_j^n) in terms of the initial condition (f_j), which is done by letting γ tend to zero in (11), and this completes the proof of Corollary 1.

Remark 3. *The above derivation of the semigroup estimate for the solution (U_j^n) to (48) heavily relies on the assumption $\|Q_0\|_{\ell^2(\mathbb{Z})\to\ell^2(\mathbb{Z})} = 1$, which in view of the consistency assumption on the A_ℓ's, is*

equivalent to $\|Q_0\|_{\ell^2(\mathbb{Z})\to\ell^2(\mathbb{Z})} \le 1$. *This property is called "strong stability" in [Str68], see also [Tad86], though "strong stability" in this context should not be mixed up with Definition 1.*

The exact same assumption on Q_0 *is the cornerstone of the analysis in [CG11]. Since [Wu95] deals with scalar problems, this assumption is also present, though hidden, in [Wu95]. However, the method we use here is completely different from the one in [Wu95, CG11] and bypasses the introduction of Dirichlet or other auxiliary boundary conditions. Unlike [Wu95, CG11], we use here the consistency of the discretized hyperbolic operator in order to derive an "integration by parts formula", which connects the time derivative of the* ℓ^2 *norm of* (U_j^n) *with the trace of* (U_j^n) *on the first space meshes.*

We aim in a near future to extend the derivation of such an "integration by parts formula" to numerical schemes with arbitrarily many time levels, which would imply, with the help of Theorem 1, a semigroup estimate for the solution to (2) and therefore a positive answer to the uniform power boundedness conjecture.

A On the non-glancing condition

The goal of this Appendix is to show that the validity of Proposition 1 is equivalent to the non-occurrence of glancing wave packets. This uses similar constructions as those in [Tre84], namely we use *discrete geometric optics* expansions. Opposite to [Tre84], we use here a *fully discrete* framework, namely we only deal with piecewise constant functions. This has a major impact on the arguments we use. While in [Tre84, Lemma 5.1], L^∞ error bounds are derived by using decay of the Fourier transform (or, equivalently, smoothness of the functions), the framework of step functions yields Fourier transforms that have no better than L^2 decay (and certainly not L^1). Hence the derivation of L^∞ error bounds is more intricate than in [Tre84], and we pay special attention to the rigorous justification of our error bound below. Our result is the following.

Proposition 2. *Let Assumptions 1 and 2 be satisfied. Assume furthermore that there exists a constant* $C > 0$ *such that for all* $\Delta t \in]0,1]$, *and for all solution to the fully discrete Cauchy problem*

$$\begin{cases} V_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} Q_\sigma\, V_j^{n-\sigma}\,, & j\in\mathbb{Z}\,, \quad n \ge s\,, \\ V_j^n = f_j^n\,, & j\in\mathbb{Z}\,, \quad n = 0,\dots,s\,, \end{cases} \tag{52}$$

there holds

$$\sum_{n\ge 0} \Delta t\, |V_0^n|^2 \le C \sum_{n=0}^{s} \sum_{j\in\mathbb{Z}} \Delta x\, |f_j^n|^2\,. \tag{53}$$

Then Assumption 3 is satisfied.

The proof of Proposition 2 is based on high frequency asymptotics for solutions to (52). We first state independently a Lemma which gives the expression of the Fourier transform of a piecewise constant "highly oscillating" function[14].

Lemma 5. *Let* a *denote a Schwartz function from* $\mathbb{R}$ *to* $\mathbb{C}^q$ *for some* $q\in\mathbb{N}$. *Given* $\underline{\xi}\in\mathbb{R}$ *and* $\Delta x > 0$, *we consider the step function*

$$\forall\, j\in\mathbb{Z}\,, \quad \forall\, x\in[j\,\Delta x, (j+1)\,\Delta x[\,, \quad a_\Delta(x) := \mathrm{e}^{i\,j\,\underline{\xi}}\, a(j\,\Delta x)\,.$$

[14] Of course the maximal frequency that is compatible with the mesh is $2\,\pi/\Delta x$ so high frequency in our discrete setting means a frequency of order $1/\Delta x$.

