THÈSE DE DOCTORAT DE

UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N^o 601

Mathématiques et Sciences et Technologies de l'Information et de la Communication

Spécialité : *Mathématiques et leurs Interactions*

Par

Thi Hoai Thuong NGUYEN

Numerical approximation of boundary conditions and stiff source terms in hyperbolic equations

Thèse présentée et soutenue à Rennes, le 13 octobre 2020

Unité de recherche : UMR CNRS 6625 Institut de Recherche Mathématiques de Rennes (IRMAR)

Rapporteurs avant soutenance :

Sylvie BENZONI-GAVAGE	Professeure, Université de Lyon 1
Christophe BESSE	Professeur, Université de Toulouse 3

Composition du Jury :

Examinateurs :	Sylvie BENZONI-GAVAGE	Professeure, Université de Lyon 1
	Christophe BESSE	Professeur, Université de Toulouse 3
	Jean-François COULOMBEL	Directeur de Recherche CNRS, Université de Toulouse
	Sonia FLISS	Maître de Conférences, ENSTA ParisTech
	Florian MEHATS	Professeur, Université de Rennes 1
Dir. de thèse :	Nicolas SEGUIN,	Professeur, Université de Rennes 1
Co-dir. de thèse :	Benjamin BOUTIN,	Maître de Conférences, Université de Rennes 1

Acknowledgements

This dissertation would not have been completed without the help, encouragement and support from many people whom I would like to send all my best wishes and all my honest thanks to.

First of all, my deepest gratitude goes to my supervisors Professor Benjamin Boutin and Professor Nicolas Seguin, who have always been there with me since my master internship, sharing with me theirs valuable advices, supporting me during the darkest moments of my Ph.D. study and inspiring me by simply being themselves. Furthermore, I would like to thank them because they gave me a very precious opportunity to collaborate with them in my three-years doctoral period. Thanks to their guidance, patience, motivation and immense knowledge, I could finish my thesis. I will never forget the days when they were sitting with me to try to correct a lot of English errors in my research statement and this long manuscript.

I greatly appreciate Professor Sylvie Benzoni-Gavage and Professor Christophe Besse, the two reviewers of this dissertation, who have been so kind to spend time to read, evaluate my works and give me valuable comments. I am also thankful to Professor Jean-François Coulombel, Professor Sonia Fliss and Professor Florian Mehats who accepted to be in my thesis committee.

I am very grateful to Institute of Mathematical Research of Rennes (IRMAR, University of Rennes 1) and Center of Henri Lebesgue (CHL) and the staff therein. A special thank is sent to Xhenxila Lachambre, Marie-Aude Verger, Florian Rogowski, Helène Rousseaux, Elodie Cottrel, and Chantal Halet, who have endlessly supported me even before the first day I was in France till the last. I also acknowledge the partial support from the ANR project NABUCO during time at the CEMRACS.

I would also like to thank my PUF close friends: Truong Hoang Ngan, Cao Van Kien, Vu Duc Thach Son, Nguyen Van Thanh, Phan Tan Binh, Nguyen Dang Phuong Lan for everything we have been through. A special thanks must go to my friends: Abraham Sylla, Sébastien Tran Tien, Khawla Msheik, Nguyen Van Duc for a wonderful time at the CEMRACS.

Besides, I have also been lucky to meet and make friends with many nice people in Rennes. Thank you my brothers and sisters Nguyen Thi Lan Anh, Chau Thi Tuyet Trang, Dao Manh Khang... who helped me many things when I have been living in France. Thank you "5 anh em sieu nhan" Nguyen Duc Tho, Nguyen Dinh Duong, Nguyen Thi Minh Phuong, Truong Thi Hao, and many others... for standing beside me in every events of my life anh giving me a warming second family. The events has gone, but the memory about each of you will stay with me forever.

Last but not least, I would like to express all my gratitude to my family for constant encouragement, unconditional love and trusting me in every decision. I am truly thankful to my best friend, Pham Nhut Truong, for being on my side whenever I need. His encouragements and supports are motivations for me in the past, the present and the future. Especially, I dedicate this thesis to my departed maternal grandparents. I just want to say to them "Con cam on Ngoai da luon yeu thuong, cham soc va quan tam con. Ngoai mai mai o trong ki uc, trong tam tri va trong tim cua con. Con yeu Ngoai nhieu”.

Résumé
Approximation numérique des conditions aux bords et des termes sources raides dans les équations hyperboliques

1. Introduction Cette thèse est consacrée l'analyse théorique et numérique de systèmes hyperboliques d'équations aux dérivées partielles et aux équations de transport, comportant des termes de relaxation et des conditions aux bords. Elle est constituée de trois sujets qui font l'objet d'une présentation synthétique dans un premier chapitre introductif et sont ensuite détaillés dans les chapitres 2 à 5.

La première étude (chapitres 2 et 3) mêle deux thèmes délicats du point de vue de l'analyse mathématique des équations aux dérivées partielles hyperboliques, à savoir les problèmes aux limites et les problèmes raides. Elle les aborde du point de vue de l'approximation numérique. L'étude porte plus précisément sur la discrétisation de l'équation des ondes unidimensionnelle amortie

$$\frac{\partial^2 w}{\partial t^2} - a\frac{\partial^2 w}{\partial x^2} = -\frac{1}{\epsilon}\frac{\partial w}{\partial t},$$

où $\sqrt{a}$ est la vitesse des ondes, considérée sur le demi-espace $\{x > 0\}$. Il s'agit de proposer des schémas d'approximation numérique et d'étudier leur stabilité, ceci de façon uniforme vis à vis du paramètre caractéristique de relaxation $\epsilon > 0$, propriété appelée en anglais « stiff stability » que nous appellerons dans la suite stabilité raide ou rigide. Dans le chapitre 2, la condition au bord numérique utilisée s'appuie sur une technique de sommation par parties tandis que dans le chapitre 3, elle emploie la technique de condition transparente.

La seconde étude (chapitre 4) concerne l'élaboration et l'analyse de schémas numériques d'ordre élevé pour l'équation de transport unidimensionnelle

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$$

posée sur un domaine borné $x \in (0, L)$. Le traitement numérique des conditions aux limites en entrée et en sortie est réalisé à un ordre élevé de sorte à ne pas dégrader l'ordre du schéma intérieur. La convergence de la solution numérique est quantifiée précisément en fonction de différents paramètres retenus.

Le troisième sujet (chapitre 5) a trait a l'étude de la stabilité des solutions stationnaires de systèmes non-conservatifs avec terme source de relaxation par des techniques d'entropie relative.

2. Schémas rigidement stables pour l'équation des ondes amortie posée dans le quart de plan Le modèle considéré peut se réécrire sous la forme d'un système de deux équations d'ordre un

$$\begin{cases} \partial_t u(t,x) + \partial_x v(t,x) = 0, \\ \partial_t v(t,x) + a\partial_x u(t,x) = -\epsilon^{-1} v(t,x), \end{cases}$$

ceci en introduisant $u = \partial_x w$ et $v = -\partial_t w$. Ce système constitue un cas particulier de la classe de modèles de relaxation étudiés par S. Jin et Z.P. Xin [51]. Aux données initiales

$$u(0,x) = u_0(x), \quad v(0,x) = v_0(x), \quad x > 0$$

vient s'adjoindre une condition au bord $x = 0$ prenant la forme

$$B_u u(t,0) + B_v v(t,0) = b(t).$$

Ici B_u et B_v désignent des constantes réelles. La structure hyperbolique du système est caractérisée par les invariants de Riemann $\sqrt{a}u \pm v$ et les vitesses caractéristiques associées $\pm\sqrt{a}$. De ce fait, le caractère bien posé du modèle continu est subordonné à la condition de Kreiss uniforme (UKC)

$$B_u + \sqrt{a} B_v \neq 0.$$

Depus les travaux de Z. Xin et W.-Q. Xu [96], cette condition de stabilité de Kreiss est connue pour ne pas être suffisante pour garantir la stabilité rigide de ce problème, i.e. indépendamment du taux de relaxation $\epsilon \in (0, +\infty)$. Afin de remédier à cette difficulté, ces auteurs introduisent la condition de Kreiss rigide (SKC) se réduisant dans ce cadre à

$$B_v = 0 \quad \text{ou} \quad \frac{B_u}{B_v} \notin [-\sqrt{a}, 0].$$

W.-A. Yong [97] démontre que cette condition s'avère être nécessaire et suffisante pour garantir le caractère rigidement stable du modèle. La motivation principale de notre étude est d'aborder l'analogue discret de cette théorie de stabilité raide, dans le cadre de l'approximation numérique par différences finies du système de relaxation considéré. Dans ce cadre discret, la condition au bord scalaire précédente doit génériquement être complétée par une condition au bord numérique artificielle. Tout l'enjeu est de la choisir de façon à ne pas introduire d'instabilité numérique.

- Dans un premier temps, l'étude porte sur l'approximation semi-discrète suivante

$$\begin{cases} \dfrac{\partial}{\partial t} U_j(t) + (QU)_j(t) = \dfrac{1}{\epsilon} S U_j(t), & j \geq 1, \ t \geq 0, \\ U_j(0) = f_j, & j \geq 0, \\ BU_0(t) = b(t), & t \geq 0, \end{cases}$$

dans laquelle on a noté

$$U = \begin{pmatrix} u \\ v \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ a & 0 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} B_u & B_v \end{pmatrix},$$

et la discrétisation en espace est obtenue par l'opérateur de différence centré

$$(\mathcal{Q}U)_j = \frac{1}{2\Delta x} A(U_{j+1} - U_{j-1}), \quad j \geq 0.$$

La technique de sommation par parties permet de considérer dans cette définition la valeur $U_{-1} = 2U_0 - U_1$. La condition discrète complémentaire au bord que nous proposons est la projection

$$\Gamma \left(\frac{\partial}{\partial t} U_0(t) + (\mathcal{Q}U)_0(t) \right) = \frac{1}{\epsilon} \Gamma S U_0(t), \quad t \geq 0,$$

où $\Gamma = \begin{pmatrix} -aB_v & B_u \end{pmatrix}$. Dans un premier temps, nous détermions une condition suffisante de stabilité du schéma ainsi constitué. Au moyen d'estimations d'énergie et de la transformée de Laplace nous démontrons le théorème suivant

Théorème. *Supposons vérifiée la condition de dissipativité stricte $B_u B_v > 0$. Pour tout $T > 0$, il existe une constante $C_T > 0$ telle que pour toute donnée initiale $f \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ et toute donnée de bord $b \in \mathcal{C}^1(\mathbb{R}^+, \mathbb{R}) \cap L^2(\mathbb{R}^+, \mathbb{R})$ la solution $(U_j(t))$ du schéma semi-discret vérifie l'estimation*

$$\int_0^T |U_0(t)|^2 \, dt + \int_0^T \sum_{j \geq 0} \Delta x |U_j(t)|^2 \, dt \leq C_T \left(\sum_{j \geq 0} \Delta x |f_j|^2 + \int_0^T |b(t)|^2 \, dt \right),$$

où la constante C_T est indépendante des données f et b ainsi que du paramètre $\epsilon \in (0, +\infty)$ et de $\Delta x \in (0, 1]$

Tandis que Z. Xin et W.-Q. Xu démontrent que le modèle continu est rigidement bien posé si et seulement si la condition SKC est vérifiée, pour le problème semi-discret que nous considérons, cette condition ne semble pas suffisante pour obtenir des estimations de stabilité uniformes. La condition de dissipativité stricte est plus restrictive mais nous ne sommes pas en mesure que de démontrer qu'elle est nécessaire. Sur la base d'une analyse en modes normaux, nous sommes en mesure de construire des solutions instables et des investigations numériques complémentaires étayent la thèse selon laquelle la condition de dissipativité stricte serait également nécessaire pour la stabilité rigide.

• Dans un second temps, nous concentrons notre étude sur la discrétisation implicite en temps du schéma précédent. L'objectif est de déterminer une condition suffisante pour la stabilité rigide de ce schéma discret. Le résultat obtenu est le suivant

Théorème. *Supposons vérifiée la condition de dissipativité stricte $B_u B_v > 0$. Pour tout $T > 0$, il existe une constante $C_T > 0$ telle que pour tout $\Delta t > 0$ et toute constante $\delta < 3\sqrt{a}/8$ avec $\Delta x = \delta \Delta t$, toute donnée initiale $f \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ et toute donnée de bord $b \in \ell^2(\mathbb{N}, \mathbb{R})$, la solution (U_j^n) du schéma considéré vérifie l'estimation*

$$\sum_{n=0}^{N} \sum_{j \geq 0} \Delta x \Delta t |U_j^n|^2 + \sum_{n=0}^{N} \Delta t |U_0^n|^2 \leq C_T \left(\sum_{j \geq 0} \Delta x |f_j|^2 + \sum_{n=0}^{N} \Delta t |b^n|^2 \right),$$

où $N = T/\Delta t$ et C_T est indépendante de $\epsilon \in (0, +\infty)$.

• Enfin, nous considérons le schéma implicite basée sur l'approximation décentrée amont en espace. Cette dernière s'appuie sur la struture hyperbolique et les champs caractéristiques du problème continu. Pour ce schéma, au moyen d'estimations d'énergie discrète nous démontrons le résultat suivant

Théorème. *Supposons que les paramètres $(B_u, B_v) \in \mathbb{R}^2$, $\Delta x \in (0, 1]$, et $\epsilon > 0$ satisfont à la condition de dissipativité stricte*

$$2a \frac{B_u}{B_v} + \frac{\Delta x}{\epsilon} \left(\frac{B_u}{B_v} \right)^2 > 0.$$

Alors il existe une constante $C > 0$ telle que pour tout $\Delta t > 0$ et toute donnée initiale $f \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ la solution (U_j^n) du schéma considéré vérifie l'estimation

$$\langle U^n, HU^n \rangle_{\Delta x} + C \Delta t \sum_{k=0}^{n} |U_0^k|^2 \leq \langle f, Hf \rangle_{\Delta x}, \quad n \geq 0.$$

De plus,

a) Si $B_u B_v > 0$, alors l'estimation précédente est uniforme au sens où C est indépendante de ϵ et de Δx.

b) Si $B_u B_v < 0$, alors considérant $\delta > -2aB_v B_u^{-1}$, il existe $C = C(\delta) > 0$ telle que l'estimation précédente soit uniforme pour $\Delta x \geq \delta \epsilon$.

3. Un schéma rigidement stable pour l'équation des ondes amortie posée dans le quart de plan utilisant une condition transparente discrète
Dans le troisième chapitre de cette thèse, nous utilisons la technique de condition au bord transparente discrète pour construire un schéma numérique rigidement stable. Notre objectif est de démontrer que la condition SKC est une condition suffisante pour la stabilité rigide du schéma implicite en temps et centré en espace.

La condition au bord transparente discrète est mise en place au point $j = 0$ de façon à ce que le schéma implicite soit stable et à ce que sa solution coïncide avec celle du schéma considéré sur l'espace $\{j \in \mathbb{Z}\}$ entier, restreinte ensuite au demi-espace $\{j \in \mathbb{N}\}$. En utilisant ette notion de condition au bord transparente, nous sommes en mesure de définir

$$U_{-1}^n = \sum_{k=0}^{n} \mathcal{C}_{n-k} U_0^k, \quad n \geq 0,$$

formule dans laquelle les coefficients $\mathcal{C}_m$ pour $m \in \mathbb{N}$ sont déterminés par une formule explicite. Le schéma discret retenu prend alors la forme

$$\begin{cases} \dfrac{U_j^{n+1} - U_j^n}{\Delta t} + \dfrac{1}{2\Delta x} A(U_{j+1}^{n+1} - U_{j-1}^{n+1}) = \dfrac{1}{\epsilon} S U_j^{n+1}, & j \geq 1, \ n \geq 0, \\ U_j^0 = f_j, & j \geq 0, \\ B U_0^n = b^n, & n \geq 0, \\ \dfrac{1}{\Delta t} \Gamma (U_j^{n+1} - U_j^n) + \dfrac{1}{2\Delta x} \Gamma A \left(U_1^{n+1} - \displaystyle\sum_{k=0}^{n+1} \mathcal{C}_{n+1-k} U_0^k \right) = \dfrac{1}{\epsilon} \Gamma S U_0^{n+1}, & n \geq 0. \end{cases}$$

L'analyse de ce schéma est effectuée au moyen de la transformée en $\mathcal{Z}$ par rapport à l'indice de temps $n \in \mathbb{N}$, qui n'est que l'analogue discret de la transformée de Laplace en temps $t \in \mathbb{R}^+$. Nous obtenons alors le théorème suivant

Théorème. *Soient $(B_u, B_v) \in \mathbb{R}^2$ vérifiant la condition SKC et $\delta \leq 3\sqrt{a}/8$. Pour tout $T > 0$, il existe une constante $C_T > 0$ telle que pour tout $\Delta t > 0$ et $\Delta x = \delta \Delta t$, toute donnée initiale $f \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ et toute donnée de bord $b \in \ell^2(\mathbb{N}, \mathbb{R})$, la solution (U_j^n) du schéma précédent vérifie l'estimation*

$$\sum_{n=0}^{N} \sum_{j \geq 0} \Delta x \Delta t |U_j^n|^2 + \sum_{n=0}^{N} \Delta t |U_0^n|^2 \leq C_T \left(\sum_{j \geq 0} \Delta x |f_j|^2 + \sum_{n=0}^{N} \Delta t |b^n|^2 \right),$$

avec $N = T/\Delta t$ et C_T indépendante de $\epsilon \in (0, +\infty)$.

La condition SKC est ainsi également une condition suffisante pour garantir la stabilité rigide du schéma numérique considéré, indépendamment de la raideur ϵ du terme source, du pas d'espace Δx et du pas de temps Δt.

4. Schémas d'ordre élevé pour l'équation de transport sur un segment
Dans le quatrième chapitre de cette thèse, nous proposons un traitement d'ordre

élevé de la condition de bord entrante pour l'équation de transport unidimensionnelle

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0, \quad t \geq 0, \quad x \in (0, L), \\ u(0,x) = f(x), \quad x \in (0, L), \\ u(t,0) = g(t), \quad t \geq 0, \end{cases}$$

posée sur un intervalle de longueur $L > 0$ pour une vitesse positive $a > 0$. Dans un travail antérieur J.F. Coulombel et F. Lagoutière [22] ont analysé des discrétisations d'ordre élevé, stables et convergentes pour ce problème dans le cas d'une donnée nulle en entrée $g = 0$. L'objectif du chapitre 4 de cette thèse est d'étendre ce résultat au cas de données non-nulles au bord entrant, les données initiale f et de bord g étant alors soumises à une hypothèse de recollement régulier au coin $(x, t) = (0, 0)$.

Le schéma intérieur considéré prend la forme d'une itération linéaire explicite à $p + r + 1$ points, supposée ℓ^2-stable et d'ordre $k \geq 1$

$$u_j^{n+1} = \sum_{\ell=-r}^{p} a_\ell u_{j+\ell}^n, \quad j = 1, \ldots, J, \quad n \in \mathbb{N}.$$

Les coefficients a_ℓ, pour $\ell = -r, \ldots, p$, ne dépendent que du paramètre $\lambda = \Delta t/\Delta x$ et de la vitesse a. La définition du schéma est complétée par la constitution de valeurs pour la solution numérique (u_j^n) dans les mailles fantômes localisées au bord du domaine de calcul. Au bord entrant $x = 0$, la stratégie employée, désignée dans la littérature sous le nom de « méthode de Lax-Wendroff inverse », permet de garantir la consistance avec la donnée de Dirichlet du problème continu et prend la forme suivante

$$u_\ell^n = \sum_{m=0}^{k-1} \frac{\Delta x^m}{(m+1)!(-a)^m} (\ell^{m+1} - (\ell-1)^{m+1}) g^{(m)}(t), \quad \ell = 1 - r, \ldots, 0, \quad n \in \mathbb{N}.$$

Au bord sortant $x = L$, la condition discrète retenue, du type extrapolatoire, prend la forme

$$(D_-^{k_b} u^n)_{J+\ell} = 0, \quad \ell = 1, \ldots, p, \quad n \in \mathbb{N},$$

où D_- désigne l'opérateur de différence décalé à gauche et $k_b \in \mathbb{N}$ est l'ordre d'extrapolation retenu au bord. Enfin le schéma est initialisé à l'aide de la projection constante par morceau de la donnée initiale continue du problème :

$$u_j^0 = \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x)\, dx, \quad j = 1, \ldots, J.$$

Nous démontrons des estimations de convergence de la solution numérique vers la solution exacte en $\Delta x^{\min(k,k_b)-1/2}$, ceci sur tout intervalle de temps borné. La preuve repose sur les estimations de stabilité du schéma et un argument de superposition qui permet de considérer séparément deux demi-problèmes : le premier posé sur $(0,+\infty)$ qui concerne l'entrée seulement ; le second posé sur $(-\infty, L)$ qui traite de la sortie seulement. L'étude est complétée par une decription plus précise de la solution numérique au voisinage du bord sortant, ceci au moyen d'un développement de couche limite s'inspirant du travail antérieur de B. Boutin et J.F. Coulombel [6]. Le terme dominant de ce développement correspond à la solution du problème continu. Ne vérifiant elle-même pas la condition d'extrapolation discrète de sortie, un terme d'erreur en $O(\Delta x^{k_b})$ apparaît au bord. Sous des hypothèses peu contraignantes concernant la structure du schéma, nous démontrons que cette erreur au bord engendre un terme supplémentaire de couche limite d'amplitude $O(\Delta x^{k_b+1/2})$ dans la norme ℓ^2 discrète. Ce gain d'un facteur $\Delta x^{1/2}$ nous permet de retrouver dans le cas $k_b < k$ un taux de convergence optimal k_b dans la norme uniforme ℓ^∞.

5. Stabilité de solutions stationnaires parmi les solutions processus entropiques La stabilité des solutions stationnaires de systèmes non-conservatifs de la forme

$$\begin{cases} \dfrac{\partial u}{\partial t} + \displaystyle\sum_{i=1}^{d} \frac{\partial}{\partial x_i} f_i(u,\alpha) + \sum_{i=1}^{d} s_i(u,\alpha) \frac{\partial \alpha}{\partial x_i} = 0, \\ \dfrac{\partial \alpha}{\partial t} = 0. \end{cases}$$

posés en dimension d d'espace a été étudiée dans un travail antérieur de N. Seguin [80]. Le propos du cinquième chapitre de cette thèse est d'envisager et d'étendre ce résultat dans le cas de systèmes non-conservatifs comprenant de surcroît un terme source de relaxation, autrement dit de la forme

$$\begin{cases} \dfrac{\partial u}{\partial t} + \displaystyle\sum_{i=1}^{d} \frac{\partial}{\partial x_i} f_i(u,\alpha) + \sum_{i=1}^{d} s_i(u,\alpha) \frac{\partial \alpha}{\partial x_i} = r(u,\alpha), \\ \dfrac{\partial \alpha}{\partial t} = 0. \end{cases}$$

Afin de pouvoir démontrer la stabilité des solutions stationnaires, satisfaisant $r(v,\alpha) = 0$, ce système est supposé disposer d'une entropie convexe partielle. Nous utilisons la notion de solution processus entropique qui généralise celui de solution faible entropique. À partir de là, suivant la méthode de A.E. Tzavaras [91], nous définissons une entropie relative permettant de comparer deux solutions associées à une même donnée géométrique α. La condition de dissipation d'entropie est le point clef permettant d'obtenir la stabilité de certains états stationnaires parmi les solutions processus entropiques.

Contents

Chapter 1

Introduction

The aim of this dissertation is to study the theoretical and numerical analysis of hyperbolic systems of partial differential equations and transport equations, with relaxation terms and boundary conditions.

1.1 Overview of hyperbolic partial differential equations

In many industrial applications, hyperbolic systems arise as basic models and especially in various branches of physics in which finite-speed propagation and conservation laws are involved. The data for the fully hyperbolic first-order partial differential equations (PDEs) include not only initial conditions (governing the so-called Cauchy problem) but also boundary conditions (leading to the so-called initial boundary value problem or IBVP for short). When discussing numerical methods for hyperbolic systems, it is usual to construct difference scheme and provide a theoretical analysis only for the initial value problem since we usually do not know how to calculate the boundary points and how to ascertain whether an algorithm for boundary points is reasonable. However, most of physical phenoma take place in bounded domain under prescribed boundary constraints. The main motivations of present study is the determination of the correct number and kind of boundary conditions that can be imposed to yield a well-posed problem. This work presents a formalism for the treatment of boundary conditions for systems of hyperbolic equations. The central concept of this work is that hyperbolic systems of equations represent the propagation of waves and, at any boundary, some of the waves are propagating into the computational domain while others are propagating out of it [34, 62]. The outward propagating waves have their behavior defined entirely by the solution at and within the boundary, and no boundary conditions can be specified for them. The inward propagating waves depend on the fields exterior to the solution domain and therefore require boundary conditions to complete the specification of their behavior [46, 58]. For a hyperbolic system of equations, considerations on characteristics show that one must be cautious about prescribing the solution on the boundary. In some particular cases, the boundary conditions can be found by physical considerations (such as a solid wall), but their derivation in the general case is not obvious. The problem of finding the suitability of boundary conditions, i.e., those that lead to a well-posed problem, is difficult in general from both the theoretical and practical points of view (proof of well-posedness, choice of the physical variables that can be prescribed). The implementation of these boundary conditions is crucial in practice, but it strongly depends on the problem at hand as shown in Godlewski and Raviart [35]. The theory developed by Kreiss [58] and others [65, 78], known as the Uniform Kreiss Condition (UKC), is one of the earliest works in this area. This theory relies on the analysis of "normal modes", which are introduced by applying a Fourier transformation in the spatial direction normal to the boundary of interest and a Laplace transform in the time variable. The main idea in the derivation of necessary conditions on the boundary data so that the problem is well-posed is to exclude the cases that can lead to an ill-posed problem by looking for particular normal modes that cannot satisfy an energy estimate. The purpose of this section is to give a brief introduction to the basic concept related to hyperbolic systems associated with the boundary conditions and the relaxation terms.

1.1.1 Characteristics and boundary conditions for hyperbolic systems in one space dimension

The prototype for all hyperbolic partial differential equations is the one way wave equation

$$\partial_t u(t,x) + a\partial_x u(t,x) = 0, \quad 0 \leq x \leq L, t \geq 0, \tag{1.1.1}$$

where a is a constant and an interval length $L > 0$. At $t = 0$, we give initial data

$$u(0,x) = f(x).$$

By the method of characteristic, the solution $u(t,x)$ to (1.1.1) is a copy of the original function and depends only the characteristic lines $x - at =$ constant.

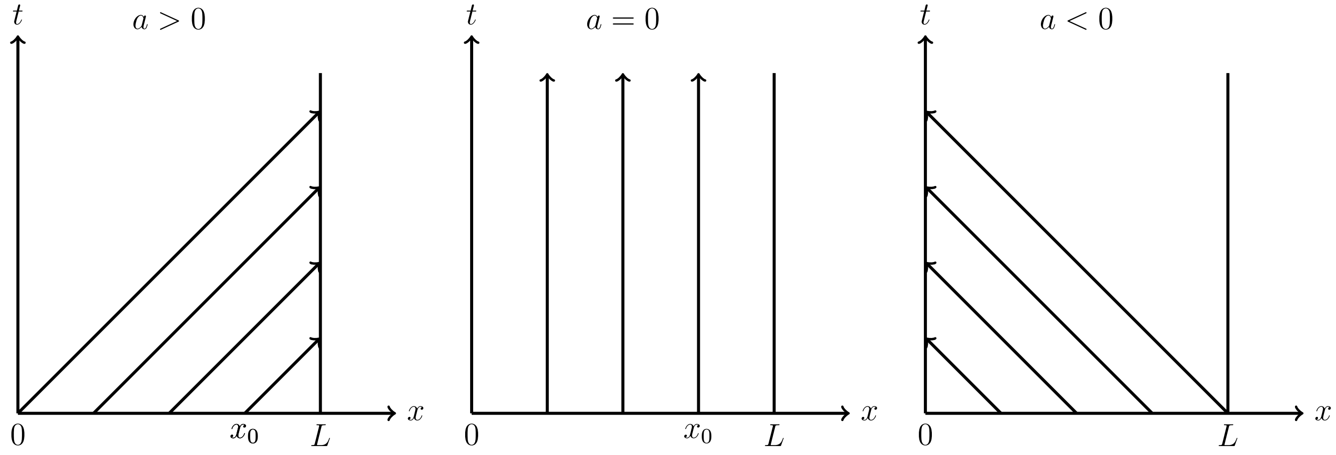

Figure 1.1.1: Characteristics of equation (1.1.1).