Then $a_\Delta \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and its Fourier transform is given by

$$\forall\, \xi \in \mathbb{R}\,, \quad \widehat{a_\Delta}(\xi) = \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m \in \mathbb{Z}} \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x}\right).$$

Observe that the function a_Δ in Lemma 5 is a piecewise constant version of the "continuous" function

$$x \in \mathbb{R} \longmapsto \mathrm{e}^{i\,x\,\underline{\xi}/\Delta x}\, a(x)\,,$$

which represents a fast oscillation at frequency $\underline{\xi}/\Delta x$ (Δx is meant to be small while $\underline{\xi}$ is fixed), with a slowly varying smooth envelope a.

Proof of Lemma 5. Due to the fast decay of a at infinity, the Fourier transform of a_Δ is given by

$$\widehat{a_\Delta}(\xi) = \sum_{j \in \mathbb{Z}} \mathrm{e}^{i\,j\,\underline{\xi}}\, a(j\,\Delta x) \int_{j\,\Delta x}^{(j+1)\,\Delta x} \mathrm{e}^{-i\,x\,\xi}\,\mathrm{d}x = \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\xi} \sum_{j \in \mathbb{Z}} \mathrm{e}^{-i\,j\,\Delta x\,(\xi - \underline{\xi}/\Delta x)}\, a(j\,\Delta x)\,,$$

and it only remains to apply the so-called Poisson summation formula to obtain the result of Lemma 5. □

Proof of Proposition 2. Let us first rewrite (52) as a scheme with two time levels for an augmented vector. Namely, we introduce $W_j^n := (V_j^{n+s}, \dots, V_j^n) \in \mathbb{C}^{N\,(s+1)}$, and rewrite (52) as

$$W_j^{n+1} = \mathscr{Q}\, W_j^n\,, \quad j \in \mathbb{Z}\,, \quad n \geq 0\,, \tag{54}$$

with the operator

$$\mathscr{Q} := \begin{pmatrix} Q_0 & \dots & \dots & Q_s \\ I & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & I & 0 \end{pmatrix}.$$

The estimate (53) can be equivalently rewritten for solutions to (54) as

$$\sum_{n \geq 0} \Delta t\, |W_0^n|^2 \leq C \sum_{j \in \mathbb{Z}} \Delta x\, |W_j^0|^2\,. \tag{55}$$

Let us consider some fixed parameter $\underline{\xi} \in [0, 2\,\pi[$, a Schwartz function a from $\mathbb{R}$ to $\mathbb{C}^{N\,(s+1)}$ and the initial sequence for (54):

$$\forall\, j \in \mathbb{Z}\,, \quad W_j^0 := \mathrm{e}^{i\,j\,\underline{\xi}}\, a(j\,\Delta x)\,.$$

For all $n \in \mathbb{N}$, the step function corresponding to the sequence $(W_j^n)_{j \in \mathbb{Z}}$ is denoted W_Δ^n. Applying the Fourier transform to (54) and using Lemma 5, we have

$$\widehat{W_\Delta^n}(\xi) = \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m \in \mathbb{Z}} \mathscr{A}\big(\mathrm{e}^{i\,\Delta x\,\xi}\big)^n\, \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x}\right).$$

We now use Assumptions 1 and 2 in order to give a detailed expansion of the amplification matrix $\mathscr{A}$ close to $\exp(i\,\underline{\xi})$: there exists an integer P such that, on the disk $\{\eta \in \mathbb{C} \,/\, |\eta - \underline{\xi}| < \delta_0\}$, $\mathscr{A}$ admits the spectral decomposition

$$\mathscr{A}(\mathrm{e}^{i\,\eta}) = \sum_{p=1}^{P} \mathrm{e}^{i\,\omega_p(\eta)}\, \Pi_p(\eta) + \mathscr{A}_\sharp(\eta)\, \Pi_\sharp(\eta)\,,$$

with (scalar) functions $\omega_1, \dots, \omega_P$, rank one projectors $\Pi_1, \dots, \Pi_P$, a rank $N(s+1) - P$ projector $\Pi_\sharp$ and a square matrix $\mathscr{A}_\sharp$ that has spectral radius less than 1 for all η. In the latter decomposition, all functions depend holomorphically on η. The functions $\omega_1, \dots, \omega_P$ satisfy

$$\forall\, p = 1, \dots, P\,, \quad \omega_p(\underline{\xi}) \in \mathbb{R}\,,$$

so the $\exp(i\,\omega_p(\eta))$ correspond to the eigenvalues of the amplification matrix that are close to the unit circle as $\exp(i\,\eta)$ is close to $\exp(i\,\underline{\xi})$. Of course, we can extend all functions to the disks $\{\eta \in \mathbb{C} \,/\, |\eta - (\underline{\xi} + 2\,m\,\pi)| < \delta_0\}$, $m \in \mathbb{Z}$, by $2\,\pi$-periodicity because $\mathscr{A}(\exp(i\,\cdot))$ is $2\,\pi$-periodic. The latter spectral decomposition of $\mathscr{A}$ only holds "microlocally", that is, locally near $\underline{\xi} + 2\,\pi\,\mathbb{Z}$. To avoid technicalities, we assume that a satisfies

$$a \in \mathscr{C}_0^\infty(\mathbb{R})\,, \quad \text{Supp}\ \widehat{a} \subset [-\delta_0/2, \delta_0/2]\,.$$

In this way, the expression of $\widehat{W_\Delta^n}$ splits into

$$\widehat{W_\Delta^n}(\xi) = \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m \in \mathbb{Z}} \sum_{p=1}^{P} \mathrm{e}^{i\,n\,\omega_p(\xi\,\Delta x)}\, \Pi_p(\xi\,\Delta x)\, \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x}\right) + \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m \in \mathbb{Z}} \mathscr{A}_\sharp(\xi\,\Delta x)^n\, \Pi_\sharp(\xi\,\Delta x)\, \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x}\right). \quad (56)$$

Following [Tre82, Tre84], we define the group velocities $\mathbf{v}_p := -\omega_p'(\underline{\xi})/\lambda$, which by Assumption 1, are known to be real (see, for instance, [Tre84, Lemma 3.2]). In the notation of Assumption 2, the group velocity is equivalently given by $\mathbf{v}_p = -\underline{\kappa}\,\zeta_p'(\underline{\kappa})/(\lambda\,\underline{z})$. In particular, Assumption 3 is valid provided that the scheme does not admit any wave packet with a vanishing group velocity. We introduce the "approximate" solution to (54) by defining:

$$\forall\,(j,n) \in \mathbb{Z} \times \mathbb{N}\,, \quad \mathscr{W}_j^n := \sum_{p=1}^{P} \mathrm{e}^{i\,(n\,\omega_p(\underline{\xi}) + j\,\underline{\xi})}\, \Pi_p(\underline{\xi})\, a(j\,\Delta x - n\,\Delta t\,\mathbf{v}_p)\,,$$

which represents a sum of highly oscillating signals with phase velocity $-\omega_p(\underline{\xi})/(\lambda\,\underline{\xi})$, and corresponding smooth envelopes that propagate at the group velocity $\mathbf{v}_p$. According to Lemma 5, the Fourier transform of the corresponding piecewise constant function is given by

$$\widehat{\mathscr{W}_\Delta^n}(\xi) = \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m \in \mathbb{Z}} \sum_{p=1}^{P} \mathrm{e}^{i\,n\,\omega_p(\underline{\xi}) + i\,n\,\Delta x\,\omega_p'(\underline{\xi})\,(\xi - (\underline{\xi} + 2\,m\,\pi)/\Delta x)}\, \Pi_p(\underline{\xi})\, \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x}\right). \quad (57)$$

We are now going to estimate the error $W_0^n - \mathscr{W}_0^n$.

Let us define the error:

$$\forall\,(j,n) \in \mathbb{Z} \times \mathbb{N}\,, \quad e_j^n := W_j^n - \mathscr{W}_j^n\,.$$