If $a > 0$ then the characteristics in this region propagate from the left to the right, as shown in Figure 1.1.1. We can see that the solution is uniquely determined if we give the boundary condition

$$u(t,0) = g(t).$$

Thus, the solution is given by

$$u(t,x) = \begin{cases} f(x-at), & \text{if } x - at > 0, \\ g(t - a^{-1}x), & \text{if } x - at < 0. \end{cases}$$

Along the characteristics given by $x - at = 0$, there will be a jump discontinuity in u if $u_0(0)$ is not equal to $g(0)$.

If $a = 0$ then we do not need any boundary conditions because $\partial_t u(t,x) = 0$ implies

$$u(t,x) = f(x).$$

If $a < 0$ then the solution of our problem is uniquely determined for

$$x \geq 0, \quad t \geq 0, \quad x - at \leq L.$$

To extend the solution for $x - at > L$, we specify a boundary condition at $x = L$:

$$u(t, L) = g_1(t).$$

For the solution to be smooth in the whole domain, it is necessary that $g_1(t)$ and $f(x)$ are smooth functions. It is also necessary that $g_1(t)$ and $f(x)$ are compatible or satisfy compatibility conditions. The most obvious necessary condition is

$$g_1(0) = f(L). \tag{1.1.2}$$

Otherwise, the solution has a jump. In that case, we only obtain a generalized solution. If the condition (1.1.2) is satisfied and $g_1 \in \mathcal{C}^1(\mathbb{R}^+)$, $f \in \mathcal{C}^1([0, L])$ then $u(t, x)$ is Lipschitz continuous. To obtain solutions belonging to $\mathcal{C}^1(\mathbb{R}^+, [0, L])$, we first note that $v(t, x) = \partial_x u(t, x)$ satisfies

$$\begin{cases} \partial_t v(t,x) + a\partial_x v(t,x) = 0, & 0 \leq x \leq L,\, t \geq 0, \\ v(0,x) = f'(x), \\ v(t,L) = (-a)^{-1}\partial_t u(t,L) = (-a)^{-1} g_1'(t). \end{cases}$$

To ensure that v be continuous everywhere, the initial and boundary values must match each other at $(t, x) = (0, L)$. This leads to the condition

$$-af'(L) = g_1'(0). \tag{1.1.3}$$

Secondly, $w(t, x) = \partial_t u(t, x)$ satisfies

$$\begin{cases} \partial_t w(t,x) + a\partial_x w(t,x) = 0, & 0 \leq x \leq L,\, t \geq 0, \\ w(0,x) = -a\partial_x u(0,x) = -af'(x), \\ w(t,L) = g_1'(t) \end{cases}$$

and the condition (1.1.3) also ensures that w is continuous everywhere. Thus, $u(t, x)$ belongs to $\mathcal{C}^1(\mathbb{R}^+, [0, L])$. Higher order derivatives satisfy the same differential equation (1.1.1) and we get higher order regularity by adding more restrictions on higher order derivatives of f and g at $(t, x) = (0, L)$. The same technique can be applied to any problem to ensure that we get smooth solutions. For systems of nonlinear equations, these compatibility conditions may become complicated. The easiest way to satisfy all of them is to require that all initial and boundary data (and forcing functions) vanish near the boundaries at $t = 0$.

Now, we consider a linear system of N conservation laws

$$\begin{cases} \partial_t U(t,x) + A\partial_x U(t,x) = 0, & 0 \leq x \leq L,\, t \geq 0, \\ U(0,x) = f(x), & 0 \leq x \leq L, \end{cases} \tag{1.1.4}$$

where $U(t, x)$ is an $N-$vector of conserved variables and A is an $N \times N$ matrix. We say that a system in the form (1.1.4) is hyperbolic if the matrix A is diagonalizable as

$$A = P\Lambda P^{-1}, \tag{1.1.5}$$

where P is the $N \times N$ matrix whose columns are composed of the eigenvectors of A and $\Lambda = \text{diag}\{\lambda_1, ..., \lambda_N\}$ is a diagonal matrix consisting of the real eigenvalues of A.

By multiplying the first equation in (1.1.4) with P from the left, we can obtain the conservation law as

$$\partial_t W(t,x) + \Lambda \partial_x W(t,x) = 0 \tag{1.1.6}$$

in terms of the variable $W = P^{-1}U$. Let us introduce

$$\begin{aligned}
\Lambda^I &= \operatorname{diag}\{\lambda_1, ..., \lambda_r\} \quad \text{with } \lambda_i > 0,\ 1 \le i \le r,\\
\Lambda^{II} &= \operatorname{diag}\{\lambda_{r+1}, ..., \lambda_{N-s}\} \quad \text{with } \lambda_i < 0,\ r+1 \le i \le N-s,\\
\Lambda^{III} &= 0
\end{aligned}$$

are diagonal matrices. Then, we obtain the system

$$\begin{aligned}
\partial_t W^I(t,x) + \Lambda^I \partial_x W^I(t,x) &= 0,\\
\partial_t W^{II}(t,x) + \Lambda^{II} \partial_x W^{II}(t,x) &= 0,\\
\partial_t W^{III}(t,x) &= 0.
\end{aligned}$$

Using the previous argument, we obtain a unique solution if we specify the initial condition

$$W(0,x) = P^{-1}f(x), \quad 0 \le x \le L$$

and the boundary condition

$$\begin{aligned}
W^I(t,0) &= g^I(t),\\
W^{II}(t,L) &= g^{II}(t).
\end{aligned}$$

With these conditions, the problem decomposes into N scalar problems. We can couple the components by generalizing the boundary conditions to

$$\begin{aligned}
W^I(t,0) &= R_0^{II} W^{II}(t,0) + R_0^{III} W^{III}(t,0) + g^I(t),\\
W^{II}(t,L) &= R_1^I W^I(t,L) + R_1^{III} W^{III}(t,L) + g^{II}(t).
\end{aligned} \tag{1.1.7}$$

Here, R_0^{II}, R_0^{III}, R_1^I and R_1^{III} are rectangular matrices that may depend on t.

It is easy to describe these conditions in geometrical terms. $\Lambda^{III} = 0$ implies that $W^{III}(t,x) = (P^{-1}f)^{III}(x)$. Thus, we need only discuss the influence of the boundary conditions on W^I and W^{II}. The equations in (1.1.7) can be written as

$$\begin{aligned}
&W^I(t,0) = R^{II} W^{II}(t,0) + \widetilde{g}^I(t), \quad W^{II}(t,L) = R^I W^I(t,L) + \widetilde{g}^{II}(t),\\
&\widetilde{g}^I(t) := R_0^{III}(P^{-1}f)^{III}(0) + g^I(t), \quad \widetilde{g}^{II}(t) := R_1^{III}(P^{-1}f)^{III}(L) + g^{II}(t), \quad ,
\end{aligned} \tag{1.1.8}$$

where $R^I := R_1^I$ and $R^{II} := R_0^{II}$. Starting with $t = 0$, the initial values for W^I and W^{II} are transported along the characteristics to the boundaries $x = L$ and $x = 0$, respectively. Using the boundary conditions, these values are transformed into values for $W^I(t,0)$ and $W^{II}(t,L)$, which are then transported along the characteristics to the boundaries $x = L$ and $x = 0$, respectively. Here, the process is repeated (see Figure 1.1.2). Because of these geometrical properties, the components of W are called characteristic variables.

The number of boundary conditions for $x = 0$ is equal to the number of negative eigenvalues of Λ, or, equivalently, the number of characteristics entering the region. Correspondingly, at $x = L$, the number of boundary conditions is equal to the number of positive eigenvalues of Λ. No boundary conditions are required, or may be given, for vanishing eigenvalues.

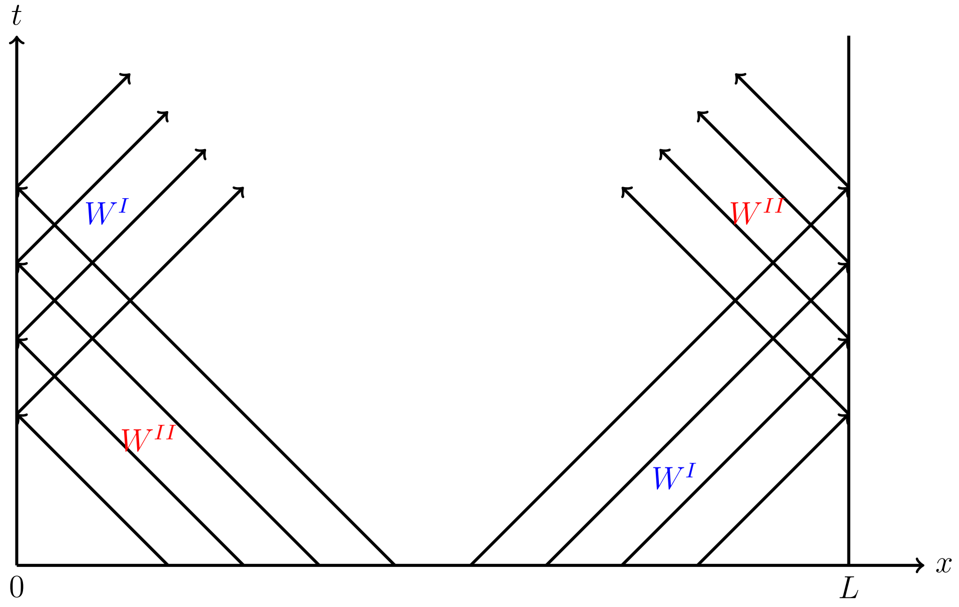

Figure 1.1.2: Characteristics and characteristic variables.

In most applications, the differential equations are given in the nondiagonal form (1.1.4) and boundary conditions are linear relations

$$L_0 U(t,0) = g_0(t), \quad L_1 U(t,L) = g_1(t). \tag{1.1.9}$$

Here,

$$L_0 = \begin{pmatrix} l_{1,1} & \cdots & l_{1,N} \\ \vdots & \vdots & \vdots \\ l_{r,1} & \cdots & l_{r,N} \end{pmatrix}, \quad L_1 = \begin{pmatrix} l_{r+1,1} & \cdots & l_{r+1,N} \\ \vdots & \vdots & \vdots \\ l_{N-s,1} & \cdots & l_{N-s,N} \end{pmatrix}$$

are rectangular matrices whose rank is equal to the number of positive and negative eigenvalues of A, respectively (or better, the number of characteristics that enter the region at the boundary).

If we use the transformation (1.1.5), the differential equations are transformed into (1.1.6) and the boundary conditions become

$$L_0 PW(t,0) = g_0(t), \quad L_1 PW(t,L) = g_1(t). \tag{1.1.10}$$

We again obtain linear relations for the characteristic variables. Our initial boundary value problem can be solved if (1.1.10) can be written in the form (1.1.7). This is equivalent to require L_0 is invertible on the entering (thus rightgoing) characteristic subspace and L_1 is invertible on the entering (thus leftgoing) characteristic subspace. Then, we can solve the relations (1.1.10) for $W^{I}(t,0)$ and $W^{II}(t,L)$, respectively.

1.1.2 The initial boundary value problems for hyperbolic system

We begin by reviewing the theory for the initial boundary value problems. The theory was introduced by Kreiss [58] for strictly hyperbolic systems. It was latter extended in many directions, yielding a huge literature on this subject, see for instance the book [3]. In this section, we focus on a hyperbolic initial boundary value problem in one-dimensional space.

Most of the material in this section can be found in [41, 43, 57]. We now consider the following problem

$$\begin{cases} \partial_t U(t,x) + A\partial_x U(t,x) = F(t,x), & 0 \le x < +\infty,\ t \ge 0, \\ U(0,x) = f(x), & 0 \le x < +\infty, \\ BU(t,0) = b(t), & t \ge 0, \end{cases} \tag{1.1.11}$$

where the unknown $U(t,x)$ is valued in $\mathbb{R}^N$ and A is an $N \times N$ matrix. The functions F, f and b are the interior source term, the initial data and the boundary data, respectively. B is an $p \times N$ matrix. We shall see in a moment that the number p of scalar boundary conditions must equal that of the positive eigenvalues of A. Now, we define well-posedness for homogeneous boundary condition

Definition 1.1.1. *The IBVP (1.1.11) with $F = 0$ and $b = 0$ is well-posed if there exists a unique solution $U(t,x)$ that satisfies*

$$\|U(t,.)\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} \le Ke^{\alpha t}\|f\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)},$$

where K and α are constants independent of f.

Since we would like to estimate U directly in terms of F, f and b, it leads to the following definition

Definition 1.1.2. *The IBVP (1.1.11) is strongly well-posed if there exists a unique solution $U(t,x)$ that satisfies*

$$\|U(t,.)\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} + \int_0^t |U(\tau,0)|^2 d\tau \le K(t)\left(\|f\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} + \int_0^t \left(\|F(\tau,.)\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} + |b(\tau)|^2\right) d\tau\right),$$

where $K(t)$ is a function that is bounded in every finite time interval and independent of F, f, b.

Now, we turn to semi-discretization of the continuous problem (1.1.11) by

$$\begin{cases} \dfrac{d}{dt}U_j(t) + (\mathcal{Q}U)_j(t) = F_j(t), & j \ge 1,\ t \ge 0, \\ U_j(0) = f_j, & j \ge 0, \\ BU_0(t) = b(t), & t \ge 0, \end{cases} \tag{1.1.12}$$

where the difference operator $(\mathcal{Q}U)_j(t)$ is a consistent approximation of the first order space-derivative $A\partial_x U(x_j,t)$ in the sense that $(\mathcal{Q}U)(x_j,t) = A\partial_x U(x_j,t) + \mathcal{O}(\Delta x^k)$, for some $k > 0$. For convenience, we define the grid functions $F_j(t) = F(x_j,t)$ and $f_j = f(x_j)$. The next definition is in analogy with Definition 1.1.1 above

Definition 1.1.3. *Let $\Delta x_0 > 0$. The approximation (1.1.12) with $(F_j)_{j\in\mathbb{N}} \equiv 0$ and $b \equiv 0$ is stable if for all $\Delta x \le \Delta x_0$, the solution $(U_j)_{j\in\mathbb{N}}(t)$ satisfies*

$$\|U(t,.)\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} \le Ke^{\alpha t}\|f\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)},$$

where K and α are constants independent of f.

The same argument as in the case of well-posedness can be used here to extend this notion to general inhomogeneous data in L^2. To do so, b must be differentiable. Corresponding to Definition 1.1.2, we make the following definition

Definition 1.1.4. *The approximation (1.1.12) is strongly stable if it is stable and the estimate*

$$\|U(t,.)\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} \le K(t)\left(\|f\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} + \max_{\tau\in[0,t]}\left(\|F(\tau,.)\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} + |b(\tau)|^2\right)\right)$$

holds. Here, $K(t)$ is a bounded function in any finite time interval and independent of F, f, b.

1.1.2.1 The Uniform Kreiss-Lopatinskii Condition

Let $r_1, \cdots, r_N$ denote a basis of eigenvectors of A associated with eigenvalues $\lambda_1, \cdots, \lambda_N$. For simplicity, we assume that A has nonzero eigenvalues. Up to reordering the eigenvalues, we may assume that

$$\exists\, p \in \mathbb{N}: \quad \lambda_1, \cdots, \lambda_p > 0, \quad \lambda_{p+1}, \cdots, \lambda_N < 0.$$

Then, we decompose the unknown U, the source term F and the initial data f on the basis $\{r_i,\ 1 \leq i \leq N\}$

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

Assuming for simplicity that the solution U is smooth, at least $\mathcal{C}^1$ with respect to (t,x), the system (1.1.11) gives

$$\forall\, 1 \leq i \leq N, \quad \frac{d}{dt}\left(U_i(t, x + \lambda_i t)\right) = F_i(t, x + \lambda_i t).$$

With the influence of the boundary condition at $x = 0$, $BU(t,0) = b(t)$, we need to be more specific on these solutions, meaning

- For $1 \leq i \leq p$, we have two separate cases according to the sign of $x - \lambda_i t$,

$$U_i(t,x) = \begin{cases} f_i(x - \lambda_i t) + \int_0^t F_i(s, x - \lambda_i(t-s))ds, & \text{if } x \geq \lambda_i t, \\ U_i(t - x/\lambda_i, 0) + \int_{t-x/\lambda_i}^t F_i(s, x - \lambda_i(t-s))ds, & \text{if } x \leq \lambda_i t. \end{cases} \tag{1.1.13}$$

- For $p+1 \leq i \leq N$, when $\lambda_i < 0$, we obtain the formula

$$U_i(t,x) = f_i(x - \lambda_i t) + \int_0^t F_i(s, x - \lambda_i(t-s))ds \tag{1.1.14}$$

and the trace of U_i on the boundary $x = 0$ could be entirely determined by the data

$$U_i(t,0) = f_i(|\lambda_i| t) + \int_0^t F_i(s, |\lambda_i|(t-s))ds. \tag{1.1.15}$$

Analyzing the formulas (1.1.13) and (1.1.14), we observe that the solution U is entirely determined provided that we can express the traces of the incoming characteristics $\{U_i(t,0),\ 1 \leq i \leq p\}$ in terms of the data F, f, b. Since the trace of the outgoing characteristic $\{U_i(t,0),\ p+1 \leq i \leq N\}$ are determined by the formula (1.1.14), the boundary condition in (1.1.11) reads

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

where $U_i(t,0)$ in the right hand side are computed in (1.1.15). This is equivalent to the linear relations (1.1.10) for the characteristic variables. Following Section 1.1.1, the IBVP (1.1.11) can be well-posed in any reasonable sense (meaning at least existence and uniqueness of a solution) if and only if B is an $p \times N$ and satisfies the following algebraic condition

$$\mathbb{R}^p = \operatorname{span}\{Br_1, \cdots, Br_p\} \tag{1.1.16}$$

A remarkable result by Kreiss [58] states that for the analogue of (1.1.11) in several space dimensions, well-posedness can still be characterized by an algebraic condition. The latter is usually referred to as the Uniform Kreiss-Lopatinskii Condition. There is however a modification between the one-dimensional case and the multi-dimensional case. In one space dimension, the condition (1.1.16) for well-posedness equivalently reads

$$\operatorname{Ker} B \cap \operatorname{span}(r_1, \cdots, r_p) = \{0\}.$$

1.1.2.2 A necessary condition for well-posedness

In chapters 2 and 3, the system depends on a damping parameter and one may require strong well-posedness independently of this parameter. We will see that the Uniform Kreiss-Lopatinskii Condition is not enough and other tools have to be introduced. Indeed, asymptotic stability has not been defined so far. We begin by the following lemma

Lemma 1.1.5 (The Lopatinskii Condition). *The IBVP* (1.1.11) *with* $F \equiv 0$ *and* $b \equiv 0$ *is not well-posed if we can find a complex number* s *with* $\mathrm{Re}\,(s) > 0$ *and initial data*

$$u(0,x) = \varphi(x), \quad \varphi \in L^2(\mathbb{R}^+, \mathbb{R}^N)$$

such that

$$u(t,x) = e^{st}\varphi(x) \tag{1.1.17}$$

is a solution.

We now give conditions such that solutions of the form (1.1.17) exists. Substituting (1.1.17) into (1.1.11), we get the following eigenvalue problem

$$\begin{cases} s\varphi + A\partial_x \varphi = 0, \quad 0 \le x < +\infty, \\ B\varphi(0) = 0, \\ \varphi \in L^2(\mathbb{R}^+, \mathbb{R}^N). \end{cases} \tag{1.1.18}$$

Let us mention that the analysis based on the Laplace transform and the connected eigenvalue problem is often called normal mode analysis. Then, we have the following lemma

Lemma 1.1.6. *There is a solution of the form* (1.1.17) *if and only if the eigenvalue problem* (1.1.18) *has an eigenvalue* s *with* $\mathrm{Re}\,(s) > 0$.

We now assume that A is nonsingular. The ordinary differential equation in (1.1.18) can be reformulated as

$$\partial_x \varphi = M\varphi, \tag{1.1.19}$$

with $M = -sA^{-1}$. If the eigenvalues of M are distinct and $(\kappa_i)_{1 \le i \le N-p}$ are the eigenvalues of M with $\mathrm{Re}\,(\kappa_i) < 0$ then the general solution to the problem (1.1.19), belonging to $L^2(\mathbb{R}^+, \mathbb{R}^N)$, can be written in the form

$$\varphi(x) = \sum_{j=1}^{N-p} \sigma_j y_j e^{\kappa_j x},$$

where $(y_j)_{1 \le j \le N-p}$ are the eigenvectors satisfying

$$My_j = \kappa_j y_j.$$

If the eigenvalues of M are not distinct then we can still write the general solution to (1.1.19), belonging to $L^2(\mathbb{R}^+, \mathbb{R}^N)$, in the form

$$\varphi(x) = \sum_j \varphi_j(x) e^{\kappa_j x},$$

where $\varphi_j(x)$ are polynomials in x with vector coefficients containing altogether $N-p$ parameters σ_j.

Substituting the expression $\varphi(x)$ into the boundary condition in (1.1.18), we obtain a linear system of equation for $\sigma = (\sigma_1, \cdots, \sigma_{N-p})$ in the form

$$C(s)\sigma = 0. \tag{1.1.20}$$

Then, the following theorem holds true.

Theorem 1.1.7. *The IBVP (1.1.11) is not well-posed if for some s with* $\mathrm{Re}\,(s) > 0$*,*

$$det(C(s)) = 0.$$

For the difference approximation IBVP (1.1.12), to derive the necessary condition for the stability, we follow the same lines as for the well-posedness of the continuous IBVP. More details are given in [41]. The test for stability is given by the following lemma

Lemma 1.1.8. *The approximation (1.1.12) with $(F_j)_{j\in\mathbb{N}} \equiv 0$ and $b \equiv 0$ is not stable if we can find a complex number s with* $\mathrm{Re}\,(s) > 0$ *and initial data*

$$U_j(0) = \varphi_j, \quad (\varphi_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^N)$$

such that

$$U_j(t) = e^{st}\varphi_j \tag{1.1.21}$$

is a solution.

Substituting (1.1.21) into (1.1.12), the eigenvalue problem associated with our approximation is

$$\begin{cases} s\varphi + (\mathcal{Q}\varphi)_j = 0, & j \geq 1,\ \mathrm{Re}\,(s) > 0, \\ B\varphi_0 = 0, \\ (\varphi_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^N). \end{cases} \tag{1.1.22}$$

By using the same argument as for the continuous problem, we can obtain the following Godunov-Ryabenkii condition, which is analogous to the Lopatinskii condition for hyperbolic initial boundary value problem

Lemma 1.1.9 (The Godunov-Ryabenkii condition)**.** *The approximation (1.1.12) is not stable if the eigenvalue problem (1.1.22) has an eigenvalue s with* $\mathrm{Re}\,(s) > 0$.

The general solution of the first equation of the problem (1.1.22) with $(\varphi_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^N)$ has the form

$$\varphi_j = \sum_{|\kappa_\nu|<1} P_\nu(j)\kappa_\nu^j(s), \quad \mathrm{Re}\,(s) > 0 \tag{1.1.23}$$

where $P_\nu(j)$ is a polynomial in j with vector coefficients. The solution $(\varphi_j)_{j\in\mathbb{N}}$ depends on Np parameters $\sigma_\sharp = [\sigma_1, \cdots, \sigma_{Np}]$. Substituting (1.1.23) into the boundary condition in (1.1.22), one yields a system of equation

$$C_\sharp(s)\sigma_\sharp = 0$$

and we can rephrase Lemma 1.1.9 in the following form

Lemma 1.1.10. *The Godunov-Ryabenkii condition is satisfied if and only if*

$$\det(C_\sharp(s)) \neq 0, \quad \text{for}\,\mathrm{Re}\,(s) > 0.$$

1.1.2.3 Stability in the generalized sense for hyperbolic system

A well-posed problem requires not only stability but also the existence of a unique smooth solution for given smooth data. The solution $U(t,x)$ to the IBVP (1.1.11) can be constructed by using the Laplace transform. Furthermore, uniqueness follows by linearity. However, in order to obtain a smooth solution, we not only need smooth initial-boundary and forcing functions but also a certain compatibility between these functions. With the proper normalization, the Kreiss condition is equivalent to

$$\det(C(s)) \neq 0, \quad \operatorname{Re} s \geq 0. \tag{1.1.24}$$

Note that $C(s)$ must always be defined for $\operatorname{Re} s = 0$ as a limit when s is approaching the imaginary axis from the right. It is natural to define stability differently when using the Laplace transform method. We now consider the following definition

Definition 1.1.11. *(i) The IBVP (1.1.11) with $f \equiv 0$ and $b \equiv 0$ is stable in the generalized sense if the solution $U(t,x)$ satisfies the following estimate*

$$\int_0^{+\infty} e^{-2\eta t} \|U(t,.)\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} dt \leq K(\eta) \int_0^{+\infty} e^{-2\eta t} \|F(t,.)\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} dt$$

for all $\eta > \eta_0$. Here, η_0 and $K(\eta)$ are constants independent of F and

$$\lim_{\eta \to +\infty} K(\eta) = 0.$$

(ii) The IBVP (1.1.11) with $f \equiv 0$ is strongly stable in the generalized sense if the following estimate

$$\int_0^{+\infty} e^{-2\eta t} \|U(t,.)\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} dt \leq K(\eta) \int_0^{+\infty} e^{-2\eta t} \left(\|F(t,.)\|^2_{L^2(\mathbb{R}^+,\mathbb{R}^N)} + |b(t)|^2 \right) dt \tag{1.1.25}$$

holds.

We have introduced the definition of stability in the generalized sense for the continuous case. The same concept will be used for the semi-discrete approximation (1.1.12). Corresponding to the analytic case, we make the following definition

Definition 1.1.12. *(i) The approximation (1.1.12) with $(f_j)_{j\in\mathbb{N}} \equiv 0$ and $b \equiv 0$ is stable in the generalized sense if for all sufficiently small Δx, the solution $(U_j)_{j\in\mathbb{N}}(t)$ satisfies*

$$\int_0^{+\infty} e^{-2\eta t} \|U(t)\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} dt \leq K(\eta) \int_0^{+\infty} e^{-2\eta t} \|F(t)\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} dt$$

for all $\eta > \eta_0$. Here, η_0 and $K(\eta)$ are constants independent of F and

$$\lim_{\eta \to +\infty} K(\eta) = 0.$$

(ii) The approximation (1.1.12) with $(f_j)_{j\in\mathbb{N}} \equiv 0$ is strongly stable in the generalized sense if the following estimate

$$\int_0^{+\infty} e^{-2\eta t} \|U(t)\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} dt \leq K(\eta) \int_0^{+\infty} e^{-2\eta t} \left(\|F(t)\|^2_{\ell^2(\mathbb{N},\mathbb{R}^N)} + |b(t)|^2 \right) dt \tag{1.1.26}$$

holds.

1.2 The linear damped wave equation in a quarter plane

In many industrial applications, models are based on hyperbolic partial differential equations which involve source terms. One of the main features of these models is related to the notation of dissipation, leading to smooth solutions and asymptotic stability. The most classical model is the linear damped wave equation, which is a particular case of the Jin-Xin relaxation model [51]. We consider here the following initial boundary value problem (IBVP) in the quarter plane

$$\begin{cases} \partial_t U(t,x) + A\partial_x U(t,x) = \varepsilon^{-1} SU(t,x), \quad x > 0,\, t > 0, \\ U(0,x) = f(x), \quad x > 0 \\ BU(t,0) = b(t), \quad t > 0, \end{cases} \tag{1.2.1}$$

where the relaxation time $\varepsilon > 0$ may be introduced to characterize the stiffness of the relaxation. Let $a > 0$ and $(B_u, B_v) \in \mathbb{R}^2$, we also define

$$U(t,x) = \begin{pmatrix} u(t,x) \\ v(t,x) \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ a & 0 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} B_u & B_v \end{pmatrix}.$$

Due to the stiff source term, the relaxation mechanism of the damped wave equation is a highly singular process and its dissipative mechanism is not guaranteed. In order for the asymptotic stability to hold, i.e, solution of the damped wave equation tending to the corresponding equilibrium as the rate of relaxation goes to zero, certain stability conditions have to be satisfied. In the case of the Cauchy problem, the most well-known is the sub-characteristic condition [64, 94]. The corresponding IBVP is much more difficult and much less is known [93, 97]. In order for the IBVP to be well-posed for a fixed rate of relaxation, the boundary condition has to satisfy the Uniform Kreiss Condition (UKC)

$$B_u + \sqrt{a} B_v \neq 0. \tag{UKC}$$

However, the sub-characteristic condition and UKC are not enough to obtain asymptotic stability and a more stringent restriction has to be imposed on the structure of the boundary condition. Indeed, Xin and Xu [96] show that a complex coupling between the sub-characteristic condition and UKC appears. This leads to the so-called the Stiff Kreiss Condition (SKC) [96]

$$B_v = 0, \quad \text{or } \frac{B_u}{B_v} \notin [-\sqrt{a}, 0] \tag{SKC}$$

which is a necessary and sufficient for the well-posedness of the IBVP for relaxation systems of balance laws, uniformly with respect to the rate of relaxation. The SKC is derived through a simplified normal mode analysis and its explicit form is obtained by the conformal mapping theorem. Firstly, Xin and Xu consider the simpler homogeneous initial data. Under the assumption of the SKC, the solution of this IBVP can be constructed by the method of Laplace transform. The stiff well-posedness can be proved rather directly in [96]. Secondly, by the asymptotic expansion, the limiting equilibrium solution and various boundary layer behaviours are identified. In case the initial data is nonzero, by linearity, Xin and Xu break up the full IBVP into two simpler problem, one with homogeneous initial condition and the other with homogeneous condition. The extra complication is the holding of the boundary terms that enter when doing the integrating or summation by parts.