The expressions (56) and (57) show that the Fourier transform $\widehat{e_\Delta^n}$ splits as:

$$\widehat{e_\Delta^n} = \sum_{p=1}^{P} \varepsilon_{1,p}^n(\xi) + \varepsilon_{2,p}^n(\xi) + \varepsilon_\sharp^n(\xi)\,,$$

with, for all $p = 1, \dots, P$,

$$\varepsilon^n_{1,p}(\xi) := \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m\in\mathbb{Z}} \left(\mathrm{e}^{i\,n\,\omega_p(\xi\,\Delta x)} - \mathrm{e}^{i\,n\,\omega_p(\underline{\xi}) + i\,n\,\omega'_p(\underline{\xi})\,(\xi\,\Delta x - \underline{\xi} - 2\,m\,\pi)} \right) \Pi_p(\underline{\xi})\,\widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x} \right), \quad (58)$$

$$\varepsilon^n_{2,p}(\xi) := \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m\in\mathbb{Z}} \mathrm{e}^{i\,n\,\omega_p(\xi\,\Delta x)} \left(\Pi_p(\xi\,\Delta x) - \Pi_p(\underline{\xi} + 2\,m\,\pi) \right) \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x} \right), \quad (59)$$

and

$$\varepsilon^n_\sharp(\xi) := \frac{1 - \mathrm{e}^{-i\,\Delta x\,\xi}}{i\,\Delta x\,\xi} \sum_{m\in\mathbb{Z}} \mathscr{A}_\sharp(\xi\,\Delta x)^n\,\Pi_\sharp(\xi\,\Delta x)\,\widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x} \right). \quad (60)$$

Let us first estimate the L^2 norm of $\varepsilon^n_{1,p}$ in (58). We fix a time $T > 0$, and consider integers n such that $n\,\Delta t \le T$. Since $\omega_p(\underline{\xi})$ and $\omega'_p(\underline{\xi})$ are real, there holds

$$|\varepsilon^n_{1,p}(\xi)| \le C\,\frac{|1 - \mathrm{e}^{-i\,\Delta x\,\xi}|}{\Delta x\,|\xi|} \sum_{m\in\mathbb{Z}} \left| \mathrm{e}^{i\,n\,\omega_p(\xi\,\Delta x) - i\,n\,\omega_p(\underline{\xi} + 2\,m\,\pi) - i\,n\,\omega'_p(\underline{\xi} + 2\,m\,\pi)\,(\xi\,\Delta x - \underline{\xi} - 2\,m\,\pi)} - 1 \right| \left| \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x} \right) \right|.$$

There is no loss of generality in assuming $\delta_0/\lambda < \pi$. Then the support property of $\widehat{a}$ shows that in the latter sum with respect to $m \in \mathbb{Z}$, at most one term is nonzero. Consequently, there holds

$$|\varepsilon^n_{1,p}(\xi)|^2 \le C\,\frac{|1 - \mathrm{e}^{-i\,\Delta x\,\xi}|^2}{\Delta x^2\,\xi^2} \sum_{m\in\mathbb{Z}} \left| \mathrm{e}^{i\,n\,\omega_p(\xi\,\Delta x) - i\,n\,\omega_p(\underline{\xi} + 2\,m\,\pi) - i\,n\,\omega'_p(\underline{\xi} + 2\,m\,\pi)\,(\xi\,\Delta x - \underline{\xi} - 2\,m\,\pi)} - 1 \right|^2 \left| \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x} \right) \right|^2,$$

and because of the limitation $n\,\Delta t \le T$, there holds[15]

$$\left| \mathrm{e}^{i\,n\,\omega_p(\xi\,\Delta x) - i\,n\,\omega_p(\underline{\xi} + 2\,m\,\pi) - i\,n\,\omega'_p(\underline{\xi} + 2\,m\,\pi)\,(\xi\,\Delta x - \underline{\xi} - 2\,m\,\pi)} - 1 \right| \le C\,T\,\Delta x \left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x} \right)^2 \le C\,T\,\Delta x\,,$$

on the support of $\widehat{a}(\xi - (\underline{\xi} + 2\,m\,\pi)/\Delta x)$. We thus derive the bound

$$\begin{aligned}
\int_{\mathbb{R}} |\varepsilon^n_{1,p}(\xi)|^2\,\mathrm{d}\xi &\le C\,T^2\,\Delta x^2 \sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \frac{|1 - \mathrm{e}^{-i\,\Delta x\,\xi}|^2}{\Delta x^2\,\xi^2} \left| \widehat{a}\left(\xi - \frac{\underline{\xi} + 2\,m\,\pi}{\Delta x} \right) \right|^2 \mathrm{d}\xi \\
&\le C\,\|\widehat{a}\|^2_{L^\infty}\,T^2\,\Delta x^2 \sum_{m\in\mathbb{Z}} \int_{(\underline{\xi} + 2\,m\,\pi)/\Delta x - \delta_0/2}^{(\underline{\xi} + 2\,m\,\pi)/\Delta x + \delta_0/2} \frac{|1 - \mathrm{e}^{-i\,\Delta x\,\xi}|^2}{\Delta x^2\,\xi^2}\,\mathrm{d}\xi \\
&\le C\,T^2\,\Delta x \sum_{m\in\mathbb{Z}} \int_{\underline{\xi} + 2\,m\,\pi - \delta_0\,\Delta x/2}^{\underline{\xi} + 2\,m\,\pi + \delta_0\,\Delta x/2} \frac{|1 - \mathrm{e}^{-i\,\eta}|^2}{\eta^2}\,\mathrm{d}\eta \\
&\le C\,T^2\,\Delta x \sum_{m\in\mathbb{Z}} \int_{\underline{\xi} + 2\,m\,\pi - \delta_0\,\Delta x/2}^{\underline{\xi} + 2\,m\,\pi + \delta_0\,\Delta x/2} \frac{1}{1 + \eta^2}\,\mathrm{d}\eta \le C\,T^2\,\Delta x^2\,,
\end{aligned}$$

[15] Here we use Assumption 1 to obtain that the imaginary part of $\omega_p(\xi\,\Delta x)$ is nonpositive.

with a constant $C > 0$ that is uniform with respect to $T > 0$ and $\Delta t \in]0, 1]$. (Recall that the ratio $\Delta t/\Delta x$ is kept fixed.) Similarly, the error $\varepsilon^n_{2,p}$ in (59) satisfies[16]

$$\int_{\mathbb{R}} |\varepsilon^n_{2,p}(\xi)|^2 \, \mathrm{d}\xi \leq C \, \Delta x^2 \,,$$

with a constant $C > 0$ that is uniform with respect to $T > 0$ and $\Delta t \in]0, 1]$.

If $\mathscr{W}^n_j$ is meant to be a good approximation of W^n_j, including for small values of n, then the term $\varepsilon^n_\sharp$ in (60) is meant to be small. In order to achieve this, we assume that a satisfies the polarization condition

$$\Pi_\sharp(\underline{\xi}) \, a = 0 \,.$$

Let us now derive an L^2 bound for $\varepsilon^n_\sharp$. Shrinking δ_0 is necessary, there is no loss of generality in assuming that the matrix $\mathscr{A}_\sharp$ in the spectral decomposition of $\mathscr{A}$ is power bounded:

$$\sup_{n \in \mathbb{N}} |\mathscr{A}_\sharp(\eta)^n| \leq C \,,$$

with a constant $C > 0$ that is uniform with respect to η as long as $|\eta - (\underline{\xi} + 2\,m\,\pi)| \leq \delta_0/2$. (We shall not even use here the exponential decay in time of the $\sharp$ component.) Performing the same kind of analysis as for the terms $\varepsilon^n_{2,p}$, the error $\varepsilon^n_\sharp$ in (60) satisfies

$$\int_{\mathbb{R}} |\varepsilon^n_\sharp(\xi)|^2 \, \mathrm{d}\xi \leq C \, \Delta x^2 \,,$$

with a constant $C > 0$ that is uniform with respect to $T > 0$ and $\Delta t \in]0, 1]$.

By Plancherel Theorem, we have proved the bound

$$\sum_{j \in \mathbb{Z}} \Delta x \, |e^n_j|^2 \leq C \, \Delta x^2 \, (1 + T^2) \,,$$

for all n such that $n \, \Delta t \leq T$ and a constant C that is uniform with respect to $T > 0$ and $\Delta x \in]0, 1]$. In particular, there holds

$$\|W^n_\Delta - \mathscr{W}^n_\Delta\|^2_{L^\infty(\mathbb{R})} = \sup_{j \in \mathbb{Z}} |W^n_j - \mathscr{W}^n_j|^2 \leq C \, \Delta x \, (1 + T^2) \,, \tag{61}$$

which gives an L^∞ bound for the error between the exact and approximate solutions provided that $\widehat{a}$ has sufficiently narrow support, and a is suitably polarized ($\Pi_\sharp(\underline{\xi}) \, a = 0$).