One of our main motivations is to study the notion of stiff stability for numerical approximations by finite differences of the IBVP for relaxation systems of balance laws. Our main interest is to apply a classical numerical scheme to the IBVP and to derive necessary conditions for stiff stability. Due to the effects of the boundary layers and the interactions of the boundary and initial layer, numerical schemes have to be properly designed in order to provide accurate approximations and consistent behaviours. Indeed, numerical boundary layers may appear and disturb the results of the continuous case. We mainly consider the semi-discrete or full discrete approximations with central or upwind schemes. For high order approximations, the physical boundary conditions are not enough to define the complete difference scheme and there is a need for numerical boundary conditions. Thus, how to formulate boundary conditions for relaxation systems to guarantee the uniform stability and to minimize or localize the artificial boundary layers are crucial to the success of relaxation schemes. Within the framework of the difference scheme in space, we propose two methods of discretization of Dirichlet boundary condition. The first is the techniques of summation by part (SBP) [42] and the second is the concept of transparent boundary condition (TBC) [2, 5, 20]. We will discuss the advantages and disadvantages of these methods in terms of simplicity of implementation and consistency with the SKC.

1.2.1 Results in Chapter 2: A stiffly stable discrete scheme for the damped wave equation using summation by parts method

In Chapter 2, we consider the continuous-in-time or time-implicit scheme with central or upwind approximations in space. The boundary is approximated using a summation by parts method and the stiff stability is proved using energy estimates and the Laplace transform.

The research on SBP operators was originally driven by applications in flow problems including turbulence and wave propagation. The general procedure for deriving higher order SBP difference operators for hyperbolic problems was first presented by Kreiss and Scherer [71, 72]. This was the start for a large number of papers on this topic. SBP operators were later developed and used for multidimensional problems including non-orthogonal grids and for implicit difference approximations, see [73, 74, 83]. Later generalizations, including differential equations containing both first and second order derivatives and wave equation in second order form, are given in [66, 67, 68].

We first study the semi-discrete approximation of the IBVP (1.2.1), which is the following system

$$\begin{cases} \partial_t U_j(t) + (\mathfrak{Q}U)_j(t) = \varepsilon^{-1} S U_j(t), & j \geq 1,\, t \geq 0, \\ U_j(0) = f_j, & j \geq 0, \\ BU_0(t) = b(t), & t \geq 0. \end{cases} \tag{1.2.2}$$

By using the technique of SBP for the central differencing scheme, the considered operator reads

$$(\mathfrak{Q}U)_j = \frac{1}{2\Delta x} A\left(U_{j+1} - U_{j-1}\right), \quad j \geq 0 \tag{1.2.3}$$

with $U_{-1} = 2U_0 - U_1$. To obtain the energy estimate, we now define the following scalar product and norm

$$\langle U, V\rangle_{\Delta x} = \frac{\Delta x}{2}\langle U_0, V_0\rangle + \Delta x \sum_{j=1}^{+\infty} \langle U_j, V_j\rangle, \quad \|U\|_{\Delta x}^2 = \langle U, U\rangle_{\Delta x}$$

with $\langle ., .\rangle$ being the usual Euclidean inner product. Let us introduce a symmetric positive definite matrix

$$H = \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}.$$

Since HA is also a symmetric matrix, one has

$$\langle U, H(\mathcal{Q}U)\rangle_{\Delta x} = -\frac{1}{2}\langle U_0, HAU_0\rangle$$

which is the discrete counterpart of the equality

$$\int_0^{+\infty} \langle U, HA\partial_x U\rangle(t,x)dx = -\frac{1}{2}\langle U, HAU\rangle(t,0)$$

available in the continuous case. Let us emphasize that the numerical scheme (1.2.2) still needs one more scalar equation at the boundary point $j = 0$ so as to be fully defined, due to the fact that the matrix B has rank one. This is actually a discrete feature only, since in the continuous case this single equation is exactly complemented by the only incoming characteristic (under UKC). We propose the following numerical

$$\Gamma\Big(\partial_t U_0(t) + (\mathcal{Q}U)_0(t)\Big) = \varepsilon^{-1}\Gamma S U_0(t), \quad t \geq 0, \tag{1.2.4}$$

where $\Gamma = \begin{pmatrix} -aB_v & B_u \end{pmatrix}$. In Section 2.2, we derive a sufficient condition for stability for the numerical scheme (1.2.2)-(1.2.4). The main result is

Theorem 1.2.1 (Theorem 1.2 in [8]). *Under the strict dissipativity condition*

$$B_u B_v > 0, \tag{1.2.5}$$

for any $T > 0$ there exists $C_T > 0$ such that for any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and any $b \in \mathcal{C}^1(\mathbb{R}^+, \mathbb{R}) \cap L^2(\mathbb{R}^+, \mathbb{R})$, the solution $(U_j)_{j\in\mathbb{N}}$ to (1.2.2)-(1.2.4) satisfies

$$\int_0^T |U_0(t)|^2\, dt + \int_0^T \sum_{j\geq 0} \Delta x |U_j(t)|^2\, dt \leq C_T \left(\sum_{j\geq 0} \Delta x |f_j|^2 + \int_0^T |b(t)|^2\, dt\right), \tag{1.2.6}$$

where the constant C_T is independent of the data f and b, but most importantly of $\varepsilon \in (0, +\infty)$ and $\Delta x \in (0, 1]$.

In [96], Xin and Xu show that the IBVP (1.2.1) is well-posed if and only if the SKC is satisfied. In the discrete IBVP (1.2.2)-(1.2.4), it seems that even the SKC is not sufficient to derive uniform stability estimates. In comparison, the strict dissipativeness condition (1.2.5) is more restrictive, but we are only able to prove that it is sufficient. Following [100] and [96], we also perform a normal mode analysis to construct unstable solutions and, based on some numerical investigations, the condition (1.2.5) would appear to be also necessary for the stiff stability.

Secondly, we focus in Section 2.3 on the time-implicit scheme of the IBVP (1.2.1), which will be precised explicitly by the following system

$$\begin{cases} U_j^{n+1} - U_j^n + \dfrac{\Delta t}{2\Delta x} A\left(U_{j+1}^{n+1} - U_{j-1}^{n+1}\right) = \dfrac{\Delta t}{\varepsilon} S U_j^{n+1}, & j \geq 1,\ n \geq 0, \\ U_j^0 = f_j, & j \geq 0, \\ BU_0^n = b^n, & n \geq 0, \\ \Gamma\left(U_0^{n+1} - U_0^n\right) + \dfrac{\Delta t}{\Delta x}\Gamma A\left(U_1^{n+1} - U_0^{n+1}\right) = \dfrac{\Delta t}{\varepsilon}\Gamma S U_0^{n+1}, & n \geq 0. \end{cases} \tag{1.2.7}$$

Our aim is to determine a sufficient condition for the stiff stability of the fully discrete scheme (1.2.7), in order words the uniform stability with respect to the stiffness of the relaxation term. The main result in Section 2.3 is

Theorem 1.2.2. *Assume that $(B_u, B_v) \in \mathbb{R}^2$ satisfies the strict dissipativity condition (1.2.5). Then, for any $T > 0$, there exists a constant $C_T > 0$ such that for all $\Delta t > 0$ and any positive constant $\delta_{xt} \le 3\sqrt{a}/8$ together with $\Delta x = \delta_{xt}\Delta t$, any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, any $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R})$, the solution $\left(U_j^n\right)_{j\in\mathbb{N}}$ to the scheme (1.2.7) satisfies*

$$\sum_{n=0}^{N}\sum_{j\ge 0}\Delta x\Delta t\left|U_j^n\right|^2 + \sum_{n=0}^{N}\Delta t\left|U_0^n\right|^2 \le C_T\left(\sum_{j\ge 0}\Delta x\left|f_j\right|^2 + \sum_{n=0}^{N}\Delta t\left|b^n\right|^2\right), \tag{1.2.8}$$

where $N := T/\Delta t$ and C_T is independent of $\varepsilon \in (0, +\infty)$.

Finally, in Section 2.4, we study the time-implicit scheme of the IBVP (1.2.1) with the upwind differencing scheme in space.

$$\begin{cases} U_j^{n+1} - U_j^n + \Delta t(\mathcal{Q}U)_j^{n+1} = \varepsilon^{-1}\Delta t S U_j^{n+1}, & j \ge 1,\ n \ge 0,\\ U_j^0 = f_j, & j \ge 0,\\ BU_0^n = 0, & n \ge 0,\\ \Gamma\left(U_0^{n+1} - U_0^n\right) + \Delta t\Gamma(\mathcal{Q}U)_0^{n+1} = \Delta t\varepsilon^{-1}\Gamma S U_0^{n+1}, & n \ge 0. \end{cases} \tag{1.2.9}$$

The considered operator $(\mathcal{Q}U)_{j\ge 0}$ is defined by

$$(\mathcal{Q}U)_j = \frac{1}{2\Delta x}\left((A - \sqrt{a}I)U_{j+1} + 2\sqrt{a}U_j - (A + \sqrt{a}I)U_{j-1}\right), \quad j \ge 0 \tag{1.2.10}$$

with $(A + \sqrt{a}I)U_{-1} = 2AU_0 - (A - \sqrt{a}I)U_1$. By means of the discrete energy method, we can prove the following statement:

Proposition 1.2.3. *Assume that the parameters $\Delta x \in\ (0, 1]$, $\varepsilon > 0$ and (B_u, B_v) satisfies the discrete strict dissipativity condition*

$$2a\frac{B_u}{B_v} + \frac{\Delta x}{\varepsilon}\left(\frac{B_u}{B_v}\right)^2 > 0. \tag{1.2.11}$$

Then, there exists a constant $C > 0$ such that for all $\Delta t > 0$ and any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, the solution $(U_j^n)_{j\in\mathbb{N}}$ to scheme (1.2.9) satisfies

$$\langle U^n, HU^n\rangle_{\Delta x} + C\Delta t\sum_{k=0}^{n}\left|U_0^k\right|^2 \le \langle f, Hf\rangle_{\Delta x}, \quad n \in \mathbb{N}. \tag{1.2.12}$$

More precisely,

a) If $B_uB_v > 0$ then (1.2.12) holds uniformly, i.e. with C independent of ε and Δx.

b) If $B_uB_v < 0$ then considering some $\delta_0 > -2aB_vB_u^{-1}$, there exists $C(\delta_0) > 0$ such that (1.2.12) holds uniformly with $C = C(\delta_0)$, as soon as $\Delta x \ge \delta_0\varepsilon$.

Let us mention that the discrete strict dissipativity condition (1.2.11) is not implied by the SKC, probably due to some numerical diffusion at the boundary. In the case homogeneous initial condition $(f_j)_{j\in\mathbb{N}} \equiv 0$ and nonzero boundary condition $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R})$, we have difficulty finding a sufficient condition for the stiff stability of the time-implicit scheme (1.2.9), but we postpone its possibility to a further work. In Appendix B, we study an example of how waves occur in modeling the action of an elastic string over time, which is a particular case of linear damped wave equation. By using the Newton's Second Law of Motion, we can derive the boundary condition $B_uB_v > 0$ for this system.

1.2.2 Results in Chapter 3: A stiffly stable fully discrete scheme for the damped wave equation using discrete transparent boundary condition

In Chapter 2, the boundary is approximated using a summation-by-parts method. Using energy estimates and Laplace transforms, the discrete approximation of the IBVP (1.2.1) is proved to be stiffly stable if the boundary condition satisfies $B_uB_v > 0$, which is only the subclass of the SKC. In this chapter, we consider the discrete transparent boundary technique to construct a stiffly stable boundary condition. The technique and its analysis has been proposed by Arnold and Ehrhardt in [2]. Besse, Noble and co-authors apply the tools to dispersive problems [4, 5, 54]. We also refer the reader to [39] for non-reflecting methods in the context of wave problems. The recent work of Coulombel [20] proposes a systematic study of transparent boundary conditions for evolution problems. Our aim here is to prove that the SKC derived in [96] is a sufficient condition for the stiff stability of the time-implicit scheme, which obtained by the the central differencing scheme in space. The discrete transparent boundary condition is designed at $j = 0$, such that the time-implicit scheme for the IBVP (1.2.1) is stable and its solution coincides with the solution of the whole space problem $\{j \in \mathbb{Z}\}$ restricted to $\{j \in \mathbb{N}\}$. By using the concept of transparent boundary condition, at $j = 0$, we define

$$U_{-1}^n = \sum_{k=0}^{n} \mathcal{C}_{n-k} U_0^k,$$

where the coefficients $\mathcal{C}_k$ will be precised explicitly in the forthcoming Definition 3.2.4. Let us mention that in our case, the values $\mathcal{C}_k$ are designed in the case of homogeneous initial data and are kept unchanged in the case of nonzero initial data.

To summarize, we study the following fully discrete approximation of the IBVP (1.2.1)

$$\begin{cases} \dfrac{U_j^{n+1} - U_j^n}{\Delta t} + \dfrac{1}{2\Delta x} A \left(U_{j+1}^{n+1} - U_{j-1}^{n+1} \right) = \dfrac{1}{\varepsilon} S U_j^{n+1}, & j \geq 1,\ n \geq 0, \\ U_j^0 = f_j, & j \geq 0, \\ BU_0^n = b^n, & n \geq 0, \\ \dfrac{1}{\Delta t} \Gamma \left(U_0^{n+1} - U_0^n \right) + \dfrac{1}{2\Delta x} \Gamma A \left(U_1^{n+1} - \displaystyle\sum_{k=0}^{n+1} \mathcal{C}_{n+1-k} U_0^k \right) = \dfrac{1}{\varepsilon} \Gamma S U_0^{n+1}, & n \geq 0. \end{cases} \tag{1.2.13}$$

By applying the $\mathcal{Z}-$transform with respect to time index $n \in \mathbb{N}$, which is discrete analogue of the Laplace transform in time $t \in \mathbb{R}^+$, we have the following theorem

Theorem 1.2.4 (Theorem 1.1 in [7]). *Assume that $(B_u, B_v) \in \mathbb{R}^2$ satisfies the SKC. Let $\delta_{xt} \leq 3\sqrt{a}/8$ be a positive number. For any $T > 0$, there exists a constant $C_T > 0$ such that for all $\Delta t > 0$ and $\Delta x = \delta_{xt} \Delta t$, any $(f_j)_{j \in \mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and $(b^n)_{n \in \mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R})$, the solution $(U_j^n)_{j \in \mathbb{N}}$ to the scheme (1.2.13) satisfies*

$$\sum_{n=0}^{N} \sum_{j \geq 0} \Delta x \Delta t |U_j^n|^2 + \sum_{n=0}^{N} \Delta t |U_0^n|^2 \leq C_T \left(\sum_{j \geq 0} \Delta x |f_j|^2 + \sum_{n=0}^{N} \Delta t |b^n|^2 \right), \tag{1.2.14}$$

where $N := T/\Delta t$ and C_T is independent of $\varepsilon \in (0, +\infty)$.

It seems that the SKC is also a sufficient condition to guarantee the uniform stability of the scheme (1.2.13) independently of the stiffness of the source term, of the space step and of the time step.

1.2.3 Illustration of numerical in stability

In this paragraph, we perform some numerical experiments to discuss the advantages and disadvantages of two methods (SBP and TBC) with the SKC. As main parameters for the experiments, we choose $a = 1$, $B_v = 1$, $\delta_{xt} = 1/3$, the space step $\Delta x = 10^{-2}$, the time step $\Delta t = \delta_{xt}^{-1}\Delta x$, the relaxation rate $\varepsilon = 10^{-2}$ and the boundary data B_u vary. The initial data is the homogeneous one $(f_j)_{j\in\mathbb{N}} \equiv 0$. The boundary data is

$$b(t) = \frac{t}{2}\sin(t).$$

We now consider the numerical solutions $(U_j^n)_{j\in\mathbb{N}}$ to the numerical schemes (1.2.7) and (1.2.13) over the time interval $t \in [0, 1.2)$ with $B_u = 1$ (see Figure 1.2.1) and $B_u = -4$ (see Figure 1.2.2).

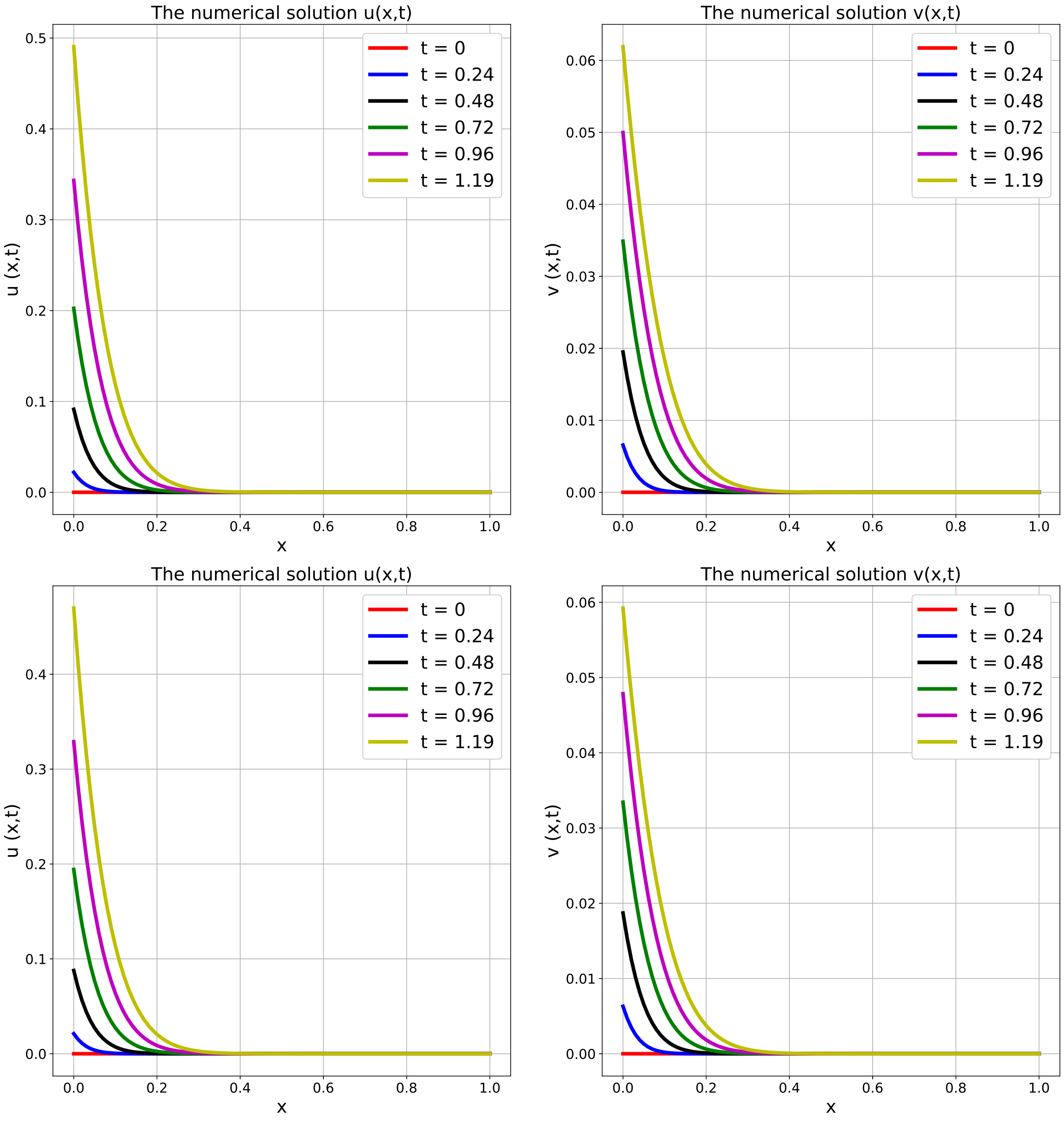

Figure 1.2.1: The numerical solution $U(x, t)$ for the numerical schemes (1.2.7) (top) and (1.2.13) (bottom). The strict dissipativity condition (1.2.5) and the SKC hold with $B_u = 1$.

Clearly, we can see that

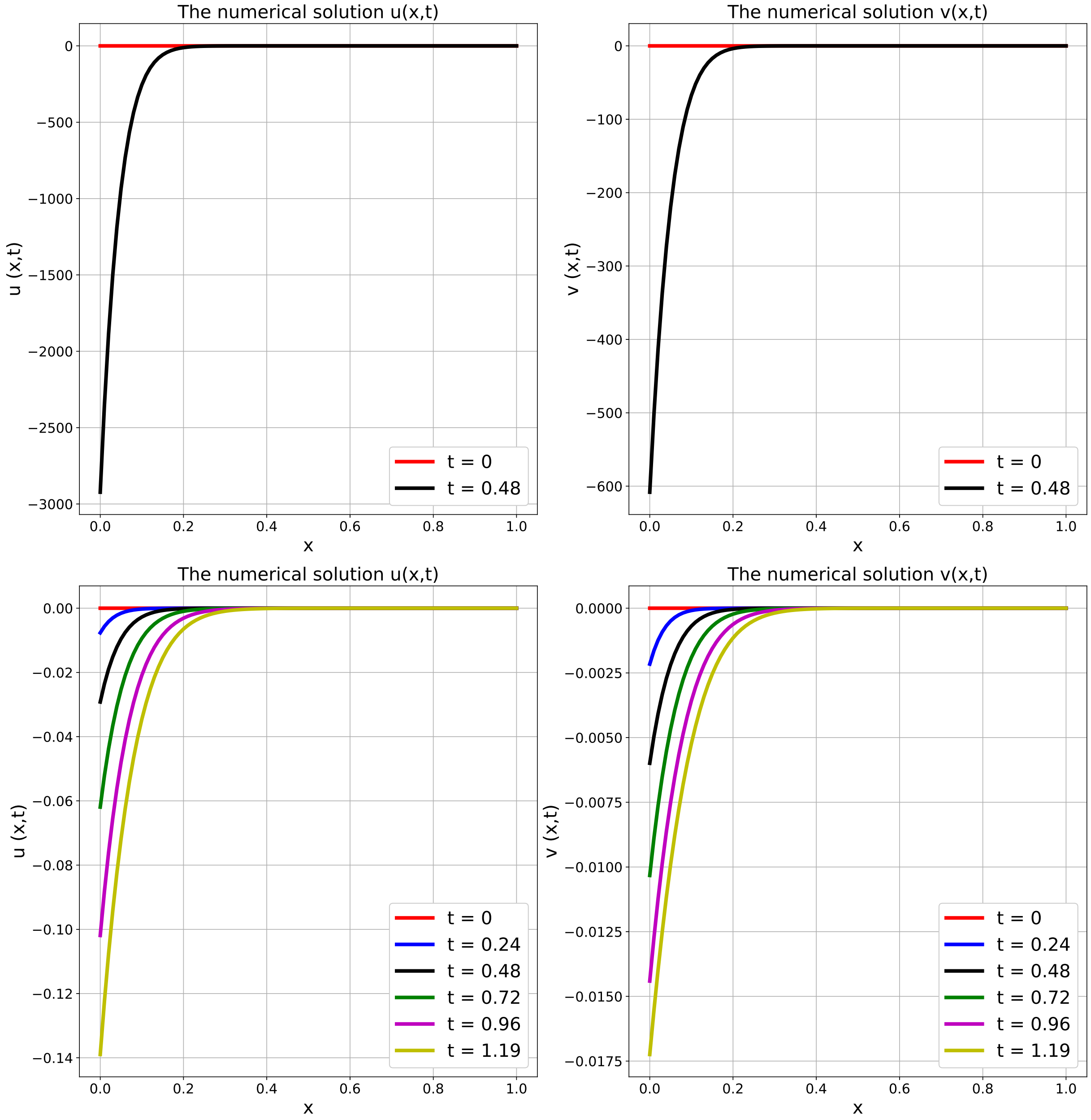

Figure 1.2.2: The numerical solution $U(x, t)$ for the numerical schemes (1.2.7) (top) and (1.2.13) (bottom). With $B_u = -4$, the strict dissipativity condition (1.2.5) fails and the SKC is valid.

- In the case $B_u B_v > 0$, at the boundary $x = 0$, the numerical solution to both of the numerical schemes (1.2.7) and (1.2.13) go up slightly. For instance, this is the case in Figure 1.2.1 with $(B_u, B_v) = (1, 1)$.

- In the case $B_u/B_v < -\sqrt{a}$, the numerical solution at the boundary $x = 0$ to the numerical scheme (1.2.7) increase suddenly. However, in the case of the numerical scheme (1.2.13), the incoming solution at $x = 0$ go slowly. This is the case for $(B_u, B_v) = (-4, 1)$, see Figure 1.2.2.

In spite of the fact that the result of the transparent boundary technique is better than the SBP method, we still have difficulty in implementing the process transparent boundary condition at $x = 0$. In our numerical approximation, the coefficients $\mathcal{C}_k$ are obtained by means of numerical quadrature. Besides, because the extra boundary condition determines U_{-1}^{n+1} as a linear function of U_0^k for the past step times only: $0 \leq k \leq n + 1$, the transparent boundary technique takes more time than the SBP method to approximate the numerical solution.

1.3 The transport equations on an interval

In Section 1.1.1, we have seen that transport equations do not require prescription of any boundary condition at an outflow boundary, that is, when the transport velocity is outgoing with respect to the boundary of the spatial domain. This can be understand by integrating the equation along the characteristics (see Section 1.1.1). However, many discretizations of the transport equation involve a stencil that includes cells of the numerical grid that are located in the downstream region. Such discretizations necessitate the prescription of numerical boundary conditions at an outflow boundary [42, 85], even though the underlying partial differential operator does not require any boundary condition for determining the solution.

1.3.1 Motivation

The transport equation on the half line with Dirichlet boundary condition has been studied, for dissipative schemes, by Kreiss and Lundqvist [59], and we here consider the equation on an interval. We are thus given a fixed constant velocity $a > 0$, an interval length $L > 0$ and we consider the (continuous) problem

$$\begin{cases} \partial_t u + a\,\partial_x u \; = \; 0\,, & t \geq 0\,, \quad x \in (0,L)\,, \\ u(0,x) \; = \; f(x)\,, & x \in (0,L)\,, \\ u(t,0) \; = \; g(t)\,, & t \geq 0\,, \end{cases} \tag{1.3.1}$$

with, at least, $u_0 \in L^2((0,L),\mathbb{R})$. By the method of characteristics, the solution to (1.3.1) is given by the explicit representation formula

$$\forall\,(t,x) \in \mathbb{R}^+ \times (0,L)\,, \quad u(t,x) \; = \; \begin{cases} f\,(x - a\,t)\;, & \text{if } x \geq a\,t\,, \\ g\left(t - \dfrac{x}{a}\right)\;, & \text{if } x \leq a\,t\,, \end{cases} \tag{1.3.2}$$

where it is understood in (1.3.2) that the initial condition f has been extended by zero to $\mathbb{R}^-$ (no extension is need on $(L,+\infty)$ since a is positive and therefore $x - at < L$ for all relevant values of t and x).

Our goal is to provide a thorough treatment of nonzero incoming boundary data and to design numerical boundary conditions that recover the optimal rate of convergence in the maximum norm (at least, the same rate of convergence as the one in [22] for zero boundary data).

1.3.2 Results in Chapter 4: High order numerical schemes for transport equations on bounded domains

High order, stable and convergent discretizations for the problem (1.3.1) have been analyzed in [22] in the case of homogeneous incoming boundary conditions ($g = 0$, in that case smoothness of the solution u is equivalent to f being flat at 0). The goal of Chapter 4 is to extend this result to the case of non-homogeneous incoming boundary conditions by developing a systematic construction and stability analysis for numerical boundary conditions arising from the so-called inverse Lax-Wendroff method, see e.g, [32, 86, 92].

It may seem a too much trivial problem to approximate the problem (1.3.1) for which an explicit solution is given, but one should keep in mind that such a representation formula ceases to be available for hyperbolic systems in several space dimensions, and our goal is to develop analytical tools which do not rely on the fact that (1.3.1) is a one-dimensional scalar problem. We therefore consider from now on an approximation of (1.3.1) by means of a finite difference scheme. We are given a positive integer J, that is meant to be large, and define the space step Δx and the grid points $(x_j)_{j\in\mathbb{Z}}$ by

$$\Delta x := \frac{L}{J}, \quad x_j := j\,\Delta x \quad (j \in \mathbb{Z}).$$

The time step Δt is then defined as $\Delta t := \lambda\,\Delta x$, where $\lambda > 0$ is a constant that is fixed so that Assumption 1.3.1 below is satisfied. The interval $(0, L)$ is divided in J cells (x_{j-1}, x_j) with $j = 1, \dots, J$, but considering the whole real line $\{j \in \mathbb{Z}\}$ will be useful in some parts of the analysis. We use from now on the notation $t^n := n\,\Delta t$, $n \in \mathbb{N}$, the quantity u_j^n will play the role of an approximation for the solution u to (4.1.1) at the time t^n on the cell (x_{j-1}, x_j). We do not wish to discriminate between finite difference or finite volume schemes for (1.3.1), so rather than deriving this or that type of numerical scheme, we consider a linear iteration for the u_j^n that reads in the interior domain:

$$u_j^{n+1} = \sum_{\ell=-r}^{p} a_\ell\, u_{j+\ell}^n\,, \quad n \in \mathbb{N}\,, \quad j = 1, \dots, J\,. \tag{1.3.3}$$

In (1.3.3), r, p are fixed non-negative integers, and the coefficients a_ℓ, $\ell = -r, \cdots, p$ may only depend on the ratio λ and the velocity a. Most of the usual linear explicit schemes, such as the upwind, Rusanov, Lax- Friedrichs and Lax-Wendroff schemes, can be put in that form.

Before describing the numerical boundary conditions we enforce for (1.3.3), let us state our main and in fact only assumption on the coefficients in (1.3.3).

Assumption 1.3.1 (Consistency and stability without any boundary)**.** *The coefficients $a_{-r}, \dots, a_p$ in* (4.1.7) *satisfy $a_{-r}\, a_p \neq 0$ (normalization), and for some integer $k \geq 1$, there holds:*

$$\forall\, m = 0, \dots, k\,, \quad \sum_{\ell=-r}^{p} \ell^m\, a_\ell = (-\lambda\, a)^m\,, \quad \text{(consistency of order } k), \tag{1.3.4}$$

$$\sup_{\theta\in[0,2\pi]} \left| \sum_{\ell=-r}^{p} a_\ell\, e^{i\,\ell\,\theta} \right| \leq 1\,, \quad (\ell^2\text{-stability on } \mathbb{Z}). \tag{1.3.5}$$

If the interior cells of the grid are labeled, as in (1.3.3), by $j \in \{1, \cdots, J\}$, the numerical approximation of (1.3.1) requires, for passing from one time index n to the next, prescribing r numerical boundary conditions on the left of the interval (that is, close to $x = 0$), and p numerical boundary conditions on the right (that is, close to $x = L$). In other words, we need to prescribe the value of the approximate solution (u_j^n) in the ghost cells located at the boundary of the interior domain. For simplicity, and in order to be consistent with the continuous problem (1.3.1), we prescribe Dirichlet homogeneous boundary conditions in conjunction with (1.3.3) on the left of the interval $(0, L)$:

$$u_\ell^n = \sum_{\kappa=0}^{k-1} \frac{\Delta x^\kappa}{(\kappa+1)\,!\,(-a)^\kappa} \left(\ell^{\kappa+1} - (\ell-1)^{\kappa+1}\right) g^{(\kappa)}(t^n)\,, \quad n \in \mathbb{N}\,, \quad 1 - r \leq \ell \leq 0\,. \tag{1.3.6}$$

The strategy is not new and is now referred to as the *inverse Lax-Wendroff method.* It consists, as detailed below, in writing Taylor expansions with respect to the space variable x close to the incoming boundary and then using the advection equation (1.3.1) to substitute the normal derivatives $\partial_x^m u(t, 0)$ for tangential derivatives $\partial_t^m u(t, 0)$, the latter being computed thanks to the boundary conditions in (1.3.1). On the right of the interval $(0, L)$, there is nothing to be done if $p = 0$, that is, in the case of an upwind discretization, for in that case, given the vector $(u_1^n, \cdots, u_J^n)$, the vector $(u_1^{n+1}, \cdots, u_J^{n+1})$ is entirely determined by (1.3.6) and (1.3.3), so we can iterate the scheme (1.3.3), (1.3.6) starting from some initial data $(u_1^0, \cdots, u_J^0)$ to any positive time level n. We therefore assume from now on $p \geq 1$, which is the interesting case where the numerical discretization of (1.3.1) necessitates an outflow numerical boundary condition while the continuous problem does not "obviously" provide with one. n this article, we shall prescribe Neumann type numerical boundary conditions (these are called extrapolation numerical boundary conditions in [36]). For ease of writing, we introduce the difference operator in space which acts on vectors $(v_j)_{j=1-r,\cdots,J+p}$ as follows:

$$\forall\, j = 2-r, \cdots, J+p, \quad (D_- v)_j \;:=\; v_j - v_{j-1}\,.$$

Higher order difference operators D_-^m, $m \geq 2$, are defined accordingly by iterating D_-. Then given a fixed integer $k_b \in \mathbb{N}$ (b stands for "boundary"), we prescribe the following numerical boundary condition in conjunction with (1.3.3):

$$(D_-^{k_b} u^n)_{J+\ell} \;=\; 0\,, \quad n \in \mathbb{N}\,, \quad \ell = 1, \dots, p\,. \tag{1.3.7}$$

The scheme (1.3.3), (1.3.6), (1.3.7) is initialized with the piecewise constant projection of the initial condition for (1.3.1), that is, for the interior cells:

$$\forall\, 1 \leq j \leq J\,, \quad u_j^0 \;:=\; \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x\,. \tag{1.3.8}$$

Our main result is the following convergence estimate for the scheme (1.3.3), (1.3.6), (1.3.7) supplemented with the initial condition (1.3.8), which is the extension of the main result in [22] to the case of nonzero boundary data.

Theorem 1.3.2 (Main convergence result [9])**.** *Let $a > 0$, $k \in \mathbb{N}^*$ and $k_b \in \mathbb{N}$. Under Assumption 1.3.1, there exists a constant $C > 0$ such that for any final time $T \geq 1$, any integer $J \in \mathbb{N}^*$, any data $f \in H^{k+1}((0,L))$ and $g \in H^{k+1}((0,T))$ satisfying the compatibility requirements at $t = x = 0$:*

$$\forall\, m = 0, \dots, k\,, \quad f^{(m)}(0) \;=\; (-a)^{-m}\, g^{(m)}(0)\,,$$

the solution (u_j^n) to (1.3.3), (1.3.6), (1.3.7), (1.3.8) satisfies:

$$\sup_{0 \leq n \leq T/\Delta t}\ \sup_{1 \leq j \leq J} \left| u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} u(t^n, x)\,\mathrm{d}x \right| \leq C\, T\, e^{C\,T/L}\, \Delta x^{\min(k,k_b)-1/2} \left(\|f\|_{H^{k+1}((0,L))} + \|g\|_{H^{k+1}((0,T))} \right), \tag{1.3.9}$$

with u the exact solution to (1.3.1), whose expression is given by (1.3.2).

Actually, the constant C in (1.3.9) is independent of $L \geq 1$, which is consistent with the convergence result we shall prove below for the half-space problem on $\mathbb{R}^+$ with inflow at $x = 0$. Following [22], we shall prove Theorem 1.3.2 by using a stability estimate for (1.3.3), (1.3.6), (1.3.7) and a superposition argument, which amounts to considering separately two half-space problems: one in which there is only inflow at $x = 0$, and one for which there is only outflow at $x = L$.

Theorem 1.3.3 (Convergence estimate for the inflow problem [9])**.** *Let $a > 0$, $k \in \mathbb{N}^\star$. Under Assumption 1.3.1, there exists a constant $C > 0$ such that for any final time $T \geq 1$, any $J \in \mathbb{N}^*$, any initial condition $f \in H^{k+1}((0,+\infty))$ and boundary source term $g \in H^{k+1}((0,T))$ satisfying the compatibility conditions*

$$\forall\, 0 \leq m \leq k\,, \quad f^{(m)}(0) \;=\; (-a)^{-m}\, g^{(m)}(0)\,, \tag{1.3.10}$$

the solution $(u_j^n)_{j\geq 1-r, n\in\mathbb{N}}$ *to the numerical scheme*

$$\begin{cases} u_j^0 \, = \, \dfrac{1}{\Delta x} \displaystyle\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x\,, & j \geq 1\,, \\ u_\ell^n \, = \, \displaystyle\sum_{\kappa=0}^{k-1} \dfrac{\Delta x^\kappa}{(\kappa+1)\,!\,(-a)^\kappa} \left(\ell^{\kappa+1} - (\ell-1)^{\kappa+1}\right) g^{(\kappa)}(t^n)\,, & 0 \leq n \leq T/\Delta t\,, \quad 1-r \leq \ell \leq 0\,, \\ u_j^{n+1} \, = \, \displaystyle\sum_{\ell=-r}^{p} a_\ell\, u_{j+\ell}^n\,, & 0 \leq n \leq T/\Delta t - 1\,, \quad j \geq 1\,, \end{cases} \tag{1.3.11}$$

satisfies

$$\sup_{0\leq n\leq T/\Delta t} \left(\sum_{j\leq J} \Delta x \left(u_j^n - \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} u(t^n,x)\,\mathrm{d}x \right)^2 \right)^{1/2} \leq \, C\,T\;\Delta x^k \left(\|f\|_{H^{k+1}((0,+\infty))} + \|g\|_{H^{k+1}((0,T))} \right),$$

where u *is the exact solution to the half-line transport problem*

$$\begin{cases} \partial_t u + a\,\partial_x u \, = \, 0\,, & t \in (0,T)\,,\ x \geq 0\,, \\ u(0,x) \, = \, f(x)\,, & x \geq 0\,, \\ u(t,0) \, = \, g(t)\,, & t \in (0,T)\,. \end{cases} \tag{1.3.12}$$

Theorem 1.3.4 (Convergence estimate for the outflow problem [22]). *Let* $a > 0$, $k \in \mathbb{N}^\star$ *and* $k_b \in \mathbb{N}$. *Under Assumption 1.3.1, there exists a constant* $C > 0$ *such that for any final time* $T \geq 1$, *for any* $J \in \mathbb{N}^*$ *and for any initial condition* $f \in H^{k+1}((-\infty,L))$, *the solution* $(u_j^n)_{j\leq J, 0\leq n\leq T/\Delta t}$ *to the numerical scheme*

$$\begin{cases} u_j^0 \, = \, \dfrac{1}{\Delta x} \displaystyle\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x\,, & j \leq J\,, \\ (D_-^{k_b} u^n)_{J+\ell} \, = \, 0\,, & 0 \leq n \leq T/\Delta t\,, \quad 1 \leq \ell \leq p\,, \\ u_j^{n+1} \, = \, \displaystyle\sum_{\ell=-r}^{p} a_\ell\, u_{j+\ell}^n\,, & 0 \leq n \leq T/\Delta t - 1\,, \quad j \leq J\,, \end{cases}$$

satisfies

$$\sup_{0\leq n\leq T/\Delta t} \left(\sum_{j\leq J} \Delta x \left(u_j^n - \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} f(x - a\,t^n)\,\mathrm{d}x \right)^2 \right)^{1/2} \leq \, C\,T\,\Delta x^{\min(k,k_b)}\, \|f\|_{H^{k+1}((-\infty,L))}\,.$$

As in [22], the loss of 1/2 in the rate of convergence of Theorem 1.3.2 looks somehow artificial and is mostly a matter of passing from the $\ell_n^\infty \ell_j^2$ topology to $\ell_{n,j}^\infty$. Our next result examines a situation where the optimal convergence rate $\min(k, k_b)$ can be obtained. In order to characterize the discrete steady states, it is natural to introduce the characteristic polynomial

$$A(X) := \sum_{\ell=-r}^{p} a_\ell X^{\ell+r} - X^r. \tag{1.3.13}$$

From the consistency property in Assumption 1.3.1, any constant sequence is a discrete steady solution for (1.3.3), the same property being available for the continuous model (namely, the transport operator). However, the discrete nature of the differentiation operator involved in the numerical scheme (1.3.3) allows the existence of many other discrete steady solutions. The latter play an important role when considering then the half-space problem with some discrete boundary conditions. To make the analysis more intelligible, we will work under the following assumption, which was already present in [6].

Assumption 1.3.5. *The characteristic polynomial A defined in (1.3.13) has a unique root (equal to 1) on the unit circle $\mathbb{S} = \{z \in \mathbb{C},\ |z| = 1\}$. In other words, we assume*

$$\bigcup_{\sigma=1}^{\tau} \{\kappa_\sigma\} \cap \mathbb{S} = \{1\}. \tag{1.3.14}$$

In other words, the technical Assumption 1.3.5, which is verified on many examples such as the Lax-Wendroff and $O3$ schemes, allow to recover the optimal rate $k_b = \min(k_b, k)$ in the case $k_b < k$. Of course, one would also like to improve the rate $\min(k_b, k) - 1/2$ in the case $k_b = k$, which is clearly the most natural choice. However, in that case, both the interior and boundary consistency errors scale like Δx^k and, in the framework of Assumption 1.3.1, stability in the interior domain is available only in the ℓ^2_j topology, so it is quite difficult to derive the convergence rate k in the ℓ^∞_j topology. In order to simplify the proof of Theorem 1.3.6, we only examine here the case of a half-space with extrapolation outflow conditions. The extension of the techniques to the case of an interval is left to the interested reader.

Theorem 1.3.6 (Optimal rate of convergence for the outflow problem [9])**.** *Let $a > 0$, $k \in \mathbb{N}^*$ and $k_b \in \mathbb{N}$. Under Assumption 1.3.1 and under the additional Assumption 1.3.5, there exists a constant $C > 0$ such that for any final time $T \geq 1$, any integer $J \in \mathbb{N}^*$, any data $f \in H^{k+1}((-\infty, L))$, the solution to the scheme*

$$\begin{cases} u_j^0 = \dfrac{1}{\Delta x} \displaystyle\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x\,, & j \leq J\,, \\ (D_-^{k_b} u^n)_{J+\ell} = 0\,, & 0 \leq n \leq T/\Delta t\,, \quad 1 \leq \ell \leq p\,, \\ u_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, u_{j+\ell}^n\,, & 0 \leq n \leq T/\Delta t - 1\,, \quad j \leq J\,, \end{cases} \tag{1.3.15}$$

satisfies the error estimate

$$\sup_{0 \leq n \leq T/\Delta t}\ \sup_{j \leq J} \left| u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x - a\, t^n)\,\mathrm{d}x \right| \leq C\, T\, \Delta x^{k_b}\, \|f\|_{H^{k+1}((0,L))}\,,$$

as long as $k_b < k$.

Theorem 1.3.6 already indicates that combining the approach of [22] with other techniques (here, boundary layer expansions) may improve some results. We hope to deal with the case $k_b = k$ in the future.

1.4 Hyperbolic conservation laws with stiff relaxation terms and entropy

1.4.1 Strong, weak, and entropy weak solution

Hyperbolic systems of conservation laws with relaxation appear in the study of a variety of physical phenomena of great practical importance such as thermally non equilibrium fluid flows, non reacting two-phase fluid flows composed of solid particles suspended in gas, viscoelasticity,... The relaxation terms are source terms whose effect is the relaxation to zero of some algebraic quantity, namely the relaxation variables. For instance, in the case of two-phase fluid flows composed of solid particles in gas, the relaxation terms model the drag force whose effect is the relaxation to zero of the relative velocity between the two phases. In the case of thermally non equilibrium fluid flows, the relaxation term depend upon the different temperatures involved in the modeling: here the thermal equilibrium of the flow is characterized by a single temperature. A relaxation time may be introduced to characterize the stiffness of the relaxation. In the case of the two-phase fluid flows considered above the relaxation time is the drag time which is proportional to inverse of the the square of the radius of the particles. In the case of thermally non equilibrium flows the relaxation time depends on the heat exchanges. Therefore, we consider a system of N conservation laws

$$\partial_t u(t,x) + \sum_{j=1}^{d} \partial_j f_j(u)(t,x) = r(u)(t,x). \tag{1.4.1}$$

System (1.4.1) is set on the whole space $x \in \mathbb{R}^d$ and for any time $t \in [0,T)$, $T > 0$. The notation ∂_j denotes the partial derivative with respect to x_j. We assume that there exists a convex bounded set of $\mathbb{R}^N$, denoted by Ω and called set of the admissible states such that

$$u(t,x) \in \Omega, \quad \forall\, (t,x) \in [0,T) \times \mathbb{R}^d. \tag{1.4.2}$$

We also assume for all $j \in \{1,...,d\}$, the function $f_j : \mathbb{R}^N \to \mathbb{R}^N$ belong to $\mathcal{C}^2(\overline{\Omega}, \mathbb{R}^N)$ and be such that $\partial_u f_j$ are diagonalizable with real eigenvalues, where ∂_u denotes the differential with respect to the variables u. System (1.4.1) is complemented with the initial condition

$$u(0,x) = u_0(x), \quad \forall\, x \in \mathbb{R}^d. \tag{1.4.3}$$

System (1.4.1) is endowed with a uniformly convex entropy $\eta \in \mathcal{C}^2(\overline{\Omega}, \mathbb{R})$ such that there exists $\beta_1 \geq \beta_0 > 0$ so that

$$\operatorname{spec}\left(\partial_u^2 \eta(u)\right) \subset [\beta_0, \beta_1], \quad \forall\, u \in \overline{\Omega}, \tag{1.4.4}$$

and the corresponding entropy flux $F \in \mathcal{C}^2(\overline{\Omega}, \mathbb{R}^d)$ satisfies for all $j \in \{1, \cdots, d\}$

$$\partial_u F_j(u) = \partial_u \eta(u) \partial_u f_j(u), \quad \forall\, u \in \Omega. \tag{1.4.5}$$

Despite it is well-known that even for smooth initial data u_0, the solution of (1.4.1)-(1.4.3) may develop discontinuities after a finite time, our study is restricted to the approximation of smooth solutions $u \in W^{1,\infty}\left(\mathbb{R}^+ \times \mathbb{R}^d, \Omega\right)$ to (1.4.1)-(1.4.3). Such solutions are called strong solutions, and they satisfy the conservation of the entropy

$$\eta(u) + \sum_{j=1}^{d} \partial_j F_j(u) = \Sigma(u), \tag{1.4.6}$$

where $\Sigma(u) = \partial_u \eta(u) \cdot r(u)$ with the notation $\cdot$ is the scalar product in the same space.

Assuming that $u_0 \in L^\infty\left(\mathbb{R}^d, \Omega\right)$, a function $u \in L^\infty\left(\mathbb{R}^+ \times \mathbb{R}^d, \Omega\right)$ is said to be a weak solution (1.4.1)-(1.4.3) if, for all $\varphi \in \mathcal{C}_c^1(\mathbb{R}^+ \times \mathbb{R}^d, \mathbb{R}^n)$, one has

$$\int\int_{\mathbb{R}^d \times \mathbb{R}^+} u \partial_t \varphi dx\, dt + \int_{\mathbb{R}^d} u_0 \varphi(0,.) dx + \int\int_{\mathbb{R}^d \times \mathbb{R}^+} \sum_{j=1}^{d} f_j(u) \partial_j \varphi dx\, dt = -\int\int_{\mathbb{R}^d \times \mathbb{R}^+} r(u) \varphi dx\, dt. \tag{1.4.7}$$

Moreover, u is said to be an entropy weak solution to (1.4.1)-(1.4.3) if u is a weak solution, i.e., u satisfies (1.4.7), and if, for all $\Psi \in \mathcal{C}_c^1(\mathbb{R}^+ \times \mathbb{R}^d, \mathbb{R}^+)$, it satisfies

$$\int\int_{\mathbb{R}^d \times \mathbb{R}^+} \eta(u) \partial_t \Psi dx\, dt + \int_{\mathbb{R}^d} \eta(u_0) \Psi(0,.) dx + \int\int_{\mathbb{R}^d \times \mathbb{R}^+} \sum_{j=1}^{d} F_j(u) \partial_j \Psi dx\, dt \geq -\int\int_{\mathbb{R}^d \times \mathbb{R}^+} r(u) \Psi dx\, dt. \tag{1.4.8}$$

1.4.2 Relative entropy

To compare two entropy weak solution, we employ the notion of the relative entropy [24, 30]. We define the relative entropy and entropy-fluxes among two solutions by

$$\begin{aligned} h(u,v) &= \eta(u) - \eta(v) - \partial_u \eta(v) \cdot (u-v), \\ q_j(u,v) &= F_j(u) - F_j(v) - \partial_u \eta(v) \cdot (f_j(u) - f_j(v)), \quad \forall j \in \{1, \cdots d\}. \end{aligned}$$

Now, let us consider an admissible weak solution u of (1.4.1) and a constant vector $v \in \mathbb{R}^N$. After some calculations, one obtains

$$\partial_t h(u,v) + \sum_{i=1}^{d} \partial_i \Big(F_i(u) - \partial_u \eta(v) \cdot f_i(u) \Big) \leq \Sigma(u) - \partial_u \eta(v) \cdot r(u). \tag{1.4.9}$$

A natural assumption is the entropy dissipation condition, see [11, 69, 99], namely, for every $u, v \in \Omega$ with $r(v,.) = 0$,

$$\Big(\partial_u \eta(u) - \partial_u \eta(v) \Big) \cdot r(u) \leq 0.$$

If we assume that the system (1.4.1) is entropy dissipative, and integrate the inequality (1.4.9) for $x \in \mathbb{R}^d$, the divergence term disappears and we have

$$\frac{d}{dt} \int_{\mathbb{R}^d} h(u,v) dx \leq 0.$$

Indeed, it is easy to check that

$$\frac{\beta_0}{2} |u-v|^2 \leq h(u,v) \leq \frac{\beta_1}{2} |u-v|^2 \tag{1.4.10}$$

where $|\cdot|$ is the Euclidian norm of $\mathbb{R}^N$. Thus, one obtains the following estimate for any $r > 0$

$$\int_{|x|<r} |v(t,x) - u(t,x)|^2 dx \leq C \int_{|x<|r+L_f t} |v_0(x) - u_0(x)|^2 dx,$$

where a constant $C > 0$ and the quantity L_f is given by

$$L_f = \sup_{j \in \{1, \cdots, d\}} \sup_{(u,v) \in \Omega^2} \sup_{w \in \mathbb{R}^N \setminus \{0\}} \left| \frac{w^T \partial_u^2 \eta(v) \partial_u f_j(u) w}{w^T \partial_u^2 \eta(v) w} \right|.$$

1.4.3 Results in Chapter 5: Stability of stationary solution for non-linear relaxation balance laws

Stability of stationary solutions of singular systems of balance laws have been analyzed in [80]. In chapter 5, we extend this result to the case of nonlinear relaxation balance laws. We first define an entropy process solution, which generalizes the concept of entropy weak solutions. After that, we construct the relative entropy to compare two states. Therefore, we are able to prove the stability of some stationary states within entropy process solutions.

We consider non-conservation systems in d space dimensions of the form

$$\begin{cases} \partial_t u(t,x) + \sum_{i=1}^{d} \partial_i f_i(u,\alpha)(t,x) + \sum_{i=1}^{d} s_i(u,\alpha)(t,x) \partial_i \alpha(t,x) = r(u,\alpha)(t,x), \\ \partial_t \alpha(t,x) = 0. \end{cases} \tag{1.4.11}$$

where

$$\begin{aligned} &u : \mathbb{R}^+ \times \mathbb{R}^d \to \Omega, && f_j : \Omega \times \mathbb{R} \to \mathbb{R}^N, \\ &\alpha : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}, && s_i : \Omega \times \mathbb{R} \to \mathbb{R}^N, \\ &r : \Omega \times \mathbb{R} \to \mathbb{R}^N. \end{aligned}$$

An initial condition is associated with (1.4.11)

$$\begin{cases} u(0,x) &= u_0(x), \\ \alpha(0,x) &= \alpha(x), \end{cases} \text{for } x \in \mathbb{R}^d. \tag{1.4.12}$$

The second equation in (1.4.11) means that α is time-independent. If α is smooth, the third term of the left-hand side of (1.4.11) can be considered as a source term. However, the analysis of our work also applies to non-smooth α, and the term $\sum_{i=1}^d s_i(u,\alpha)\partial_i\alpha$ is a non-conservative product. System (1.4.11) is endowed with an entropy pair (η, F), which depends on (u,α) and satisfies the following assumptions

(H1) The function $\eta = \eta(u,\alpha) \in \mathcal{C}^2(\Omega \times \mathbb{R}, \mathbb{R})$ is convex with respect to its first variable and there exist two positive constants $\beta_0 < \beta_1$ such that

$$\sigma\left(\partial_u^2\eta\right) \subset [\beta_0, \beta_1], \quad \text{on } \Omega \times \mathbb{R}, \tag{1.4.13}$$

where σ denotes the matrix spectrum.

(H2) There exists an entropy flux $F = (F_i(u,\alpha))_{1\le i\le d}$ such that

$$\forall 1 \le i \le d, \quad \partial_u\eta\partial_u f_i = \partial_u F_i \quad \text{and} \quad \partial_u\eta(\partial_\alpha f_i + s_i) = \partial_\alpha F_i.$$

To investigate the stability of system (1.4.11), we assume a natural assumption, which is the following entropy dissipation condition:

(H3) For every $u, v \in \Omega$ with $r(v,.) = 0$,

$$\Big(\partial_u\eta(u,.) - \partial_u\eta(v,.)\Big) \cdot r(u,.) \le 0$$

Instead of using the assumption (H3), in [91], Tzavaras studied the following condition

(H3′) For every $u, v \in \Omega$ with $r(v,\cdot) = 0$, there exists $\gamma > 0$ such that

$$-\Big(\partial_u\eta(u,\cdot) - \partial_u\eta(v,\cdot)\Big) \cdot \Big(r(u,\cdot) - r(v,\cdot)\Big) \ge \gamma|u-v|^2,$$

which may be more restrictive than the entropy dissipation condition (H3). However, it is the key to prove the asymptotic stability of system (1.4.11).

In this work, we extend the analysis in Section 1.4.2 to systems of the form (1.4.11). For a given α, we are able to compare an entropy process solution to some particular stationary solutions. We first assume that the products $s_i\partial_i\alpha$ can be described by means of vector-valued Radon measures $\mu_i \in \mathcal{M}\left(\mathbb{R}^+ \times \mathbb{R}^d \times (0,1)\right)^N$[1] which satisfy at least the following properties

[1] More precisely, $\mathcal{M}(X)$ denotes the set of locally bounded Radon measures on a set X, i.e. $\mathcal{M}(X) = (\mathcal{C}_c(X))'$.