The proof of Proposition 2 is now almost complete. Indeed, let us assume that Assumption 3 is not valid. Up to reordering, this means that for some $\underline{\xi}$, the group velocity $\mathbf{v}_1$ is zero. We use the previous construction of high frequency solutions to (54). Choosing a such that the (more restrictive) polarization condition $\Pi_1(\underline{\xi}) \, a = a$ holds, the expression of the approximate solution $\mathscr{W}$ reduces to

$$\forall \, (j, n) \in \mathbb{Z} \times \mathbb{N} \,, \quad \mathscr{W}^n_j := \mathrm{e}^{i \, (n \, \omega_1(\underline{\xi}) + j \, \underline{\xi})} \, a(j \, \Delta x) \,.$$

Let us consider some time $T > 0$. The trace estimate (55) gives

$$\begin{aligned} \sum_{0 \leq n \leq T/\Delta t} \Delta t \, |\mathscr{W}^n_0|^2 &\leq 2 \sum_{0 \leq n \leq T/\Delta t} \Delta t \, |W^n_0|^2 + 2 \sum_{1 \leq n \leq T/\Delta t} \Delta t \, |\mathscr{W}^n_0 - W^n_0|^2 \\ &\leq C \sum_{j \in \mathbb{Z}} \Delta x \, |W^0_j|^2 + C \, \Delta x \, (1 + T^2) \, T \,. \end{aligned}$$

[16] Here we use again Assumption 1 in order to have $|\exp(i \, n \, \omega_p(\xi \, \Delta x))| \leq 1$ uniformly in n.

By the smoothness of a, there holds

$$\sum_{j\in\mathbb{Z}} \Delta x\, |W_j^0|^2 = \sum_{j\in\mathbb{Z}} \Delta x\, |a(j\,\Delta x)|^2 \le C \sum_{j\in\mathbb{Z}} \|a\|^2_{L^2([j\,\Delta x,(j+1)\,\Delta x[)} + \Delta x^2\, \|a'\|^2_{L^2(L^2([j\,\Delta x,(j+1)\,\Delta x[)} \le C\, \|a\|^2_{H^1(\mathbb{R})}\,,$$

uniformly with respect to $\Delta t \in]0,1]$. Summing up, we have shown that, for a suitable constant $C>0$ that is uniform with respect to $T>0$ and $\Delta t\in]0,1]$, there holds

$$(N_T+1)\,\Delta t\, |a(0)|^2 \le C + C\,\Delta x\,(1+T^2)\,T\,,$$

with N_T the largest integer such that $N_T\,\Delta t \le T$. By first passing to the limit $\Delta t\to 0$, we get

$$T\,|a(0)|^2 \le C\,,$$

and by passing to the limit $T\to+\infty$, we get $a(0)=0$, which is obviously a contradiction because one can construct the function a that meets all previous requirements (support of $\widehat{a}$, smoothness and polarization), together with $a(0)\neq 0$. □

Remark 4. *The above argument is actually simpler in the PDE context because an accurate description of high frequency asymptotics (including L^∞ error bounds) is available for hyperbolic systems, say with constant multiplicity. Consider for instance the Cauchy problem*

$$\partial_t u + \sum_{j=1}^d A_j\,\partial_{x_j} u = 0\,,$$

with a hyperbolic operator of constant multiplicity, that is:

$$\forall\,\xi=(\xi_1,\dots,\xi_d)\in\mathbb{R}^d\setminus\{0\}\,,\quad \det\Big[\tau\,I+\sum_{j=1}^d \xi_j\,A_j\Big] = \prod_{k=1}^q \big(\tau+\lambda_k(\xi)\big)^{\nu_k}\,,$$

with (real valued) real analytic semi-simple eigenvalues $\lambda_1,\dots,\lambda_q$. The validity of the trace estimate

$$\int_{\mathbb{R}^+}\int_{\mathbb{R}^{d-1}} |u(t,y,0)|^2\,\mathrm{d}y\,\mathrm{d}t \le C\,\|u(0,\cdot)\|^2_{L^2(\mathbb{R}^d)}\,,$$

is equivalent to the fact that there is no glancing wave packet, namely:

$$\forall\,\xi\neq 0\,,\quad \forall\,k=1,\dots,q\,,\quad \frac{\partial\lambda_k(\xi)}{\partial\xi_d}\neq 0\,.$$

The latter condition is basically never satisfied in dimension $d\ge 2$, and this is one reason why the derivation of semigroup estimates in [Kaj72, Rau72] and followers is so involved.

Acknowledgments The discussion in Appendix A originates from a discussion in Les Houches with Guy Métivier, whom I warmly thank for pointing out several "well-known" results. I also thank Mark Williams for providing and clarifying parts of [MT]. Eventually, it is a pleasure to thank Denis Serre, whose interest and encouragements over the years have been very stimulating.

References

[Aud11] C. Audiard. On mixed initial-boundary value problems for systems that are not strictly hyperbolic. *Appl. Math. Lett.*, 24(5):757–761, 2011.

[BGS07] S. Benzoni-Gavage and D. Serre. *Multidimensional hyperbolic partial differential equations*. Oxford University Press, 2007. First-order systems and applications.

[Car43] F. Carlson. Quelques inégalités concernant les fonctions analytiques. *Ark. Mat. Astr. Fys.*, 29B(11):6, 1943.

[CG11] J.-F. Coulombel and A. Gloria. Semigroup stability of finite difference schemes for multidimensional hyperbolic initial boundary value problems. *Math. Comp.*, 80(273):165–203, 2011.

[Cou09] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. *SIAM J. Numer. Anal.*, 47(4):2844–2871, 2009.

[Cou11] 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.

[Cou13] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. In *HCDTE Lecture Notes. Part I. Nonlinear Hyperbolic PDEs, Dispersive and Transport Equations*, pages 97–225. American Institute of Mathematical Sciences, 2013.

[GKO95] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods*. John Wiley & Sons, 1995.

[GKS72] 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.

[Kaj72] K. Kajitani. Initial-boundary value problems for first order hyperbolic systems. *Publ. Res. Inst. Math. Sci.*, 7:181–204, 1971/72.

[Kre68] H.-O. Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comp.*, 22:703–714, 1968.

[Kre70] H.-O. Kreiss. Initial boundary value problems for hyperbolic systems. *Comm. Pure Appl. Math.*, 23:277–298, 1970.

[KW93] H.-O. Kreiss and L. Wu. On the stability definition of difference approximations for the initial-boundary value problem. *Appl. Numer. Math.*, 12(1-3):213–227, 1993.

[Mét14] G. Métivier. On the L^2 well posedness of hyperbolic initial boundary value problems. *Preprint*, 2014.

[MT] R. Melrose and M. Taylor. *Boundary problems for wave equations with grazing and gliding rays*. Unpublished notes.

[Osh69a] S. Osher. Stability of difference approximations of dissipative type for mixed initial boundary value problems. I. *Math. Comp.*, 23:335–340, 1969.

[Osh69b] S. Osher. Systems of difference equations with general homogeneous boundary conditions. *Trans. Amer. Math. Soc.*, 137:177–201, 1969.

[Rau72] J. Rauch. $\mathscr{L}^2$ is a continuable initial condition for Kreiss' mixed problems. *Comm. Pure Appl. Math.*, 25:265–285, 1972.

[Sar65] L. Sarason. On hyperbolic mixed problems. *Arch. Rational Mech. Anal.*, 18:310–334, 1965.

[Sar77] L. Sarason. Hyperbolic and other symmetrizable systems in regions with corners and edges. *Indiana Univ. Math. J.*, 26(1):1–39, 1977.

[Str68] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[Tad86] E. Tadmor. Complex symmetric matrices with strongly stable iterates. *Linear Algebra Appl.*, 78:65–77, 1986.

[TE05] L. N. Trefethen and M. Embree. *Spectra and pseudospectra*. Princeton University Press, 2005. The behavior of nonnormal matrices and operators.

[Tre82] L. N. Trefethen. Group velocity in finite difference schemes. *SIAM Rev.*, 24(2):113–136, 1982.

[Tre84] L. N. Trefethen. Instability of difference models for hyperbolic initial boundary value problems. *Comm. Pure Appl. Math.*, 37:329–367, 1984.

[Wu95] L. Wu. The semigroup stability of the difference approximations for initial-boundary value problems. *Math. Comp.*, 64(209):71–88, 1995.