(P1) On any open set $B = B_t \times B_x \subset \mathbb{R}^+ \times \mathbb{R}^d$ such that $\alpha \in W^{1,\infty}(B_x)$, the measures $\mu_i, 1 \leq i \leq d$, satisfy

$$\forall \varphi \in \mathcal{C}_c^\infty(B), \forall 1 \leq i \leq d, \int_0^1 \int_B \varphi d\mu_i(t,x,\lambda) = \int_0^1 \int_B \varphi s_i(\nu,\alpha)\partial_i \alpha dt dx d\lambda.$$

(P2) For any component $1 \leq k \leq N$ and any dimension index $1 \leq i \leq d$,

$$s_i^{(k)} \equiv 0 \Rightarrow \mu_i^{(k)} \equiv 0.$$

Let us now define the entropy process solution

Definition 1.4.1. *Let $u_0 \in BV(\mathbb{R}^d,\Omega)^N$, $\alpha \in BV(\mathbb{R}^d)$ and $T > 0$. A function $\nu \in L^\infty([0,T) \times \mathbb{R}^d \times (0,1),\Omega)$ is an entropy process solution of the Cauchy problem (1.4.11)-(1.4.12) if there exists $(\mu_i)_{1\leq i \leq d} \subset \mathcal{M}(\mathbb{R}^+ \times \mathbb{R}^d \times (0,1))$ satisfying assumptions (P1) and (P2) such that, for all $\varphi \in \mathcal{C}_c^\infty([0,T) \times \mathbb{R}^d)$,*

$$\begin{aligned}-\int_0^1\int_0^T\int_{\mathbb{R}^d}\left(\nu(t,x,\lambda)\partial_t\varphi + \sum_{i=1}^d f_i(\nu,\alpha)\partial_i\varphi\right)dxdtd\lambda + \int_0^1\int_0^T\int_{\mathbb{R}^d}\varphi d\mu(t,x,\lambda)\\ -\int_{\mathbb{R}^d}u_0(x)\varphi(0,x)dx = \int_0^1\int_0^T\int_{\mathbb{R}^d} r(\nu,\alpha)\varphi dxdtd\lambda,\end{aligned} \tag{1.4.14}$$

and, for all non-negative $\varphi \in \mathcal{C}_c^\infty([0,T) \times \mathbb{R}^d)$,

$$\begin{aligned}-\int_0^1\int_0^T\int_{\mathbb{R}^d}\left(\eta(\nu,\alpha)\partial_t\varphi + \sum_{i=1}^d F_i(\nu,\alpha)\partial_i\varphi\right)dxdtd\lambda - \int_{\mathbb{R}^d}\eta(u_0,\alpha)(x)\varphi(0,x)dx\\ \leq \int_0^1\int_0^T\int_{\mathbb{R}^d}\Sigma(\nu,\alpha)\varphi dxdtd\lambda.\end{aligned} \tag{1.4.15}$$

Let us remark that from an entropy weak solution $u(t,x)$ to problem (1.4.11)-(1.4.12), one may easily construct an entropy process solution to problem (1.4.11)-(1.4.12) by setting $\nu(t,x,\lambda) = u(t,x)$ for any $\lambda \in (0,1)$. Reciprocally, if ν is an entropy process solution to problem (1.4.11)-(1.4.12) such that there exists $u \in L^\infty([0,T) \times \mathbb{R}^d)$ such that $\nu(t,x,\lambda) = u(t,x)$ for a.e. $(t,x,\lambda) \in [0,T) \times \mathbb{R}^d \times (0,1)$, then u is an entropy weak solution to problem (1.4.11)-(1.4.12). We now define the relative entropy and entropy-fluxes between two solutions (ν,α) and (v,β)

Definition 1.4.2. *Let $\nu, v \in \Omega$. The relative entropy of ν with respect to v is defined by*

$$\begin{aligned}h:\ \Omega\times\Omega\times\mathbb{R} &\to \mathbb{R}^+\\ (\nu,v,\alpha) &\mapsto \eta(\nu,\alpha) - \eta(v,\alpha) - \partial_u\eta(v,\alpha)\cdot(\nu - v)\end{aligned} \tag{1.4.16}$$

and the corresponding relative entropy fluxes $q : \Omega \times \Omega \times \mathbb{R} \to \mathbb{R}^d$ are

$$q_i(\nu,v,\alpha) = F_i(\nu,\alpha) - F_i(v,\alpha) - \partial_u\eta(v,\alpha)\cdot\Big(f_i(\nu,\alpha) - f_i(v,\alpha)\Big), \quad \forall i \in \{1,...,d\}. \tag{1.4.17}$$

For a given $\alpha \in \mathcal{C}^1(\mathbb{R}^d,\mathbb{R})$, consider a smooth, and thus entropy conservative, entropy process solution ν of (1.4.11), and a time-independent function v. After some calculations, we have

$$\begin{aligned}\partial_t h(\nu,v,\alpha) = &-\sum_{i=1}^d \partial_i\left[F_i(\nu,\alpha) - \partial_u\eta(v,\alpha)\cdot f_i(\nu,\alpha)\right] + \Sigma(\nu,\alpha) - \partial_u\eta(v,\alpha)\cdot r(\nu,\alpha)\\ &-\sum_{i=1}^d \partial_i\left[\partial_u\eta(v,\alpha)\right]\cdot f_i(\nu,\alpha) + \partial_u\eta(v,\alpha)\cdot\sum_{i=1}^d s_i(\nu,\alpha)\partial_i\alpha.\end{aligned}$$

Since the two last terms are not in conservation form, we need to add some assumptions on v in order to make them vanishing. Thus, for any given constant vector $H_0 \in \mathbb{R}^N$, we introduce $\mathcal{S}(H_0)$, the set of $(v, \alpha) \in \Omega \times \mathbb{R}$ such that

(S1) $\partial_u \eta(v, \alpha) = H_0$.

(S2) For all $1 \leq i \leq d$ and $1 \leq k \leq N$, $H_0^{(k)} s_i^{(k)} \equiv 0$.

Let us mention that the above discussion on the smooth solution is extended to entropy process solutions by the theorems 1.4.3 and 1.4.4. Firstly, we use the entropy dissipation condition (H3) to state the following stability result

Theorem 1.4.3. *Let $H_0 \in \mathbb{R}^N$ and consider the set $\mathcal{S}(H_0)$ defined by (S1)-(S2), assumed to be nonempty. Consider $\alpha \in BV(\mathbb{R}^d)$ and a function $v \in BV(\mathbb{R}^d, \Omega)$ such that $(v, \alpha) \in \mathcal{S}(H_0)$ almost everywhere and satisfy the entropy dissipation condition (H3). Then, v is a stationary entropy process solution of system (1.4.11).*

Moreover, let $T > 0$, $u_0 \in BV(\mathbb{R}^d, \Omega)^N$, and $\nu \in L^\infty\big([0, T) \times \mathbb{R}^d \times (0, 1), \Omega)\big)$ an associated entropy process solution. Then, there exists a positive constant L_f, independent of ν, v and α such that the following nonlinear stability property holds for all $R > 0$ and for almost every $t \in [0, T]$

$$\int_0^1 \int_{B(0,R)} h(\nu(t, x, \lambda), v(x), \alpha(x)) dx d\lambda \leq \int_{B(0,R+L_f t)} h(u_0(x), v(x), \alpha(x)) dx. \qquad (1.4.18)$$

Secondly, we consider the assumption (H3′) to prove the following asymptotic stability

Theorem 1.4.4. *Let $H_0 \in \mathbb{R}^N$ and consider the set $\mathcal{S}(H_0)$ defined by (S1)-(S2), assumed to be nonempty. Consider $\alpha \in BV(\mathbb{R}^d)$ and a function $v \in BV(\mathbb{R}^d, \Omega)$ such that $(v, \alpha) \in \mathcal{S}(H_0)$ almost everywhere and satisfy the entropy dissipation condition (H3′). Then, v is a stationary entropy process solution of system (1.4.11).*

Moreover, let $T > 0$, $u_0 \in BV(\mathbb{R}^d, \Omega)^N$, and $\nu \in L^\infty\big([0, T) \times \mathbb{R}^d \times (0, 1), \Omega)\big)$ an associated entropy process solution. Then, there exist positive constants L_f and γ, independent of ν, v and α such that the following nonlinear stability property holds for all $R > 0$ and for almost every $t \in [0, T]$

$$\int_0^1 \int_{B(0,R)} h(\nu(t, x, \lambda), v(x), \alpha(x)) dx d\lambda + \gamma \int_0^1 \int_0^t \int_{B(0,R+L_f(t-\tau))} |\nu(\tau, x, \lambda) - v(x)|^2 dx d\tau d\lambda \leq \int_{B(0,R+L_f t)} h(u_0(x), v(x), \alpha(x)) dx. \qquad (1.4.19)$$

1.5 Links with the scientific production

This thesis is mostly contributed to the following submitted and preprint papers.

1. B. Boutin, T. H. T. Nguyen, and N. Seguin. *A stiffly stable semi-discrete scheme for the characteristic linear hyperbolic relaxation with boundary.* ESAIM: Mathematical Modelling and Numerical Analysis, 2020. In this paper, we study a stiffly stable semi-discrete scheme for the damped wave equation using SBP method. This result is given more details in Section 2.2. Furthermore, we focus in Section 2.3 on the time-implicit scheme of the IBVP (1.2.1). By using energy estimate and Laplace transform, this approximation is proved to be stiffly stable if the boundary condition satisfies $B_u B_v > 0$. Finally, in

Section 2.4, we study the fully discrete approximation of the IBVP (1.2.1) with homogeneous boundary condition $b^n \equiv 0$, for any $n \in \mathbb{N}$, obtained by upwind scheme in space and the implicit scheme in time. Let us remark that we have difficulty finding a sufficient condition for the stiff stability of the numerical scheme with nonzero boundary data, but we propose its possibility to a further work.

2. B. Boutin, T. H. T. Nguyen, and N. Seguin. *A stiffly stable fully discrete scheme for the damped wave equation using discrete transparent boundary condition.* Preprint, May 2020. By using the concept of transparent boundary condition, we show here that the SKC is a sufficient condition for the stiff stability of the time-implicit central differencing scheme for the linear damped wave equation with boundary. It is given more details in Chapter 3.

3. B. Boutin, T. H. T. Nguyen, A. Sylla, S. Tran-Tien, and J.-F. Coulombel. *High order numerical schemes for transport equations on bounded domains.* Preprint, Dec. 2019. The goal of this article is to extend the result in [22] to the case of nonzero incoming boundary data by developing a systematic construction and stability analysis for numerical boundary conditions arising from the so-called inverse Lax-Wendroff method. We illustrate the results with the Lax-Wendroff and $O3$ schemes. This is the main result in Chapter 4.

4. In Chapter 5, we study the stability of stationary solutions of non-conservation systems with the source term. We aim at proving that stationary solutions are stable among entropy process solution, which generalizes the concept of entropy weak solutions and can be obtained by passing to the limit of solution of the numerical scheme. To prove the stability of some stationary states within entropy process solution, we construct an associated relative entropy which allows to compare two states, and use the entropy dissipation condition. Besides, we also consider another assumption in [91] to prove the asymptotic stability of stationary solution of nonlinear relaxation balance laws.

Chapter 2

A stiffly stable discrete scheme for the damped wave equation using summation by parts method

We study the discretization of the linear damped wave equation in a quarter plane. Because of the work of Z. Xin and W. Xu in [96], we know that the stability condition of Kreiss is not sufficient to ensure the uniform stability of the initial boundary value problem (IBVP) for relaxation system independent of the stiffness of the source term. To remedy this problem, they then introduce a so-called Stiff Kreiss Condition (SKC), which turn out to be a necessary and sufficient condition on the boundary. In this chapter, we also exhibit a sufficient condition on the boundary to guarantee the uniform stability of the IBVP for relaxation system independent of the stiffness of the source term and of the space step. The boundary is approximated using a summation-by-parts method and the stiff stability is proved by energy estimates and Laplace transform. We also investigate if the condition is also necessary, following the continuous case studied by Z. Xin and W. Xu in [96].

2.1 Introduction

In many physical situations, we are interested in hyperbolic systems of partial differential equations with relaxation terms [3]. Such systems are found in relaxing gas theory [12], water waves [82, 94] and reactive flows [13]. One of the main features of these models is related to the notion of dissipation, leading to smooth solutions and asymptotic stability. The study of the zero relaxation limit for such systems has caught much interest, both from a theoretical and numerical point of view, after the works of Liu [64], Chen, Levermore and Liu [11], Hanouzet and Natalini [44], Yong [97, 98]. In this chapter, we are concerned with the numerical treatment of the boundary for hyperbolic relaxation systems by using the SBP method. Due to the presence of boundary layers and to the possible interaction of the boundary and initial layers, numerical schemes have to be properly designed so as to provide accurate approximations and consistent behaviors.

One of the simplest linear hyperbolic systems with relaxation is the linear damped wave equation on $u^\varepsilon, v^\varepsilon \in \mathbb{R}$

$$\begin{cases} \partial_t u^\varepsilon + \partial_x v^\varepsilon = 0, \\ \partial_t v^\varepsilon + a\partial_x u^\varepsilon = -\varepsilon^{-1} v^\varepsilon, \end{cases} \tag{2.1.1}$$

where $a > 0$ and the relaxation time $\varepsilon > 0$ characterizes the stiffness of the relaxation process. When ε goes to zero, the model may be simplified. We expect indeed that for any position x

and any time t, the solution $(u^\varepsilon, v^\varepsilon)$ tends to $(u(x), 0)$, which is the solution of the corresponding equilibrium system [11, 96].

In order to determine a unique solution to the problem (2.1.1) in the quarter plane $x > 0$, $t > 0$, it is necessary to specify values of the solution at initial time

$$u^\varepsilon(x, 0) = u_0(x), \quad v^\varepsilon(x, 0) = v_0(x), \tag{2.1.2}$$

and to impose conditions on the solution at the boundary

$$B_u u^\varepsilon(0, t) + B_v v^\varepsilon(0, t) = b(t), \tag{2.1.3}$$

where B_u and B_v are constants. For simplicity, we also assume the initial data $f(x) = (u_0(x), v_0(x))$ and the boundary data $b(t)$ to be compatible at the space-time corner $(x, t) = (0, 0)$, i.e.

$$f(0) = f'(0) = 0, \quad b(0) = b'(0) = 0. \tag{2.1.4}$$

In some cases, the suitable boundary conditions comes from physical considerations. At a solid wall that bounds the flow of a fluid, for example, one sets the normal component of the fluid velocity equal to zero (if effects of viscosity are to be considered, the tangential component must also vanish). In other situations, the choice of boundary conditions is not so obvious. This is the case when considering artificial boundaries, which do not correspond to a well-identifies physical phenomenon. In general, not any boundary condition is suitable for a given hyperbolic problem. In the case of the problem (2.1.1), which is a particular case of the Jin-Xin relaxation model in one space dimension [51], the hyperbolic structure is related to the Riemann invariants $\sqrt{a}u^\varepsilon \pm v^\varepsilon$ and to the characteristic velocities $\pm\sqrt{a}$. Therefore the boundary condition (2.1.3) has to satisfy the Uniform Kreiss Condition (UKC)

$$B_u + \sqrt{a} B_v \neq 0. \tag{2.1.5}$$

Only under this assumption, the incoming flow $\sqrt{a}u^\varepsilon + v^\varepsilon$ at the boundary $x = 0$ can be deduced from the outgoing flow $\sqrt{a}u^\varepsilon - v^\varepsilon$ and the data $b(t)$. Therefore the initial boundary value problem (IBVP) (2.1.1)-(2.1.3) is well-posed for each fixed ε (see [3, 96, 97]).

In [96], Z. Xin and W. Xu study the asymptotic equivalence of a general linear system of one-dimensional conservation laws and the corresponding relaxation model proposed by S. Jin and Z. Xin [51] in the limit of a small relaxation rate ε. The main issue is to extend and precise this asymptotic equivalence in the presence of physical boundaries. Within the same problematic, W.-A. Yong in [97] proposed a Generalized Kreiss Condition (GKC) for general multi-dimensional linear constant coefficient relaxation systems, or one-dimensional nonlinear systems, with non-characteristic boundaries. This condition enables uniform stability estimates and a reduced boundary condition for the corresponding equilibrium system. For the special Jin-Xin system (2.1.1) with boundary condition (2.1.3) but with stiff source terms of the form $\varepsilon^{-1}(\lambda u^\varepsilon - v^\varepsilon)$ for some λ, Z. Xin and W. Xu identify and rigorously justify a necessary and sufficient condition (which they call the Stiff Kreiss Condition, or SKC in short) on the boundary condition to guarantee the uniform well-posedness of the IBVP, independently of the relaxation parameter. In addition to the work in [97], their study also covers the characteristic case and provides optimal asymptotic expansions for the limit process, handling with boundary and/or initial layers. In the case of our system (2.1.1), the parameter $\lambda = 0$ so that the boundary is characteristic for limit equation, and the SKC in [96] then simply reduces to

$$B_v = 0, \qquad \text{or} \qquad \frac{B_u}{B_v} \notin \left[-\sqrt{a}, 0\right]. \tag{2.1.6}$$

The motivation of the present study is to analyze the counterpart of the above results but now for the difference approximation of the IBVP (2.1.1)-(2.1.3). The major issues in the theory of the relaxation approximations to equilibrium system of conservation laws is the appearance of stiff boundary layers in the presence of physical or numerical boundary conditions due to the additional characteristic speeds introduced in the relaxation systems. On the other hand, the stability estimate obtained for a certain approximation is the key to the proper error estimates. Thus, the way of formulating boundary conditions for the relaxation systems so as to guarantee the uniform stability and to minimize the artificial boundary layer is a crucial issue to the success of the schemes.

In this chapter, we first study a stiffly stable semi-discrete scheme for the IBVP (2.1.1)-(2.1.3) obtained by the central differencing scheme in Section 2.2. After that, we focus in Section 2.3 on the time-implicit scheme of the IBVP (2.1.1)-(2.1.3). By using energy estimate and Laplace transform, we prove that $B_u B_v > 0$ is a sufficient condition for the stiff stability of this scheme. Finally, in Section 2.4, we consider the fully discrete approximation of the IBVP (2.1.1)-(2.1.3) with homogeneous boundary condition $b^n \equiv 0$, for any $n \in \mathbb{N}$, obtained by upwind scheme in space and the implicit scheme in time.

2.2 The semi-discrete central scheme

Let $\Delta x > 0$ be the space step and $U(x_j, t) = (u^\varepsilon, v^\varepsilon)^T(x_j, t)$ with $x_j = j\Delta x$, for any $j \in \mathbb{N}$. The solution to the IBVP (2.1.1)-(2.1.3) is approximated by a sequence $U_j(t) = (u_j(t), v_j(t))^T$ (where we omit the explicit dependence on ε). We focus in this section on the semi-discrete approximation of the IBVP obtained by the central differencing scheme and derive a sufficient condition for its stiff stability. Let A, S and B be the following matrices:

$$A = \begin{pmatrix} 0 & 1 \\ a & 0 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} B_u & B_v \end{pmatrix}. \tag{2.2.1}$$

A first step towards the semi-discrete approximation of the IBVP (2.1.1)-(2.1.3) is the following system

$$\begin{aligned} &\partial_t U_j(t) + (\mathcal{Q}U)_j(t) = \frac{1}{\varepsilon} S U_j(t), \quad j \geq 1, \\ &U_j(0) = f_j, \\ &BU_0(t) = b(t), \end{aligned} \tag{2.2.2}$$

with the discrete Cauchy data $f_j = U(x_j, 0)$. The difference operator $(\mathcal{Q}U)_j(t)$ is a consistent approximation of the first order space-derivative $A\partial_x U(x_j, t)$ in the sense that $(\mathcal{Q}U)(x_j, t) = A\partial_x U(x_j, t) + \mathcal{O}(\Delta x^p)$, for some $p > 0$. It is defined at any discrete point including the boundary point $j = 0$.

The summation by parts (SBP) finite difference operators were first derived in [71, 72]. In [83], the analysis was revisited and exact expressions for the finite difference coefficients were obtained. In the case of the central scheme, the modification of the difference operator $\mathcal{Q}U$ at $j = 0$ can also be interpreted as the use of an extra boundary condition. It means that we use the centered approximation at the boundary point $j = 0$ but supply another boundary condition that determines a ghost value U_{-1} through the identity

$$U_1 - 2U_0 + U_{-1} = 0.$$

If we eliminate U_{-1}, then we obtain a one-sided approximation. In [42], the corresponding energy estimate is obtained by using the scalar product and norm

$$\langle U, V \rangle_{\Delta x} = \frac{\Delta x}{2} \langle U_0, V_0 \rangle + \Delta x \sum_{j=1}^{\infty} \langle U_j, V_j \rangle, \qquad \|U\|_{\Delta x}^2 = \langle U, U \rangle_{\Delta x}. \tag{2.2.3}$$

with $\langle ., . \rangle$ being the usual Euclidean inner product. The considered difference operator reads

$$(\mathcal{Q}U)_j = \begin{cases} \dfrac{1}{2\Delta x} A(U_{j+1} - U_{j-1}), & j \geq 1, \\ \dfrac{1}{\Delta x} A(U_1 - U_0), & j = 0, \end{cases} \tag{2.2.4}$$

which uses a noncentered approximation at the boundary, so that the difference operator is defined at all gridpoints including the boundary point $j = 0$.

Let us emphasize that the numerical scheme (2.2.2) still needs one more scalar equation at the boundary point $j = 0$ so as to be fully defined, due to the fact that the matrix B has rank one. This is actually a discrete feature only, since in the continuous case this single equation is exactly complemented by the only incoming characteristic (under UKC). We choose to define the remaining discrete boundary value in agreement with the dissipativeness of the source term. We then use a symmetric form of the problem, based on the matrix P and on the symmetric positive definite matrix H_P below

$$P = \begin{pmatrix} B_u & B_v \\ 1 & 0 \end{pmatrix}, \qquad H_P = \begin{pmatrix} 1 & -B_u \\ -B_u & aB_v^2 + B_u^2 \end{pmatrix}. \tag{2.2.5}$$

As a consequence, the matrix $P^T H_P P$ is symmetric positive definite, $H_P PAP^{-1}$ is symmetric and $H_P PSP^{-1}$ is negative semi-definite. Since $P^T H_P PA$ is also a symmetric matrix, one has

$$\left\langle U, P^T H_P P(\mathcal{Q}U) \right\rangle_{\Delta x} (t) = -\frac{1}{2} \left\langle U_0, P^T H_P PAU_0 \right\rangle (t),$$

which is the discrete counterpart of the equality

$$\int_0^{+\infty} \left\langle U, P^T H_P PA\partial_x U \right\rangle (x,t) dx = -\frac{1}{2} \left\langle U, P^T H_P PAU \right\rangle (0,t)$$

available in the continuous case. Moreover, at the boundary $j = 0$, we obtain

$$\left\langle \partial_t U_0, P^T H_P PU_0 \right\rangle (t) + \left\langle (\mathcal{Q}U)_0, P^T H_P PU_0 \right\rangle (t) = \frac{1}{\varepsilon} \left\langle SU_0, P^T H_P PU_0 \right\rangle (t). \tag{2.2.6}$$

Inserting now the homogeneous boundary condition $BU_0(t) = 0$ and introducing the matrix $\Pi_2 = \begin{pmatrix} 0 & 1 \end{pmatrix}$, the previous equality (2.2.6) can be reformulated as

$$\partial_t \left(\Pi_2 H_P PU_0\right)(t) \left(\Pi_2 PU_0\right)(t) + (\Pi_2 H_P P(\mathcal{Q}U)_0)(t)(\Pi_2 PU_0)(t) = \frac{1}{\varepsilon} \left(\Pi_2 H_P PSU_0\right)(t) \left(\Pi_2 PU_0\right)(t).$$

We therefore propose the following numerical approximation at the boundary

$$\partial_t(\Pi_2 H_P PU_0)(t) + \Pi_2 H_P P(\mathcal{Q}U)_0(t) = \frac{1}{\varepsilon} \Pi_2 H_P PSU_0(t).$$

To summarize, along the rest of the chapter, we will study the following semi-discrete approximation of the IBVP (2.1.1)-(2.1.3):

$$\begin{cases} \partial_t U_j(t) + (\mathcal{Q}U)_j(t) = \varepsilon^{-1} SU_j(t), & j \geq 1, \ t \geq 0, \\ U_j(0) = f_j, & j \geq 0, \\ BU_0(t) = b(t), & t \geq 0, \\ \partial_t(\Pi_2 H_P PU_0)(t) + \Pi_2 H_P P(\mathcal{Q}U)_0(t) = \varepsilon^{-1} \Pi_2 H_P PSU_0(t), & t \geq 0. \end{cases} \tag{2.2.7}$$

Main result: For the continuous IBVP (2.1.1)-(2.1.3), the UKC (2.1.5) is not enough and a more stringent restriction has to be imposed. Our aim is to determine a sufficient condition for the stiff stability of the above semi-discrete IBVP (2.2.7), in other words the uniform stability with respect to the stiffness of the relaxation term. First of all, let us address the question of the existence and uniqueness of a solution to the infinite-dimensional ODE system (2.2.7) through the next result.

Lemma 2.2.1. *Let us consider some fixed parameters $(B_u, B_v) \in \mathbb{R}^2$ with $B_v \neq 0$, and $\varepsilon, \Delta x > 0$. For any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and any $b \in \mathcal{C}^1(\mathbb{R}_+, \mathbb{R})$ there exists a unique solution $(U_j)_{j\in\mathbb{N}} \in \mathcal{C}^1([0, +\infty[, \ell^2(\mathbb{N}, \mathbb{R}^2))$ to (2.2.7).*

Proof. The proof rests on the common linear Cauchy-Lipschitz theorem in the Banach space $\ell^2(\mathbb{N}, \mathbb{R}^2)$. Let us bring some precisions concerning the solvability of the two rank-one boundary equations. The first algebraic equation reads simply

$$B_u u_0(t) + B_v v_0(t) = b(t),$$

while the second differential one reads equivalently, for some linear operator $\mathsf{L} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, as

$$aB_v^2 u_0'(t) - B_u B_v v_0'(t) + \mathsf{L}(U_0(t), U_1(t)) = 0.$$

Eliminating v_0 from the algebraic boundary condition, we get thus $B_v v_0'(t) = b'(t) - B_u u_0'(t)$ and therefore

$$(aB_v^2 + B_u^2)u_0'(t) = -\mathsf{L}(U_0(t), U_1(t)) + B_u b'(t).$$

The solvability of the whole ODE system is therefore deduced by $B_v \neq 0$ together with $aB_v^2 + B_u^2 \neq 0$. The details are left to the reader. □

Theorem 2.2.2 (Main result)**.** *Under the strict dissipativity condition*

$$B_u B_v > 0, \tag{2.2.8}$$

for any $T > 0$ there exists $C_T > 0$ such that for any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and any $b \in \mathcal{C}^1(\mathbb{R}^+, \mathbb{R}) \cap L^2(\mathbb{R}^+, \mathbb{R})$, the solution $(U_j)_{j\in\mathbb{N}}$ to (2.2.7) satisfies

$$\int_0^T |U_0(t)|^2\, dt + \int_0^T \sum_{j\geq 0} \Delta x |U_j(t)|^2\, dt \leq C_T \left(\sum_{j\geq 0} \Delta x |f_j|^2 + \int_0^T |b(t)|^2\, dt \right), \tag{2.2.9}$$

where the constant C_T is independent of the data f and b, but most importantly of $\varepsilon \in (0, +\infty)$ and $\Delta x \in (0, 1]$.

The proof of Theorem 2.2.2 is based on two main ingredients, by assembling a result for the case of homogeneous boundary data and another for the case with homogeneous initial data. We state successively hereafter these two statements.

Proposition 2.2.3 (Homogeneous boundary condition)**.** *Assume that the parameters $\Delta x \in (0, 1]$, $\varepsilon > 0$ and (B_u, B_v) satisfy the discrete strict dissipativity condition*

$$2a\frac{B_u}{B_v} + \frac{\Delta x}{\varepsilon}\left(\frac{B_u}{B_v}\right)^2 > 0. \tag{2.2.10}$$

Then there exists a constant $C > 0$ such that for any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, the solution $(U_j)_{j\in\mathbb{N}}$ to (2.2.7) with $b \equiv 0$ satisfies

$$\left\langle U, P^T H_P P U \right\rangle_{\Delta x}(T) + C\int_0^T |U_0|^2(t)dt \leq \left\langle f, P^T H_P P f \right\rangle_{\Delta x}. \tag{2.2.11}$$

More precisely,

a) If $B_u B_v > 0$ then (2.2.11) holds uniformly, i.e. with C independent of ε and Δx.

b) If $B_u B_v < 0$ then considering some $\delta_0 > -2aB_v B_u^{-1}$, there exists $C(\delta_0) > 0$ such that (2.2.11) holds uniformly with $C = C(\delta_0)$, as soon as $\Delta x \geq \delta_0 \varepsilon$.

Proposition 2.2.4 (Homogeneous initial condition)**.** *Assume that the boundary condition is strictly dissipative, satisfying* (2.2.8)*. Then, there exists a constant* $C > 0$ *such that for any* $\alpha > 0$ *there exists* $\Delta x_0 > 0$ *such that the following property holds. For any* $b \in \mathcal{C}^1(\mathbb{R}^+, \mathbb{R}) \cap L^2(\mathbb{R}^+, \mathbb{R})$ *and* $\Delta x \le \Delta x_0$*, the solution* $(U_j)_{j\in\mathbb{N}}$ *to* (2.2.7) *with* $(f_j)_{j\in\mathbb{N}} \equiv 0$ *satisfies*

$$\alpha \Delta x \int_0^\infty \sum_{j\ge 0} e^{-2\alpha t}|U_j(t)|^2 dt + \int_0^\infty e^{-2\alpha t}|U_0(t)|^2 dt \le C \int_0^\infty e^{-2\alpha t}|b(t)|^2 dt. \tag{2.2.12}$$

Remark 2.2.5. *Since* $P^T H_P P$ *is a symmetric positive definite matrix, the following inequality holds for some constants* $m, n > 0$*, independent of* Δx*:*

$$m \left\langle U, P^T H_P P U \right\rangle_{\Delta x}(t) \le \langle U, U \rangle_{\Delta x}(t) \le n \left\langle U, P^T H_P P U \right\rangle_{\Delta x}(t),$$

which will be useful to prove the estimate (2.2.9) *with weighted-in-time norms from* (2.2.11)*.*

Xin and Xu in [96] considered the IBVP for the Jin-Xin relaxation model [51] and derived the SKC (2.1.6) for its stiff well-posedness. They show in particular that the IBVP is well-posed if and only if (2.1.6) holds. In the discrete IBVP (2.2.7), it seems that even the SKC is not sufficient to derive uniform stability estimates. In comparison, the strict dissipativity condition (2.2.8) is more restrictive, but we are only able to prove that it is sufficient. Following [97] and [96], we also perform a normal mode analysis to construct unstable solutions and, based on some numerical investigations, the condition (2.2.8) would appear to be also necessary for the stiff stability. Let us mention that the discrete strict dissipativity condition (2.2.10) is not implied by the SKC (2.1.6), probably due to some numerical diffusion at the boundary.

The Proposition 2.2.3 is studied in Section 2.2.1.1 by means of the discrete energy method. In order to illustrate the relevance of the discrete strict dissipativity condition (2.2.10), we present in Section 2.2.1.2 some numerical results, for various values of the parameters (B_u, B_v) and show that the energy $\left\langle U(t), P^T H_P P U(t) \right\rangle_{\Delta x}$ increases if the condition (2.2.10) does not hold. In Section 2.2.2.1, we want to address the question of the existence of unstable solutions in order to derive necessary condition for stability by using the normal mode analysis. In Section 2.2.2.2, we present numerical results and show that $B_u B_v > 0$ seems to be necessary to ensure the stiff stability of the discrete IBVP. Even if the boundary condition (2.1.3) satisfies the SKC, there exist unstable solutions of the discrete IBVP (2.2.7). To isolate the effects of a possible boundary layer and avoid the complicated interactions of boundary and initial layers, in Section 2.2.3, we consider the IBVP (2.2.7) with homogeneous initial data and nonzero boundary data $b(t)$. In Section 2.2.3.1, the numerical solution $(U_j(t))_{j\in\mathbb{N}}$ can be constructed by Laplace transform. By using the Parseval's identity, under assumption $B_u B_v > 0$, the Proposition 2.2.4 is proved in Section 2.2.3.2 .

2.2.1 Stiff stability of the semi-discrete IBVP with homogeneous boundary condition

In this section, we consider the IBVP (2.2.7) for homogeneous boundary condition $b \equiv 0$, nonzero Cauchy data $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and prove Proposition 2.2.3 by means of the discrete energy method.

2.2.1.1 The energy method

In the continuous case, the energy estimates are obtained using integration by parts rules. Therefore, we make use of the similar SBP rules for the discrete approximations of $\partial/\partial x$ [42]. The sufficient condition (2.2.10) is then deduced directly from discrete energy estimates.

According to the scalar product (2.2.3), we can see that

$$\left\langle \partial_t U, P^T H_P P U \right\rangle_{\Delta x}(t) = \frac{\Delta x}{2} \left\langle \partial_t \left(H_P P U_0 \right), P U_0 \right\rangle(t) + \Delta x \sum_{j=1}^{\infty} \left\langle \partial_t U_j, P^T H_P P U_j \right\rangle(t). \quad (2.2.13)$$

Since $P^T H_P P$ is a symmetric positive definite matrix and using the homogeneous boundary condition $BU_0(t) = 0$ and thus $PU_0 = (0, \Pi_2 P U_0)^T$, the previous equality (2.2.13) can be reformulated as

$$\frac{1}{2}\partial_t \left\langle U, P^T H_P P U \right\rangle_{\Delta x}(t) = \frac{\Delta x}{2}\partial_t \left(\Pi_2 H_P P U_0 \right)(t) \left(\Pi_2 P U_0 \right)(t) + \Delta x \sum_{j=1}^{\infty} \left\langle \partial_t U_j, P^T H_P P U_j \right\rangle(t). \quad (2.2.14)$$

Now, we show how the difference operator $(\mathfrak{Q}U)_{j\in\mathbb{N}}$ can be applied for the IBVP (2.2.7) for the homogeneous boundary condition at all gridpoints including the boundary point $j = 0$

$$\partial_t U_j(t) = \frac{1}{\varepsilon} S U_j(t) - \frac{1}{2\Delta x} A \left(U_{j+1}(t) - U_{j-1}(t) \right), \quad j \geq 1,$$

and

$$\partial_t \left(\Pi_2 H_P P U_0 \right)(t) = \frac{1}{\varepsilon} \Pi_2 H_P P S U_0(t) - \frac{1}{\Delta x} \Pi_2 H_P P A \left(U_1 - U_0 \right)(t).$$

As a consequence, the equation (2.2.14) can be represented as

$$\begin{aligned}
\partial_t \left\langle U, P^T H_P P U \right\rangle_{\Delta x}(t) = {} & \frac{\Delta x}{\varepsilon} \left(\Pi_2 H_P P S U_0 \right)(t) \left(\Pi_2 P U_0 \right)(t) + \frac{2\Delta x}{\varepsilon} \sum_{j=1}^{\infty} \left\langle S U_j, P^T H_P P U_j \right\rangle(t) \\
& + \left(\Pi_2 H_P P A U_0 \right)(t) \left(\Pi_2 P U_0 \right)(t) - \left(\Pi_2 H_P P A U_1 \right)(t) \left(\Pi_2 P U_0 \right)(t) \\
& - \sum_{j=1}^{\infty} \left\langle A U_{j+1}, P^T H_P P U_j \right\rangle(t) + \sum_{j=1}^{\infty} \left\langle A U_{j-1}, P^T H_P P U_j \right\rangle(t).
\end{aligned} \quad (2.2.15)$$

On the other hand, the last term in the right hand side becomes

$$\sum_{j=1}^{\infty} \left\langle A U_{j-1}, P^T H_P P U_j \right\rangle(t) = \left\langle A U_0, P^T H_P P U_1 \right\rangle(t) + \sum_{j=1}^{\infty} \left\langle A U_j, P^T H_P P U_{j+1} \right\rangle(t).$$

Since $H_P P A P^{-1}$ is symmetric and $PU_0 = (0, \Pi_2 P U_0)^T$, one gets

$$\left\langle A U_0, P^T H_P P U_1 \right\rangle(t) = \left(\Pi_2 H_P P A U_1 \right)(t) \left(\Pi_2 P U_0 \right)(t),$$

and then

$$\sum_{j=1}^{\infty} \left\langle A U_{j-1}, P^T H_P P U_j \right\rangle(t) = \left(\Pi_2 H_P P A U_1 \right)(t) \left(\Pi_2 P U_0 \right)(t) + \sum_{j=1}^{\infty} \left\langle A U_{j+1}, P^T H_P P U_j \right\rangle(t). \quad (2.2.16)$$

Substituting (2.2.16) into (2.2.15), the three last terms in (2.2.15) vanish. After some calculations, we obtain

$$\partial_t \left\langle U, HU \right\rangle_{\Delta x}(t) + 2a\frac{B_u}{B_v} u_0^2(t) + \frac{\Delta x}{\varepsilon} v_0^2(t) = -\frac{2\Delta x}{\varepsilon} \sum_{j=1}^{\infty} v_j^2(t),$$

where the symmetric positive definite matrix $H = B_v^{-2} P^T H_P P$ is simply

$$H = \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.2.17}$$

In order for the energy method to work, the boundary condition has to satisfy

$$2a\frac{B_u}{B_v}u_0^2(t) + \frac{\Delta x}{\varepsilon}v_0^2(t) \geq C|U_0(t)|^2$$

for some constant $C > 0$ whenever $B_u u_0(t) + B_v v_0(t) = 0$. This leads to the following sufficient condition

$$2a\frac{B_u}{B_v} + \frac{\Delta x}{\varepsilon}\left(\frac{B_u}{B_v}\right)^2 > 0,$$

under which we directly get the inequality

$$\partial_t \langle U, HU \rangle_{\Delta x}(t) + C|U_0|^2(t) \leq 0,$$

and thus finally

$$\langle U, HU \rangle_{\Delta x}(T) + C\int_0^T |U_0|^2(t)dt \leq \langle f, Hf \rangle_{\Delta x}. \tag{2.2.18}$$

More into the details, the following cases occur:

- If $B_u B_v > 0$ then there exists $C \leq 2aB_uB_v\left(B_u^2 + B_v^2\right)^{-1}$ such that the inequality (2.2.18) holds uniformly.
- If $B_u B_v < 0$, consider some $\delta_0 > -2aB_vB_u^{-1}$. Then there exists $C(\delta_0) > 0$ such that the inequality (2.2.18) holds uniformly as soon as $\Delta x \geq \delta_0\varepsilon$ with $C = C(\delta_0)$. For example, if we choose $\delta_0 = -3aB_vB_u^{-1}$ then there exists $C \leq -aB_uB_v\left(B_u^2 + B_v^2\right)^{-1}$ such that (2.2.18) holds uniformly.

This ends the proof of Proposition 2.2.3.

Let us mention that, assuming the strict dissipativity condition (2.2.8) of the main theorem 2.2.2 to be fulfilled, the discrete strict dissipativity condition (2.2.10) is then automatically satisfied. Then, from the inequality (2.2.18), for any $T > 0$, there exists a constant $C_T > 0$ such that the following inequality holds

$$\int_0^T \sum_{j\geq 0} \Delta x|U_j(t)|^2 dt + \int_0^T |U_0(t)|^2 dt \leq C_T \sum_{j\geq 0} \Delta x|f_j|^2. \tag{2.2.19}$$

This will be used to prove Theorem 2.2.2.

2.2.1.2 Numerical experiments

In this section we perform some numerical experiments and observe the effective behavior (i.e. the time evolution) of the energy

$$E(t) := \langle U(t), HU(t) \rangle_{\Delta x},$$

according to whether or not the discrete strict dissipativity condition (2.2.10) is valid. We also have a look at the degenerate case when the UKC (2.1.5) does not hold (and thus, none of the other stability conditions). As discussed in the previous section and in the calculations of Xin and Xu [96], we expect to observe the decrease of the energy $E(t)$ as soon as $B_u B_v > 0$. What happens in the case $B_u B_v < 0$, but while the the discrete strict dissipativity condition (2.2.10) still holds, is also experimented.

As main parameters for the experiments, we fix the space step $\Delta x = 10^{-2}$, choose $a = 4$ and let ε and the boundary parameter (B_u, B_v) vary. Our purpose is not to discuss the choice of a time integrator for the ODE system, let us mention that in any case we make use of the integrated solver ode45 of MATLAB (explicit variable time-step Runge-Kutta $(4,5)$ formula, the Dormand-Prince pair).

The test case we consider concerns the following data. The boundary data is the homogeneous one $b \equiv 0$. The initial data is

$$f_j = \begin{cases} (0,0), & \text{if } x_j = 0, \\ (15,10)^T, & \text{if } 0 < x_j \leq 1/2, \\ (0,0), & \text{if } x_j > 1/2. \end{cases}$$

Let us first observe that these data are compatible in the corner $(x,t) = (0,0)$ in the sense that $Bf_0(0) = b(0)$. Moreover the choice of an initial data with support in $[0,1/2]$ is motivated by the property of finite speed of propagation available at the continuous side (2.1.1). More precisely, the exact solution we approximate has characteristic velocities ± 2 and therefore vanishes outside some space interval $[0,0.9]$ for small times in $[0,0.2]$. Thus we choose for our experiments the space interval $[0,1]$ and the time interval $[0,T]$ with $T = 0.2$. Let us however mention that, strictly speaking, this analysis is actually wrong at the semi-discrete level and that in addition we have to define some discrete right boundary condition at $x = 1$. The most natural choice in this situation is to select the homogeneous Neumann boundary condition $U_{J+1}(t) = U_J(t)$ at the rightmost cell J. We here don't address the precise analysis of this choice but the numerical experiments seem to behave correctly, for example when extending the space-domain to $[0,2]$. Other strategies exist in the litterature, with for example discrete transparent boundary conditions (see for example [5]), but we postpone these possibilities to a further work.

Firstly, we choose a set of values (B_u, B_v) such that the discrete strict dissipativity condition (2.2.10) is satisfied with $\varepsilon = 10^{-2}$ and also with $\varepsilon = 10^2$. The Figure 2.2.1 shows the evolution of the energy $E(t)$ over the time interval $t \in [0, 0.2]$.

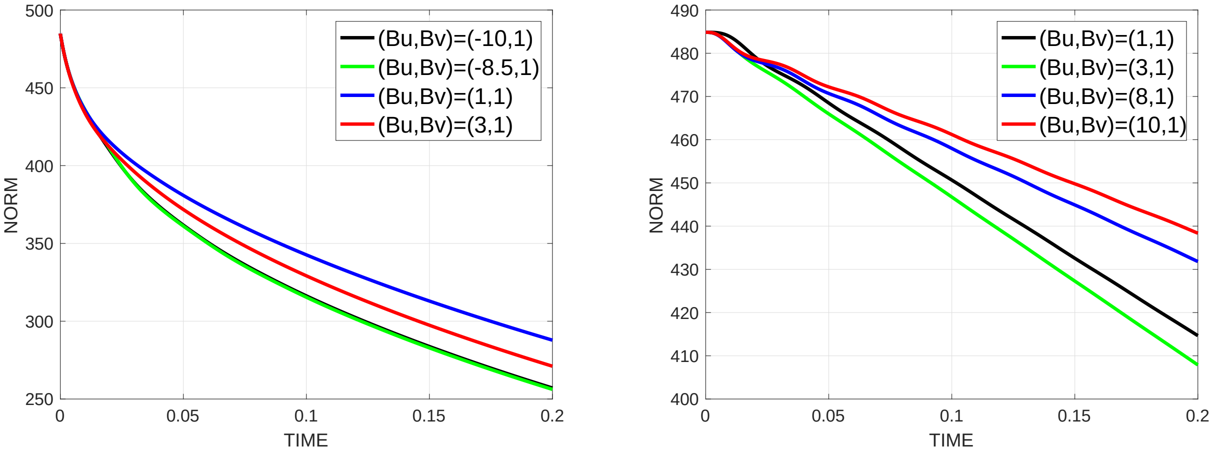

Figure 2.2.1: Energy evolution with the discrete strict dissipativity condition (2.2.10), for $\varepsilon =\ 10^{-2}$ (left) and $\varepsilon = 10^2$ (right).

- We proved that for any $\varepsilon \in (0, +\infty)$ and (B_u, B_v) satisfying the discrete strict dissipativity condition (2.2.10), $E(t)$ is decreasing. This is strongly supported by the experiments.

Observe also that the decrease of $E(t)$ is true even in the case $B_u B_v < 0$ provided the discrete strict dissipativity condition (2.2.10) is true. This is the case for example for $\varepsilon = 10^{-2}$ together with the parameters $(B_u, B_v) = (-8.5, 1)$.

- In the case $\varepsilon = 10^{-2}$ the energy $E(t)$ go down suddenly for small $t > 0$. This is due to the initial relaxation of the solution to the equibrium system. In the case $\varepsilon = 10^2$, the decrease seems to be linear. It is not so much influenced by the relaxation source term but more by the boundary dissipation.

Secondly, we choose a set of values (B_u, B_v) such that the discrete strict dissipativity condition (2.2.10) is not satisfied with $\varepsilon = 10^{-2}$ nor with $\varepsilon = 10^2$. Besides, we also present the evolution of the energy for parameters such that the Uniform Kreiss Condition (2.1.5) is wrong. The Figure 2.2.2 shows the evolution of $E(t)$ over the time interval $t \in [0, 0.2]$.

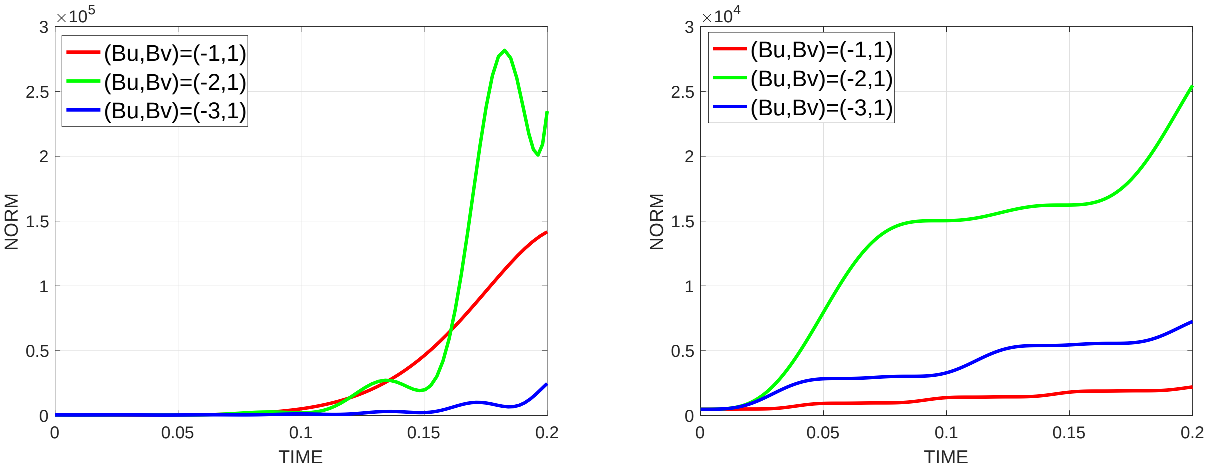

Figure 2.2.2: Energy evolution without the discrete strict dissipativity condition (2.2.10), for $\varepsilon = 10^{-2}$ (left) and $\varepsilon = 10^2$ (right).

- On the boundary $x = 0$, for all $\varepsilon > 0$, if the boundary condition (2.1.3) with homogeneous boundary condition $b(t) \equiv 0$ does not satisfy the UKC, then $v^\varepsilon = \sqrt{a} u^\varepsilon$. Therefore, the numerical scheme of the IBVP is not stable for each fixed ε. For $\varepsilon = 10^{-2}$ and $\varepsilon = 10^2$, if we choose $(B_u, B_v) = (-2, 1)$ then the values of $E(t)$ increase quickly.
- When the discrete strict dissipativity condition (2.2.10) fails, then we observe for any $t \in (0, 0.2]$ the inequality $E(t) > E(0)$. In the particular case $\varepsilon = 10^{-2}$, the evolution is non-monotone and there exists $0 < t_1 < t_2$ such that $E(t_1) > E(t_2)$. However, after that the values of $E(t)$ increase rapidly. In the case $\varepsilon = 10^2$, the values of $E(t)$ rise gradually.

Clearly, the numerical results show that the energy $E(t)$ increases in time as soon as the discrete strict dissipativity condition (2.2.10) does not hold. The behavior is even worse when the UKC (2.1.5) is not satisfied. It seems that the condition (2.2.10) is also necessary to ensure the non-increase of the energy, but let us stress that an increasing energy with respect to time may not be in contradiction with the stiff stability.

2.2.2 Stiff strong stability of the semi-discrete IBVP

In the continuous case, the IBVP (2.1.1)-(2.1.3) is stiffly well-posed if and only if the boundary condition satisfies the SKC (2.1.6). Now, we want to address the question of existence of unstable solutions in order to derive a necessary condition for the stability of the discrete IBVP (2.2.7). Following W.-A. Yong in [97] and Z. Xin and W. Xu in [96], we shall apply the normal mode analysis to derive the strict dissipativity condition (2.2.8).

2.2.2.1 Strictly dissipative boundary conditions

We look for (nontrivial) solutions of (2.2.7) satisfying the homogeneous boundary condition $BU_0(t) = 0$ and of the form

$$U_j(t) = e^{\xi t/\varepsilon}\phi_j, \tag{2.2.20}$$

with $\xi \in \mathbb{C}$ such that $\mathrm{Re}\,(\xi) > 0$, and $(\phi_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{C}^2)$. Such solutions, if they exist, clearly violate the $\varepsilon-$uniform ℓ^2 estimates in (2.2.9). Our goal is to find a sufficient condition to ensure that they do not exist.

Substituting (2.2.20) into (2.2.7), we have to solve the following problem

$$\phi_{j+1} - \phi_{j-1} = 2\Delta x\varepsilon^{-1}M(\xi)\phi_j, \quad j > 0 \tag{2.2.21a}$$

$$B\phi_0 = 0, \tag{2.2.21b}$$

$$\Pi_2 P^{-T}HA\left(\phi_1 - \left(I + \Delta x\varepsilon^{-1}M(\xi)\right)\phi_0\right) = 0, \tag{2.2.21c}$$

where we denote the following matrix $M(\xi)$, already used in [96]:

$$M(\xi) = A^{-1}(S - \xi I) = \frac{1}{a}\begin{pmatrix} 0 & -(1+\xi) \\ -a\xi & 0 \end{pmatrix}. \tag{2.2.22}$$

For convenience in the notations, we recall that the eigenvalues and eigenvectors of $M(\xi)$ can be easily found to be respectively

$$\mu_\pm(\xi) = \pm\sqrt{\frac{\xi(1+\xi)}{a}}, \qquad r_\pm(\xi) = \begin{pmatrix} 1 \\ \dfrac{a\mu_\mp(\xi)}{1+\xi} \end{pmatrix}. \tag{2.2.23}$$

According to Lemma 2.5.1 with the property $\mathrm{Re}(\xi) > 0$, we can prove

$$\mathrm{Re}\,(\mu_-(\xi)) \leq -\frac{\mathrm{Re}\,(\xi)}{\sqrt{a}} < 0, \tag{2.2.24}$$

while, as a consequence,

$$\mathrm{Re}\,(\mu_+(\xi)) \geq \frac{\mathrm{Re}\,(\xi)}{\sqrt{a}} > 0.$$

Let $\overline{P}(\xi)$ be the 2×2 matrix whose columns are composed by the component of the vector $r_\pm(\xi)$:

$$\overline{P}(\xi) = \begin{pmatrix} 1 & 1 \\ \dfrac{a\mu_-(\xi)}{1+\xi} & \dfrac{a\mu_+(\xi)}{1+\xi} \end{pmatrix},$$

so that $M(\xi) = \overline{P}(\xi)D(\xi)\overline{P}^{-1}(\xi)$ with $D(\xi) = \mathrm{diag}(\mu_+(\xi), \mu_-(\xi))$. Let us also define

$$\psi_j = \left(\psi_j^I, \psi_j^{II}\right)^T = \overline{P}^{-1}(\xi)\phi_j.$$

Now, the two-dimensional linear second order recurrence relations (2.2.21a) reads also under the form of two decoupled scalar second order linear recurrence relations

$$\psi_{j+1}^I - \psi_{j-1}^I = \frac{2\mu_+(\xi)\Delta x}{\varepsilon}\psi_j^I, \tag{2.2.25a}$$

$$\psi_{j+1}^{II} - \psi_{j-1}^{II} = \frac{2\mu_-(\xi)\Delta x}{\varepsilon}\psi_j^{II}. \tag{2.2.25b}$$

Firstly, we look at the solution $(\psi_j^I)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N},\mathbb{C})$ to (2.2.25a), and assume first that the solution has the form

$$\psi_j^I = z(\xi)R_1, \tag{2.2.26}$$

for some $|z(\xi)| < 1$ and $R_1 \in \mathbb{C}$. Substituting the ansatz (2.2.26) into (2.2.25a), we then obtain $z(\xi)$ among the values

$$z_\pm(\xi) = \frac{\mu_+(\xi)\Delta x}{\varepsilon} \pm \sqrt{\left(\frac{\mu_+(\xi)\Delta x}{\varepsilon}\right)^2 + 1}. \tag{2.2.27}$$

Applying Lemma 2.5.2 with the property $\mathrm{Re}\,(\mu_+(\xi)) > 0$ for $\mathrm{Re}\,(\xi) > 0$, we can prove

$$|z_-(\xi)| = \left|\frac{-\mu_+(\xi)\Delta x}{\varepsilon} + \sqrt{\left(\frac{-\mu_+(\xi)\Delta x}{\varepsilon}\right)^2 + 1}\right| < 1,$$

while, as a consequence, $|z_+(\xi)| > 1$. Thus, the solution in $\ell^2(\mathbb{N},\mathbb{C})$ of (2.2.25a) can be represented as

$$\psi_j^I = z_-^j(\xi)R_1.$$

Similarly, the solution of (2.2.25b) can be represented as

$$\psi_j^{II} = w_+^j(\xi)R_2,$$

with $R_2 \in \mathbb{C}$ and

$$w_\pm(\xi) = \frac{\mu_-(\xi)\Delta x}{\varepsilon} \pm \sqrt{\left(\frac{\mu_-(\xi)\Delta x}{\varepsilon}\right)^2 + 1}, \tag{2.2.28}$$

that satisfies $|w_+(\xi)| < 1$ by again using Lemma 2.5.2 together with the property $\mathrm{Re}\,(\mu_-(\xi)) < 0$ for $\mathrm{Re}\,(\xi) > 0$. Again the other root satisfies $|w_-(\xi)| > 1$.

Finally, the solution $(\phi_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N},\mathbb{C}^2)$ of the two-dimensional problem (2.2.21a) has the following form

$$\phi_j = \overline{P}(\xi)Z^j(\xi)R, \tag{2.2.29}$$

with $Z(\xi) = \mathrm{diag}(z_-(\xi), w_+(\xi))$, and some $R = (R_1, R_2)^T \in \mathbb{C}^2$ that remains undetermined at this level.

Plugging now (2.2.29) into the boundary conditions (2.2.21b) and (2.2.21c), R has to satisfy the equations

$$\begin{aligned} &B\overline{P}R = 0, \\ &\Pi_2 P^{-T}HA\overline{P}\left(Z(\xi) - \left(I + \frac{\Delta x}{\varepsilon}D\right)\right)R = 0. \end{aligned} \tag{2.2.30}$$

Let us introduce the following quantities

$$\begin{aligned} g(\xi) &= \frac{a\mu_+(\xi)}{1+\xi}, \quad k(\xi) = \frac{a\mu_-(\xi)}{1+\xi}, \\ \delta_1(\xi) &= z_-(\xi) - \left(1 + \frac{\mu_+(\xi)\Delta x}{\varepsilon}\right), \\ \delta_2(\xi) &= w_+(\xi) - \left(1 + \frac{\mu_-(\xi)\Delta x}{\varepsilon}\right). \end{aligned} \tag{2.2.31}$$

Thus, the system (2.2.30) can be reformulated simply as a linear system

$$N(\xi)R = 0,$$

where we set

$$N(\xi) = \begin{pmatrix} B_u + k(\xi)B_v & B_u + g(\xi)B_v \\ -a\delta_1(\xi)(B_u - k(\xi)B_v) & -a\delta_2(\xi)(B_u - g(\xi)B_v) \end{pmatrix}. \tag{2.2.32}$$

Proposition 2.2.6. *Assume $B_uB_v > 0$ and consider some parameters $\Delta x, \varepsilon > 0$. For any $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) \geq 0$, one has $\det N(\xi) \neq 0$.*

In other words, the proposition states that, under the sufficient condition $B_uB_v > 0$, the scheme (2.2.7) with homogeneous boundary condition does not admit unstable solution of the form (2.2.20) in $\ell^2(\mathbb{N}, \mathbb{C}^2)$.

Proof. We again omit the explicit reference to ξ in the notations, assuming $\mathrm{Re}\,(\xi) \geq 0$ all along this proof. From the definition (2.2.32) and observing $k = -g$, the quantity $\det N$ reads also

$$\det N = a\left(\delta_1(B_u + gB_v)^2 - \delta_2(B_u - gB_v)^2\right),$$

Therefore, we have

$$\det N \neq 0 \Leftrightarrow \left|1 - \frac{\delta_2}{\delta_1}\left(\frac{B_u - gB_v}{B_u + gB_v}\right)^2\right| |\delta_1| \left|B_u + gB_v\right|^2 \neq 0. \tag{2.2.33}$$

- Firstly, we prove that

$$\left|\frac{\delta_2}{\delta_1}\left(\frac{B_u - gB_v}{B_u + gB_v}\right)^2\right| \neq 1.$$

Let $\delta = \Delta x/\varepsilon$, then since $\mu_- = -\mu_+$, we have

$$\left|\frac{\delta_2}{\delta_1}\right| \leq 1 \Leftrightarrow \mathrm{Re}\,\left(\sqrt{(\mu_+\delta)^2 + 1}\right) \geq 0. \tag{2.2.34}$$

Furthermore, the complex function $g(\xi)$ is analytic and bounded in $\mathrm{Re}\,(\xi) \geq 0$. By the conformal mapping theorem, $g(\xi)$ maps the half plane $\mathrm{Re}\,(\xi) \geq 0$ to a simply connected closed bounded domain $\Omega \subset \mathbb{C}$ whose boundary corresponds to the image of the imaginary axis $\mathrm{Re}\,(\xi) = 0$ under g. The boundary curve

$$g(i\beta) = \frac{\sqrt{-a\beta^2 + a\beta i}}{1 + i\beta}, \quad -\infty \leq \beta \leq \infty$$

is a closed curve which intersects the real axis only at $\beta = 0$ and at $\beta = \pm\infty$ with $g(0) = 0$, $g(\pm i\infty) = \sqrt{a}$. Besides, the curve is transversal to the real axis.

Since $B_uB_v > 0$, $\mathrm{Re}\,(g(\xi)) \geq 0$ in $\mathrm{Re}\,(\xi) \geq 0$, we observe that

$$\left|\frac{B_u - gB_v}{B_u + gB_v}\right|^2 \leq 1 \Leftrightarrow \mathrm{Re}\,(g) \geq 0. \tag{2.2.35}$$

According to (2.2.34) and (2.2.35), we obtain

$$\left|\frac{\delta_2}{\delta_1}\left(\frac{B_u - gB_v}{B_u + gB_v}\right)^2\right| \leq 1. \tag{2.2.36}$$

Now, we assume by contradiction that for some point ξ with $\operatorname{Re}(\xi) \geq 0$, the following occurs

$$\left|\frac{\delta_2}{\delta_1}\left(\frac{B_u - gB_v}{B_u + gB_v}\right)^2\right| = 1 \Leftrightarrow \begin{cases} \left|\dfrac{\delta_2}{\delta_1}\right| = 1, \\ \left|\dfrac{B_u - gB_v}{B_u + gB_v}\right|^2 = 1 \end{cases} \Leftrightarrow \begin{cases} \operatorname{Re}\left(\sqrt{(\mu_+\delta)^2 + 1}\right) = 0, \\ \xi = 0. \end{cases}$$

Since $\xi = 0$, we have $\operatorname{Re}\left(\sqrt{(\mu_+\delta)^2 + 1}\right) = 1$ and thus we conclude that

$$\left|\frac{\delta_2}{\delta_1}\left(\frac{B_u - gB_v}{B_u + gB_v}\right)^2\right| \neq 1. \tag{2.2.37}$$

According to (2.2.36) and (2.2.37) we have

$$\left|1 - \frac{\delta_2}{\delta_1}\left(\frac{B_u - gB_v}{B_u + gB_v}\right)^2\right| \neq 0.$$

- Secondly, with $\delta = \Delta x/\varepsilon$, we get $|\delta_1| = |1 + \sqrt{(\mu_+\delta)^2 + 1}| \geq 1$ and

$$|B_u + gB_v|^2 \geq B_u^2 > 0,$$

due to the facts $B_u B_v > 0$ and $\operatorname{Re}(g(\xi)) \geq 0$ for $\operatorname{Re}(\xi) \geq 0$.
Therefore, we proved $\det N \neq 0$. □

2.2.2.2 Numerical experiments for the necessity of the boundary condition

Using the normal mode analysis, we prove that the strict dissipativity condition (2.2.8) is sufficient to preclude the existence of unstable solutions of the form (2.2.20). However, we are not able to prove that this condition is also necessary, i.e. that there exists an unstable solution to (2.2.7) with homogeneous boundary condition as soon as $B_u B_v < 0$. We first present hereafter numerical results in this advantageous case (2.2.8), concerning the quantity $|\det N|$ introduced in (2.2.32). Then we also perform a numerically study for situations with $B_u B_v < 0$, and more importantly when $B_u/B_v < -\sqrt{a}$, which is a sub-case of the SKC (2.1.6).

Let us denote the following quantity of interest, depending on $\xi \in \mathbb{C}$, $\delta = \Delta x/\varepsilon > 0$ and the boundary parameters (B_u, B_v) through their ratio:

$$F\left(\xi, \delta, \frac{B_u}{B_v}\right) = \delta_1\left(\frac{B_u}{B_v} + g(\xi)\right)^2 - \delta_2\left(\frac{B_u}{B_v} - g(\xi)\right)^2,$$

with quantities introduced in (2.2.31), and recall that we have

$$\det N \neq 0 \Leftrightarrow F\left(\xi, \delta, \frac{B_u}{B_v}\right) \neq 0.$$

Our numerical study is based on two complementary methods. The first one corresponds to the display of three-dimensional data in two dimensions using contours or color-coded regions. We draw contour lines of the quantity $|F(\xi, \delta, B_u/B_v)|$ in the complex plane for ξ, thus computed from a grid of $\operatorname{Re}(\xi)$ values in the horizontal axis and a grid of $\operatorname{Im}(\xi)$ values in the vertical axis. For each fixed parameters δ and B_u/B_v, a contour line is then a curve in the ξ-plane along which the function $|F(\xi, \delta, B_u/B_v)|$ has a constant value, so that any curve joins points with equal values.

To know whether or not the function F vanishes at some point ξ, which is a property that the contour lines may hardly support, we also test numerically the argument principle for the following contour integral

$$I(\xi_0, R, \delta, B_u/B_v) = \frac{1}{2\pi i}\int_{\mathcal{D}} \frac{F'(\xi, \delta, B_u/B_v)}{F(\xi, \delta, B_u/B_v)} d\xi. \tag{2.2.38}$$

The involved contour curve is some positively oriented circle $\mathcal{D} \subset \{\xi \in \mathbb{C} \ : \ \mathrm{Re}\,(\xi) > 0\}$ defined by

$$\mathcal{D} = \{\xi \in \mathbb{C}, |\xi - \xi_0| = R\} = \left\{\xi_0 + Re^{i\theta}, \theta \in (0, 2\pi]\right\},$$

where the parameters ξ_0 and $R > 0$ are chosen by hand from the contour plots. The numerical approximation of the integral (2.2.38) is obtained thanks to the trapezoidal rule on a uniformly distributed grid $\xi(\theta_j) = \xi_0 + Re^{i\theta_j}$, where $\theta_j = 2j\pi/N$ for $1 \leq j \leq N$ for some large integer N. This computation benefits from the well-known spectral accuracy of the method for periodic integrand (see for example [89]). The numerator of (2.2.38) with values $F'(\xi(\theta_j), \delta, B_u/B_v)$ is approximated thanks to a spectral differentiation method [88]. We thus obtain approximations that we denote $\widehat{DF}(\xi(\theta_j), \delta, B_u/B_v)$. This approximation uses the discrete Fourier transform and only the pointwise evaluation of F on the grid. It also has spectral accuracy for large N. Finally, as an approximation of $I(\xi_0, R, \delta, B_u/B_v)$, we consider the following quantity:

$$I_N(\xi_0, R, \delta, B_u/B_v) := \frac{-i}{N}\sum_{j=1}^{N} \frac{\widehat{DF}(\xi(\theta_j), \delta, B_u/B_v)}{F(\xi(\theta_j), \delta, B_u/B_v)}.$$

The function F being holomorphic in the half plane $\mathrm{Re}\,(\xi) > 0$, this approximation precisely counts the number of zeros (with multiplicities) inside the contour $\mathcal{D}$.

In any case, we choose $a = 4$ so that SKC condition reads $B_u/B_v \notin [-2, 0]$.

Firstly, for all $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) > 0$, we consider the values of $|F\,(\xi, \delta, B_u/B_v)|$ with parameters $\delta = 10$ and $B_u/B_v = 1/40 > 0$ (Figure 2.2.3 left) and then with parameters $\delta = 10^{-2}$ and $B_u/B_v = 1 > 0$ (Figure 2.2.3 right).

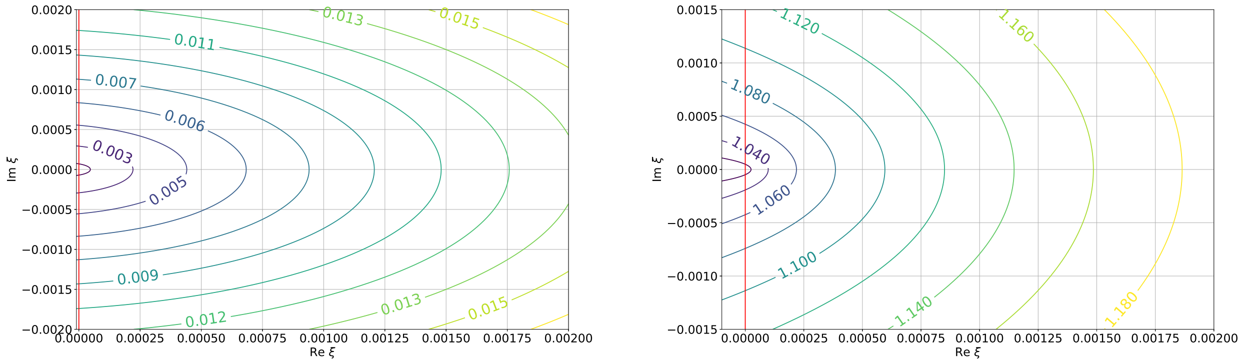

Figure 2.2.3: Contour plot of $\xi \mapsto |F\,(\xi, \delta, B_u/B_v)\,|$. The parameters are $\delta = 10$, $B_u/B_v = 1/40$ (left) and $\delta = 10^{-2}$, $B_u/B_v = 1$ (right).

In both case, we observe that there exists a constant $C > 0$ such that $|F\,(\xi, \delta, B_u/B_v)\,| \geq C$ in the half plane $\mathrm{Re}\,(\xi) > 0$. From our experiments, the constant $C = 10^{-3}$ seems suitable in the first case and $C = 1$ in the second one. Actually, the above observations will be confirmed rigorously in Section 2.2.3.2.

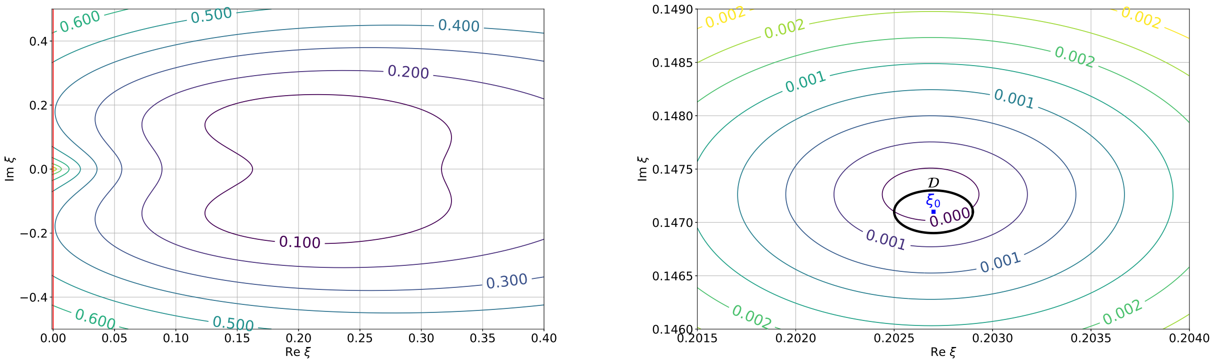

Figure 2.2.4: Contour plot of $\xi \mapsto |F(\xi, 1, -1)|$, for $\operatorname{Re}(\xi) > 0$ (left) and a close-up near a supposed zero (right).

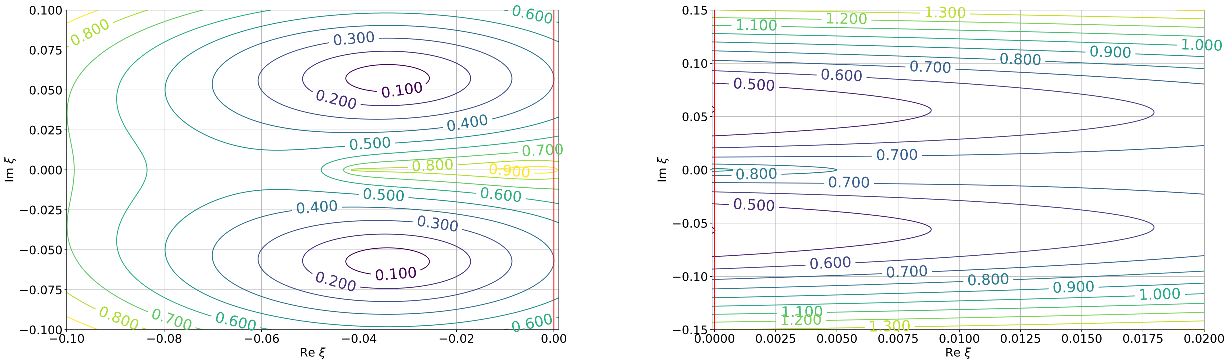

Figure 2.2.5: Contour plot of $\xi \mapsto |F(\xi, 10, -1)|$, for $\operatorname{Re}(\xi) < 0$ (left) and $\operatorname{Re}(\xi) > 0$ (right).

Secondly, for all $\xi \in \mathbb{C}$ with $\operatorname{Re}(\xi) > 0$, we consider the values of $|F(\xi, \delta, B_u/B_v)|$ with parameters $B_u/B_v \in [-\sqrt{a}, 0]$ and various values for $\delta > 0$. More precisely with choose $B_u/B_v = -1$ together with $\delta = 1$ (Figure 2.2.4) and $\delta = 10$ (Figure 2.2.5).

In the first case, the contour lines promote the existence of some $\xi \in \mathbb{C}$ with $\operatorname{Re}(\xi) > 0$ satisfying $|F(\xi, 1, -1)| \ll 1$. Therefore, we consider the circled curve $\mathcal{D}$ with parameters $\xi_0 = 0.2027 + 0.1471i$ and $R = 2 \times 10^{-4}$. According to Table 2.1, we compute the contour integral and for large integers N, we get $I_N(\xi_0, R, 1, -1) = 1$ up to the machine epsilon. Thus, there exists exactly one complex number ξ inside the contour $\mathcal{D}$ such that $F(\xi, 1, -1) = 0$ and a corresponding unstable solution. In the case $\delta = 10$, for any $\operatorname{Re}(\xi) > 0$ then $|F(\xi, 10, -1)| \neq 0$ (Figure 2.2.5). Therefore, we can not prove that for any $\delta > 0$, $\xi \in \mathbb{C}$, $\operatorname{Re}(\xi) > 0$, if $B_u/B_v \in [-\sqrt{a}, 0]$ then $|F(\xi, \delta, B_u/B_v)| \neq 0$.

Thirdly, for all $\xi \in \mathbb{C}$ with $\operatorname{Re}(\xi) > 0$, we consider the boundary parameter $B_u/B_v = -3.5$ so that $B_u/B_v < -\sqrt{a}$. Let us recall that in the continuous case, Z. Xin and W. Xu [96] proved that there is no unstable solution in that case. Contrasting with this result, for the discrete IBVP (2.2.7), the next numerical experiments support the following conjecture to hold true.

Conjecture 2.2.7. *Consider the case $B_u/B_v < -\sqrt{a}$. There exist $\delta > 0$ and $\xi \in \mathbb{C}$ with $\operatorname{Re}(\xi) > 0$ such that $\det N(\xi) = 0$. In other words, there exists an unstable solution of (2.2.7) of the form (2.2.20).*

Now, we study the behavior of $|F(\xi, \delta, -3.5)|$ with successively $\delta = 10$ and $\delta = 10^{-2}$ (Figure 2.2.6). We can see that in the case $\delta = 10$, for all $\xi \in \mathbb{C}$ with $\operatorname{Re}(\xi) > 0$, the quantity $|F(\xi, \delta, -3.5)|$ seems to be positively bounded from below.

In the case $\delta = 10^{-2}$ however the contour lines promote the existence of some $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) > 0$ satisfying $|F(\xi, 10^{-2}, -3.5)| \ll 1$. Therefore we consider the circled curve $\mathcal{D}$ with parameters $\xi_0 = 0.23 + 101.55i$ and $R = 10^{-2}$. According to Table 2.1, we compute the contour integral and for large integers N, we get $I_N(\xi_0, R, 10^{-2}, -3.5) = 1$ up to the machine epsilon. Thus, there exists a complex number ξ inside the contour $\mathcal{D}$ such that $F(\xi, 10^{-2}, -3.5) = 0$ and a corresponding unstable solution.

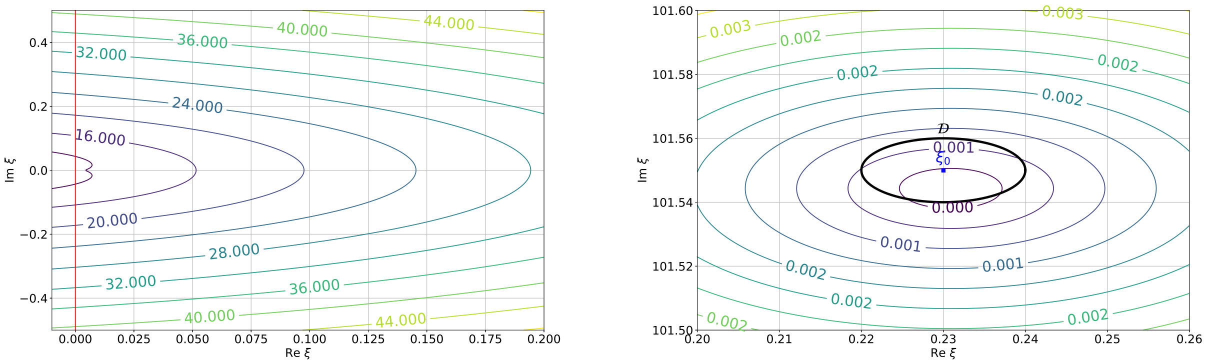

Figure 2.2.6: Contour plot of $\xi \mapsto |F(\xi, \delta, -3.5)|$ with $\delta = 10$ (left) and $\delta = 10^{-2}$ (right).

N	$I_N(0.2027 + 0.1471i, 2 \times 10^{-4}, 1, -1)$	$I_N(0.23 + 101.55i, 10^{-2}, 10^{-2}, -3.5)$
20	$0.9948572383921 + 0.019072730887644i$	$0.9999842664632257 - 2.3902006024 \times 10^{-8}i$
40	$0.9996520507698 - 0.000184801108269i$	$1.00000000024755 + 7.52237161449 \times 10^{-13}i$
80	$1.0000000869301 + 1.287245215 \times 10^{-7}i$	$0.9999999999999999 + 4.440892098 \times 10^{-17}i$
160	$0.9999999999999 + 2.225615525 \times 10^{-14}i$	$0.9999999999999999 + 3.747002708 \times 10^{-17}i$
320	$1.0000000000000007 + 1.30104 \times 10^{-17}i$	$0.9999999999999999 + 1.061650767 \times 10^{-16}i$

Table 2.1: The contour integral I_N.

2.2.3 Stiff stability of the semi-discrete IBVP with homogeneous initial condition

For convenience in the forthcoming discussions, we recall that the semi-discrete approximation of the IBVP (2.2.7) with homogeneous initial condition reads

$$\begin{cases} \partial_t U_j(t) + A\dfrac{U_{j+1}(t) - U_{j-1}(t)}{2\Delta x} = \varepsilon^{-1} S U_j(t), & j \geq 1,\ t \geq 0, \\ U_j(0) = 0, & j \geq 0, \\ BU_0(t) = b(t), & t \geq 0, \\ \partial_t \left(\Pi_2 P^{-T} H U_0\right)(t) + \Pi_2 P^{-T} H A \dfrac{U_1(t) - U_0(t)}{\Delta x} = \varepsilon^{-1} \Pi_2 P^{-T} H S U_0(t), & t \geq 0. \end{cases} \tag{2.2.39}$$

Dealing with difference approximations, the Laplace transform is already the more powerful tool for problems in one space dimension. It is used to determine stability features when the energy method is not sufficient. Under the strict dissipativity condition (2.2.8), the numerical solution $(U_j(t))_{j\in\mathbb{N}}$ can be constructed by the method of Laplace transform. By using the Parseval's identity, we get the expected result of Proposition 2.2.4

2.2.3.1 Solution by Laplace transform

The numerical solution $U_j(t)$ of the IBVP (2.2.39) can be constructed by the method of Laplace transform. Let

$$\widetilde{U}_j(\xi) = \mathcal{L}U_j = \int_0^\infty e^{-\xi t} U_j(t)dt, \quad \mathrm{Re}\,(\xi) > 0.$$

With $U_j(0) \equiv 0$, we have

$$\mathcal{L}(\partial_t U_j) = \xi \widetilde{U}_j(\xi) - U_j(0) = \xi \widetilde{U}_j(\xi)$$

and therefore the system (2.2.39) becomes

$$\widetilde{U}_{j+1}(\xi) - \widetilde{U}_{j-1}(\xi) = \frac{2\Delta x}{\varepsilon} M(\varepsilon\xi)\widetilde{U}_j(\xi), \quad j > 0, \tag{2.2.40}$$

$$B\widetilde{U}_0(\xi) = \widetilde{b}(\xi), \tag{2.2.41}$$

$$\Pi_2 P^{-T} HA \left(\widetilde{U}_1(\xi) - \left(I + \frac{\Delta x}{\varepsilon} M(\varepsilon\xi) \right) \widetilde{U}_0(\xi) \right) = 0, \tag{2.2.42}$$

where

$$\widetilde{b}(\xi) = \mathcal{L}b = \int_0^\infty e^{-\xi t} b(t)dt$$

and the matrix $M(\varepsilon\xi)$ is the same as in (2.2.22).

Note that the eigenvalues $\mu_\pm(\xi)$ of the matrix $M(\xi)$ satisfy

$$\mathrm{Re}\,\mu_-(\xi) < 0, \quad \mathrm{Re}\,\mu_+(\xi) > 0, \quad \text{for } \mathrm{Re}\,\xi > 0.$$

One can proceed as in (2.2.25)-(2.2.29) to find the solution $\widetilde{U}_j(\xi)$ of (2.2.40). For some vector $R \in \mathbb{C}^2$, it takes the form

$$\widetilde{U}_j(\xi) = \overline{P}(\varepsilon\xi) Z^j(\varepsilon\xi) R.$$

The value of R can be determined easily from the boundary condition (2.2.41) and (2.2.42)

$$R = \frac{\widetilde{b}(\xi)}{\det N(\varepsilon\xi)} N_1(\varepsilon\xi),$$

where

$$N_1(\xi) = a \begin{pmatrix} -\delta_2(\xi)(B_u - g(\xi)B_v) \\ \delta_1(\xi)(B_u + g(\xi)B_v) \end{pmatrix} \tag{2.2.43}$$

and the matrix $N(\varepsilon\xi)$ is the same as in (2.2.32). Therefore,

$$\widetilde{U}_j(\xi) = \frac{\widetilde{b}(\xi)}{\det N(\varepsilon\xi)} \overline{P}(\varepsilon\xi) Z^j(\varepsilon\xi) N_1(\varepsilon\xi).$$

With $\widetilde{U}_j(\xi)$ found, the numerical solution $U_j(t)$ of (2.2.39) can be obtained by inverting the Laplace transform

$$U_j(t) = \mathcal{L}^{-1}\widetilde{U}_j(\xi) = \frac{1}{2\pi} \int_{-\infty}^\infty e^{(\alpha + i\beta)t} \widetilde{U}_j(\alpha + i\beta) d\beta, \quad \alpha > 0.$$

2.2.3.2 Stiff stability analysis

Under the strict dissipativity condition $B_uB_v > 0$, we consider Proposition 2.2.4 with homogeneous initial condition $(f_j)_{j\in\mathbb{N}} \equiv 0$ and nonzero boundary data $b(t)$. Actually, we will need a more stringent version of the estimate (2.2.37) uniform in $\delta > 0$ and $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) \geq 0$. This is the object of the next lemma.

Lemma 2.2.8. *Assume $B_uB_v > 0$. There exists $c \in (0,1)$ such that for any $\delta = \dfrac{\Delta x}{\varepsilon} > 0$ and $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) \geq 0$*

$$\left| \frac{\delta_2(\xi)}{\delta_1(\xi)} \left(\frac{B_u - g(\xi)B_v}{B_u + g(\xi)B_v} \right)^2 \right| \leq 1 - c. \tag{2.2.44}$$

where g, δ_1 and δ_2 are defined in (2.2.31).

Proof. Firstly, from (2.2.34) and (2.2.35) we already observed, assuming $B_uB_v > 0$, that for any $\delta > 0$ and $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) \geq 0$

$$\left| \frac{\delta_2(\xi)}{\delta_1(\xi)} \right| \leq 1 \text{ and } |\tau(g(\xi))| \leq 1,$$

where we denote

$$\tau(g) = \frac{B_u - gB_v}{B_u + gB_v}.$$

Furthermore, the function $g(\xi)$ maps the half plane $\mathrm{Re}\,(\xi) \geq 0$ to a simply connected closed bounded domain Ω. Thus, $|\tau(g(\xi))|$ tends to 1 only if $\mathrm{Re}\,(g(\xi))$ goes to 0.

Secondly, let $\xi = \alpha + i\beta$ with $\alpha \geq 0$ and $\beta \in \mathbb{R}$, after some calculations, one obtains

$$\mathrm{Re}\,(g(\xi)) = \sqrt{\frac{a\left(p + \sqrt{p^2+q^2}\right)}{(1+\alpha)^2 + \beta^2}},$$

where

$$p = \alpha(1+\alpha) + \beta^2, \quad q = \beta.$$

Thus, for all $\alpha \geq 0$, $\beta \in \mathbb{R}$, $\mathrm{Re}\,(g(\xi))$ goes to 0 only if ξ tends to 0. Therefore, outside a neighborhood of 0 in $\mathrm{Re}\,(\xi) \geq 0$

$$|\tau(g(\xi))| \leq c < 1.$$

Moreover, for any $\delta > 0$ and $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) \geq 0$, the quantity $\left| \dfrac{\delta_2(\xi)}{\delta_1(\xi)} \right|$ tends to 1 only if $\mathrm{Re}\,\left(\sqrt{(\mu_+(\xi)\delta)^2 + 1}\right)$ goes to 0. However, in a neighborhood of 0 in $\mathrm{Re}\,(\xi) \geq 0$, for any $\delta > 0$, $\mathrm{Re}\,\left(\sqrt{(\mu_+(\xi)\delta)^2 + 1}\right) \geq c_1 > 0$ (for the details, we refer the reader to the technical Lemma 2.5.3). Thus,

$$\left| \frac{\delta_2(\xi)}{\delta_1(\xi)} \right| \leq c_2 < 1,$$

and the result follows. □

Proposition 2.2.9. *Assume $B_uB_v > 0$. There exists $C > 0$ such that for any $\delta = \dfrac{\Delta x}{\varepsilon} > 0$ and $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) \geq 0$*

$$\frac{|\det N(\xi)|^2}{\|N_1(\xi)\|^2} \geq CB_u^2.$$

Before we prove the above result, let us notice that it easily implies the previous Proposition 2.2.6. Actually, the reader has to understand this result as being the Uniform version of the previous one, in the same way the UKC is the uniform version of the Kreiss Condition for continuous hyperbolic PDEs, or the discrete UKC is the uniform version of the Godunov Ryabenkii condition for the (semi-)discrete IBVP, except now we also deal with the parameters ε and Δx (or equivalently with the single parameter δ).

Proof. From (2.2.32) and (2.2.43), omitting the explicit dependence in ξ, the we can compute on the one hand

$$|\det N|^2 = a^2 \left| 1 - \frac{\delta_2}{\delta_1} \left(\frac{B_u - gB_v}{B_u + gB_v} \right)^2 \right|^2 |\delta_1|^2 \, |B_u + gB_v|^4$$

and on the other hand

$$\|N_1\|^2 = a^2 \left(1 + \left| \frac{\delta_2}{\delta_1} \left(\frac{B_u - gB_v}{B_u + gB_v} \right)^2 \right| \right) |\delta_1|^2 \, |B_u + gB_v|^2 .$$

Thus we have the explicit formula

$$\frac{|\det N|^2}{\|N_1\|^2} = \left| 1 - \frac{\delta_2}{\delta_1} \left(\frac{B_u - gB_v}{B_u + gB_v} \right)^2 \right|^2 \left(1 + \left| \frac{\delta_2}{\delta_1} \left(\frac{B_u - gB_v}{B_u + gB_v} \right)^2 \right| \right)^{-1} |B_u + gB_v|^2 .$$

Let us investigate separately any of the above terms. According to Lemma 2.2.8, there exists $c > 0$ such that for any $\xi \in \mathbb{C}$, $\mathrm{Re}\,(\xi) \geq 0$ and $\delta > 0$,

$$\left| 1 - \frac{\delta_2}{\delta_1} \left(\frac{B_u - gB_v}{B_u + gB_v} \right)^2 \right|^2 \geq c$$

and from (2.2.36), we have

$$\left(1 + \left| \frac{\delta_2}{\delta_1} \left(\frac{B_u - gB_v}{B_u + gB_v} \right)^2 \right| \right)^{-1} \geq 1/2.$$

Since $B_u B_v > 0$ and $\mathrm{Re}\,(g(\xi)) \geq 0$ for $\mathrm{Re}\,(\xi) \geq 0$, we finally get

$$|B_u + gB_v|^2 \geq B_u^2.$$

Therefore, there exists $C > 0$ such that

$$\frac{|\det N|^2}{\|N_1\|^2} \geq C B_u^2.$$

□

Now, we prove the uniform ℓ^2 estimate (2.2.12). By an application of the following Parseval's identity [42][85]:

$$\int_0^\infty e^{-2\alpha t} |U_j(t)|^2 dt = \frac{1}{2\pi} \oint_{-\infty}^{\infty} |\widetilde{U}_j(\alpha + i\beta)|^2 d\beta, \quad \alpha > 0,$$

we have

$$\begin{aligned}\int_0^\infty e^{-2\alpha t}|U_0(t)|^2dt &= \frac{1}{2\pi}\oint_{-\infty}^{\infty}|\widetilde{U}_0(\alpha+i\beta)|^2d\beta\\ &= \frac{1}{2\pi}\oint_{-\infty}^{\infty}|\widetilde{b}(\xi)|^2\frac{\|N_1(\varepsilon\xi)\|^2}{|\det N(\varepsilon\xi)|^2}|\overline{P}(\varepsilon\xi)|^2d\beta.\end{aligned}$$

where $\xi=\alpha+i\beta$. We fix $\alpha>0$ from now on.
According to Proposition 2.2.9, there exists $C_1>0$ such that for any $\delta>0$, $\xi\in\mathbb{C}$, $\mathrm{Re}\,(\xi)\geq 0$,

$$\frac{\|N_1(\varepsilon\xi)\|^2}{|\det N(\varepsilon\xi)|^2}\leq C_1.$$

On the other hand, since $k(\xi)=-g(\xi)$ is uniformly bounded in $\mathrm{Re}\,(\xi)\geq 0$, we obtain

$$\begin{aligned}\int_0^\infty e^{-2\alpha t}|U_0(t)|^2dt &\lesssim \frac{1}{2\pi}\oint_{-\infty}^{\infty}|\widetilde{b}(\alpha+i\beta)|^2d\beta\\ &\lesssim \int_0^\infty e^{-2\alpha t}|b(t)|^2dt.\end{aligned} \tag{2.2.45}$$

This, together with a consequence of the hyperbolicity of (2.1.1) by using the classical argument of changing the data b to zero after time T and unchanged before time T, we obtain the desired boundary estimate

$$\int_0^T|U_0(t)|^2dt\leq K_T\int_0^T|b(t)|^2dt. \tag{2.2.46}$$

Similarly, by an application of the Parseval's identity, we have

$$\begin{aligned}\int_0^\infty\sum_{j\geq 0}e^{-2\alpha t}|U_j(t)|^2dt &= \frac{1}{2\pi}\oint_{-\infty}^{\infty}\sum_{j\geq 0}|\widetilde{U}_j(\alpha+i\beta)|^2d\beta\\ &= \frac{1}{2\pi}\oint_{-\infty}^{\infty}\sum_{j\geq 0}|\widetilde{b}(\xi)|^2\frac{\|N_1(\varepsilon\xi)\|^2}{|\det N(\varepsilon\xi))|^2}|\overline{P}(\varepsilon\xi)|^2(|z_-(\varepsilon\xi)|^{2j}+|w_+(\varepsilon\xi)|^{2j})d\beta,\end{aligned}$$

where $z_-(\varepsilon\xi)$ and $w_+(\varepsilon\xi)$ are the same as in (2.2.27) and (2.2.28).

Since $k(\varepsilon\xi)=-g(\varepsilon\xi)$ is uniformly bounded in $\mathrm{Re}\,(\xi)\geq 0$, $\varepsilon>0$ and using Proposition 2.2.9, we obtain

$$\int_0^\infty\sum_{j\geq 0}e^{-2\alpha t}|U_j(t)|^2dt\lesssim\frac{1}{2\pi}\oint_{-\infty}^{\infty}\sum_{j\geq 0}|\widetilde{b}(\xi)|^2(|z_-(\varepsilon\xi)|^{2j}+|w_+(\varepsilon\xi)|^{2j})d\beta.$$

On the other hand, since $\mu_-(\varepsilon\xi)=-\mu_+(\varepsilon\xi)$ in $\mathrm{Re}\,(\xi)\geq 0$, $\varepsilon>0$, we get $|z_-(\varepsilon\xi)|=|w_+(\varepsilon\xi)|$, and thus

$$\int_0^\infty\sum_{j\geq 0}e^{-2\alpha t}|U_j(t)|^2dt\lesssim\frac{1}{2\pi}\oint_{-\infty}^{\infty}\sum_{j\geq 0}|w_+(\varepsilon\xi)|^{2j}|\widetilde{b}(\xi)|^2d\beta.$$

According to (2.2.24), for all $\varepsilon>0$, $\xi\in\mathbb{C}$, $\mathrm{Re}\,(\xi)>0$, we have

$$\mathrm{Re}\ (\mu_-(\varepsilon\xi))\leq-\frac{\varepsilon\mathrm{Re}\,(\xi)}{\sqrt{a}}<0.$$

Furthermore, we can prove

$$\left(\frac{\mathrm{Re}\ (\mu_-(\varepsilon\xi))\,\Delta x}{\varepsilon}+\sqrt{\left(\frac{\mathrm{Re}\ (\mu_-(\varepsilon\xi))\,\Delta x}{\varepsilon}\right)^2+1}\right)^2\leq\left(\eta\Delta x+\sqrt{\eta^2\Delta x^2+1}\right)^2, \tag{2.2.47}$$

where $\eta = -\dfrac{\mathrm{Re}\,(\xi)}{\sqrt{a}}$. According to Lemma 2.5.2 and (2.2.47), we have now

$$\begin{aligned}\sum_{j\geq 0} |w_+(\varepsilon\xi)|^{2j} &= \left(1 - \left|\frac{\mu_-(\varepsilon\xi)\Delta x}{\varepsilon} + \sqrt{\left(\frac{\mu_-(\varepsilon\xi)\Delta x}{\varepsilon}\right)^2 + 1}\right|^2\right)^{-1} \\ &\leq \left(1 - \left(\eta\Delta x + \sqrt{\eta^2\Delta x^2 + 1}\right)^2\right)^{-1}.\end{aligned}$$

If we assume that $\Delta x \leq -\dfrac{3}{4\eta}$ then

$$\sum_{j\geq 0} |w_+(\varepsilon\xi)|^{2j} \leq -\eta^{-1}\Delta x^{-1},$$

and therefore, by an application of the Parseval's identity

$$\alpha\Delta x \int_0^\infty \sum_{j\geq 0} e^{-2\alpha t}|U_j(t)|^2 dt \lesssim \frac{1}{2\pi}\oint_{-\infty}^{\infty} |\widetilde{b}(\xi)|^2 d\beta \lesssim \int_0^\infty e^{-2\alpha t}|b(t)|^2 dt. \tag{2.2.48}$$

According to (2.2.45) and (2.2.48), there exists $C > 0$ such that

$$\alpha\Delta x \int_0^\infty \sum_{j\geq 0} e^{-2\alpha t}|U_j(t)|^2 dt + \int_0^\infty e^{-2\alpha t}|U_0(t)|^2 dt \leq C\int_0^\infty e^{-2\alpha t}|b(t)|^2 dt. \tag{2.2.49}$$

This ends the proof of Proposition 2.2.4.

To complete the proof of Theorem 2.2.2, observe that from (2.2.46), (2.2.48) and the hyperbolicity of (2.1.1), for any $T > 0$, there exists a constant $C_T > 0$ such that

$$\int_0^T \sum_{j\geq 0} \Delta x|U_j(t)|^2 dt + \int_0^T |U_0(t)|^2 dt \leq C_T \int_0^T |b(t)|^2 dt. \tag{2.2.50}$$

By linearity, we can break up the IBVP (2.2.7) into two simpler problems, one with homogeneous initial condition and the other with homogeneous boundary condition. Finally from (2.2.19) and (2.2.50), we get the expected result of Theorem 2.2.2.

2.3 The time-implicit scheme

Let $\Delta t > 0$ be the time step. The space step $\Delta x > 0$ will always be chosen so that the parameter $\delta_{xt} = \Delta x \Delta t^{-1}$ is fixed. Letting now $U_j^n = \left(u_j^n, v_j^n\right)^T$ denote the approximation of the exact solution to (2.1.1)-(2.1.3) at the grid point $(x_j, t^n) = (j\Delta x, n\Delta t)$, for any $(j,n) \in \mathbb{N}\times\mathbb{N}$ (where we omit the explicit dependence on ε). We focus in this section on the fully discrete approximation of the IBVP (2.1.1)-(2.1.3) obtained by the central differencing scheme in space and the implicit scheme in time

$$\begin{cases} U_j^{n+1} - U_j^n + \Delta t(\mathcal{Q}U)_j^{n+1} = \Delta t\varepsilon^{-1}SU_j^{n+1}, & j\geq 1, n\geq 0, \\ U_j^0 = f_j, & j \geq 0, \\ BU_0^n = b^n, & n \geq 0, \\ \Pi_2 H_P P\left(U_0^{n+1} - U_0^n\right) + \Delta t\Pi_2 H_P P(\mathcal{Q}U)_0^{n+1} = \Delta t\varepsilon^{-1}\Pi_2 H_P PSU_0^{n+1}, & n \geq 0. \end{cases} \tag{2.3.1}$$

with S, B, the difference operator $(\mathcal{Q}U)_{j\in\mathbb{N}}$, P^T, H_P are the same as in (2.2.1), (2.2.4) and (2.2.5), respectively. The projection matrix Π_2 is defined by $(0 \quad 1)$.

In Section 2.2, we show that the strict dissipativity condition (2.2.8) is sufficient to derive uniform stability estimates of the semi-discrete central scheme (2.2.7). Our aim is to determine a sufficient condition for the stiff stability of the above fully discrete (2.3.1), in order words the uniform stability with respect to the stiffness of the relaxation term.

Theorem 2.3.1 (Main results)**.** *Assume that $(B_u, B_v) \in \mathbb{R}^2$ satisfies the strict dissipativity condition (2.2.8). Then, for any $T > 0$, there exists a constant $C_T > 0$ such that for all $\Delta t > 0$ and any positive constant $\delta_{xt} \leq 3\sqrt{a}/8$ together with $\Delta x = \delta_{xt}\Delta t$, any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, any $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R})$, the solution $\left(U_j^n\right)_{j\in\mathbb{N}}$ to the scheme (2.3.1) satisfies*

$$\sum_{n=0}^{N}\sum_{j\geq 0} \Delta x \Delta t \left|U_j^n\right|^2 + \sum_{n=0}^{N} \Delta t \left|U_0^n\right|^2 \leq C_T \left(\sum_{j\geq 0} \Delta x \left|f_j\right|^2 + \sum_{n=0}^{N} \Delta t \left|b^n\right|^2\right), \tag{2.3.2}$$

where $N := T/\Delta t$ and C_T is independent of $\varepsilon \in (0, +\infty)$.

By linearity, the numerical scheme of the IBVP (2.3.1) can be broken up into two simpler problems, one with homogeneous initial condition $(f_j)_{j\in\mathbb{N}} \equiv 0$ and the other with homogeneous boundary condition $b^n \equiv 0$, for any $n \in \mathbb{N}$. Following the continuous case, the proof of Theorem 2.3.1 is based on two main ingredients, by assembling a result for the case of homogeneous boundary data and another for the case with homogeneous initial condition. We state successively hereafter these two statements.

Proposition 2.3.2 (Homogeneous boundary condition)**.** *Assume that the parameters $\Delta x \in (0, 1]$, $\varepsilon > 0$ and (B_u, B_v) satisfies the discrete strict dissipativity condition (2.2.10). Then, there exists a constant $C > 0$ such that for all $\Delta t > 0$ and any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, the solution $(U_j^n)_{j\in\mathbb{N}}$ to (2.3.1) with $(b^n)_{n\in\mathbb{N}} \equiv 0$ satisfies*

$$\langle U^n, HU^n\rangle_{\Delta x} + C\Delta t \sum_{k=0}^{n} \left|U_0^k\right|^2 \leq \langle f, Hf\rangle_{\Delta x}, \quad n \in \mathbb{N}. \tag{2.3.3}$$

More precisely,

a) If $B_u B_v > 0$ then (2.3.3) holds uniformly, i.e. with C independent of ε and Δx.

b) If $B_u B_v < 0$ then considering some $\delta_0 > -2aB_v B_u^{-1}$, there exists $C(\delta_0) > 0$ such that (2.3.3) holds uniformly with $C = C(\delta_0)$, as soon as $\Delta x \geq \delta_0 \varepsilon$.

Proposition 2.3.3 (Homogeneous initial condition)**.** *Assume that the boundary condition is strictly dissipative, thus satisfying (2.2.8). Then, there exists a constant $C > 0$ such that for any $\alpha > 0$ and any positive constant $\delta_{xt} \leq 3\sqrt{a}/8$, the following property holds. For any $\Delta t > 0$ together with $\Delta x = \delta_{xt}\Delta t$ and any $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R})$, the solution $(U_j^n)_{j\in\mathbb{N}}$ to scheme (2.3.1) with $(f_j)_{j\in\mathbb{N}} \equiv 0$ satisfies*

$$\frac{\alpha}{1 + \alpha\Delta t} \sum_{n\geq 0}\sum_{j\geq 0} e^{-2\alpha n\Delta t} \Delta t \Delta x \left|U_j^n\right|^2 + \sum_{n\geq 0} e^{-2\alpha n\Delta t} \Delta t \left|U_0^n\right|^2 \leq C \sum_{n\geq 0} e^{-2\alpha n\Delta t} \Delta t \left|b^n\right|^2, \tag{2.3.4}$$

where C is independent of $\varepsilon \in (0, +\infty)$.

It seems that $B_u B_v > 0$ is also sufficient to ensure the stiff stability of the fully discrete scheme (2.3.1). The Proposition 2.3.2 is studied in Section 2.3.1 by means of the discrete energy method. In Section 2.3.2, we perform a normal mode analysis to construct unstable solutions, in order to derive necessary condition for stability. To isolate the effects of a possible boundary layer and avoid the complicated interactions of boundary and initial layers, in Section 2.3.3, we consider the IBVP (2.3.1) with homogeneous initial data and nonzero boundary data b^n, for any $n \in \mathbb{N}$. In Section 2.3.3.1, the numerical solution $\left(U_j^n\right)_{j\in\mathbb{N}}$ can be constructed by Laplace transform. By using the Plancherel's theorem, under assumption $B_u B_v > 0$, the Proposition 2.3.3 is proved in Section 2.3.3.2.

2.3.1 Stiff stability of the fully discrete IBVP with homogeneous boundary condition

In this section, we consider the discrete IBVP (2.3.1) for homogeneous boundary condition $b^n \equiv 0$, for any $n \in \mathbb{N}$, nonzero Cauchy data $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and prove Proposition 2.3.2 by means of the discrete energy method.

Using the scalar product (2.2.3) with $P^T H_P P$, we obtain

$$\begin{aligned}\frac{1}{\Delta t}\left\langle U^{n+1} - U^n, P^T H_P P U^{n+1}\right\rangle_{\Delta x} &= \frac{\delta_{xt}}{2}\left\langle H_P P\left(U_0^{n+1} - U_0^n\right), P U_0^{n+1}\right\rangle \\ &\quad + \delta_{xt}\sum_{j=1}^{+\infty}\left\langle U_j^{n+1} - U_j^n, P^T H_P P U_j^{n+1}\right\rangle.\end{aligned} \tag{2.3.5}$$

Since we use the homogeneous boundary condition $BU_0^n = 0$, for any $n \in \mathbb{N}$, and thus, $PU_0^{n+1} = \left(0, \Pi_2 P U_0^{n+1}\right)^T$, the previous equation (2.3.5) can be reformulated as

$$\begin{aligned}\frac{1}{\Delta t}\left\langle U^{n+1} - U^n, P^T H_P P U^{n+1}\right\rangle_{\Delta x} &= \frac{\delta_{xt}}{2}\Big(\Pi_2 H_P P\left(U_0^{n+1} - U_0^n\right)\Big)\left(\Pi_2 P U_0^{n+1}\right) \\ &\quad + \delta_{xt}\sum_{j=1}^{+\infty}\left\langle U_j^{n+1} - U_j^n, P^T H_P P U_j^{n+1}\right\rangle.\end{aligned} \tag{2.3.6}$$

Now, we show how the difference operator $(\mathcal{Q}U)_{j\in\mathbb{N}}$ can be applied for the discrete IBVP (2.3.1) for the homogeneous boundary condition at all gridpoints including the boundary point $j = 0$

$$\frac{1}{\Delta t}\left(U_j^{n+1} - U_j^n\right) = \frac{1}{\varepsilon} S U_j^{n+1} - \frac{1}{2\Delta x} A\left(U_{j+1}^{n+1} - U_{j-1}^{n+1}\right), \quad j \geq 1,$$

and

$$\frac{1}{\Delta t}\Pi_2 H_P P\left(U_0^{n+1} - U_0^n\right) = \frac{1}{\varepsilon}\Pi_2 H_P P S U_0^{n+1} - \frac{1}{\Delta x}\Pi_2 H_P P A\left(U_1^{n+1} - U_0^{n+1}\right).$$

As a consequence, the equation (2.3.6) can be represented as

$$\begin{aligned}&\frac{1}{\Delta t}\left\langle U^{n+1} - U^n, P^T H_P P U^{n+1}\right\rangle_{\Delta x} \\ &= \frac{\Delta x}{2\varepsilon}\left(\Pi_2 H_P P S U_0^{n+1}\right)\left(\Pi_2 P U_0^{n+1}\right) + \frac{\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left\langle S U_j^{n+1}, P^T H_P P U_j^{n+1}\right\rangle \\ &+ \frac{1}{2}\left(\Pi_2 H_P P A U_0^{n+1}\right)\left(\Pi_2 P U_0^{n+1}\right) - \frac{1}{2}\left(\Pi_2 H_P P A U_1^{n+1}\right)\left(\Pi_2 P U_0^{n+1}\right) \\ &- \frac{1}{2}\sum_{j=1}^{+\infty}\left\langle A U_{j+1}^{n+1}, P^T H_P P U_j^{n+1}\right\rangle + \frac{1}{2}\sum_{j=1}^{+\infty}\left\langle A U_{j-1}^{n+1}, P^T H_P P U_j^{n+1}\right\rangle.\end{aligned} \tag{2.3.7}$$

On the other hand, the last term in the right hand side becomes

$$\sum_{j=1}^{+\infty}\left\langle AU_{j-1}^{n+1}, P^T H_P P U_j^{n+1}\right\rangle = \left\langle AU_0^{n+1}, P^T H_P P U_1^{n+1}\right\rangle + \sum_{j=1}^{+\infty}\left\langle AU_{j+1}^{n+1}, P^T H_P P U_j^{n+1}\right\rangle.$$

Since $H_P P A P^{-1}$ is symmetric and $PU_0^{n+1} = \left(0, \Pi_2 P U_0^{n+1}\right)^T$, one gets

$$\left\langle AU_0^{n+1}, P^T H_P P U_1^{n+1}\right\rangle = \left(\Pi_2 H_P P A U_1^{n+1}\right)\left(\Pi_2 P U_0^{n+1}\right)$$

and then

$$\sum_{j=1}^{+\infty}\left\langle AU_{j-1}^{n+1}, P^T H_P P U_j^{n+1}\right\rangle = \left(\Pi_2 H_P P A U_1^{n+1}\right)\left(\Pi_2 P U_0^{n+1}\right) + \sum_{j=1}^{+\infty}\left\langle AU_{j+1}^{n+1}, P^T H_P P U_j^{n+1}\right\rangle. \tag{2.3.8}$$

Substituting (2.3.8) into (2.3.7), the three last terms in (2.3.7) vanish. After some calculations, we obtains

$$\frac{1}{\Delta t}\left\langle U^{n+1} - U^n, HU^{n+1}\right\rangle_{\Delta x} + a\frac{B_u}{B_v}\left(u_0^{n+1}\right)^2 + \frac{\Delta x}{2\varepsilon}\left(v_0^{n+1}\right)^2 = -\frac{\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left(v_j^{n+1}\right)^2. \tag{2.3.9}$$

where the symmetric positive definite H is the same as in (2.2.17).
Since H is a symmetric positive definite matrix, we can see that

$$\begin{aligned}\left\langle U^{n+1} - U^n, HU^{n+1}\right\rangle_{\Delta x} &= \frac{1}{2}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x} - \left\langle U^n, HU^n\right\rangle_{\Delta x} \\ &\quad + \left\langle U^{n+1} - U^n, H\left(U^{n+1} - U^n\right)\right\rangle_{\Delta x}\Big) \\ &\geq \frac{1}{2}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x} - \left\langle U^n, HU^n\right\rangle_{\Delta x}\Big).\end{aligned} \tag{2.3.10}$$

According to (2.3.9) and (2.3.10), one gets

$$\frac{1}{\Delta t}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x} - \left\langle U^n, HU^n\right\rangle_{\Delta x}\Big) + 2a\frac{B_u}{B_v}\left(u_0^{n+1}\right)^2 + \frac{\Delta x}{\varepsilon}\left(v_0^{n+1}\right)^2 \leq -\frac{2\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left(v_j^{n+1}\right)^2.$$

In order for the energy method to work, the boundary condition has to satisfy

$$2a\frac{B_u}{B_v}\left(u_0^{n+1}\right)^2 + \frac{\Delta x}{\varepsilon}\left(v_0^{n+1}\right)^2 \geq C\left|U_0^{n+1}\right|^2,$$

for some constant $C > 0$ whenever $B_u u_0^n + B_v v_0^n = 0$, for any $n \in \mathbb{N}$. This leads to the following sufficient condition

$$2a\frac{B_u}{B_v} + \frac{\Delta x}{\varepsilon}\left(\frac{B_u}{B_v}\right)^2 > 0,$$

under which we directly get the inequality

$$\frac{1}{\Delta t}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x} - \left\langle U^n, HU^n\right\rangle_{\Delta x}\Big) + C\left|U_0^{n+1}\right|^2 \leq 0$$

and the initial data is compatible at the space-time corner $(x_j, t^n) = (0, 0)$, i.e, $U_0^0 = 0$. Thus,

$$\left\langle U^n, HU^n\right\rangle_{\Delta x} + C\Delta t\sum_{k=0}^{n}\left|U_0^k\right|^2 \leq \left\langle f, Hf\right\rangle_{\Delta x}, \quad \text{for any } n > 0. \tag{2.3.11}$$

More into the details, the following cases occur:

- If $B_uB_v > 0$ then there exists $C \leq 2aB_uB_v \left(B_u^2 + B_v^2\right)^{-1}$ such that the inequality (2.3.11) holds uniformly.
- If $B_uB_v < 0$, consider some $\delta_0 > -2aB_vB_u^{-1}$. Then there exists $C(\delta_0) > 0$ such that the inequality (2.3.11) holds uniformly as soon as $\Delta x \geq \delta_0\varepsilon$ with $C = C(\delta_0)$. For example, if we choose $\delta_0 = -3aB_vB_u^{-1}$ then there exists $C \leq -aB_uB_v \left(B_u^2 + B_v^2\right)^{-1}$ such that (2.3.11) holds uniformly.

This ends the proof of Proposition 2.3.2.

Let us mention that, assuming the condition (2.2.8) of the main theorem to be fulfilled, the discrete strict dissipativity condition (2.2.10) is then automatically satisfied. Then, from the inequality (2.3.11), for any $T > 0$, there exists a constant $C_T > 0$ such that the following inequality holds

$$\sum_{n=0}^{N}\sum_{j\geq 0} \Delta x \Delta t \left|U_j^n\right|^2 + \sum_{n=0}^{N} \Delta t \left|U_0^n\right|^2 \leq C_T \sum_{j\geq 0} \Delta x \left|f_j\right|^2, \tag{2.3.12}$$

where $N := T/\Delta t$.

2.3.2 Stiff strong stability of the fully discrete IBVP

In the continuous case, the IBVP (2.1.1)-(2.1.3) is stiffly well-posed if and only if the boundary condition satisfies the SKC (2.1.6). In Section 2.2.2.1, we show that $B_uB_v > 0$ seems to be necessary to ensure the stiff stability of the semi-discrete IBVP (2.2.7). Now, we want to address the question of existence of unstable solutions in order to derive a necessary condition for the stability of the discrete IBVP (2.3.1). Following W.-A. Yong in [97] and Z. Xin and W. Xu in [96], we shall apply the normal mode analysis to derive the strict dissipativity condition (2.2.8).

To do that, we look for (nontrivial) solution of (2.3.1) satisfying the homogeneous boundary condition $BU_0^n = 0$, for any $n \in \mathbb{N}$ and of the form

$$U_j^n = \left(1 + \frac{\xi \Delta t}{\varepsilon}\right)^n \phi_j, \quad n \in \mathbb{N}, \tag{2.3.13}$$

with $\xi \in \mathbb{C}$ such that $\mathrm{Re}\,(\xi) > 0$, and $(\phi_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{C}^2)$. Since $\mathrm{Re}\,(\xi) > 0$, we can see that $|1 + \xi\Delta t\varepsilon^{-1}| > 1$ for any $\Delta t > 0$ and $\varepsilon > 0$. Such solutions, if they exist, clearly violate the $\varepsilon-$uniform ℓ^2 estimate in (2.3.1). Our goal is to find a sufficient condition to ensure that they do not exist.

Substituting (2.3.13) into (2.3.1), we have to solve the following problem

$$\phi_{j+1} - \phi_{j-1} = 2\Delta x \varepsilon^{-1} M(\xi_1)\phi_j, \quad j > 0 \tag{2.3.14a}$$

$$B\phi_0 = 0, \tag{2.3.14b}$$

$$\Pi_2 P^{-T} HA \left(\phi_1 - \left(I + \Delta x\varepsilon^{-1} M(\xi_1)\right)\phi_0\right) = 0, \tag{2.3.14c}$$

where

$$\xi_1 = \xi \left(1 + \frac{\xi\Delta t}{\varepsilon}\right)^{-1} \tag{2.3.15}$$

and the matrix $M(\xi_1)$ is the same as in (2.2.22).

Let $\xi = \alpha + i\beta$, $\alpha \geq 0$, $\beta \in \mathbb{R}$ and $\Delta\bar{t} = \Delta t\varepsilon^{-1}$. Then,

$$\mathrm{Re}\,(\xi_1) = \overline{\alpha} = \frac{\alpha\left(1 + \Delta\bar{t}\alpha\right) + \Delta\bar{t}\beta^2}{\left(1 + \Delta\bar{t}\alpha\right)^2 + \Delta\bar{t}^2\beta^2} > 0.$$

According to Lemma 2.5.1 with the property $\operatorname{Re}(\xi_1) > 0$, we can prove

$$\operatorname{Re}\left(\mu_-(\xi_1)\right) < 0 \quad \text{and} \quad \operatorname{Re}\left(\mu_+(\xi_1)\right) > 0$$

with $\mu_\pm(\xi_1)$ is defined in (2.2.23).

One can proceed as in (2.2.25)-(2.2.29) to find the solution $(\phi_j)_{j>0}$ of (2.3.14a). For some vector $R \in \mathbb{C}^2$, it takes the form

$$\phi_j = \overline{P}(\xi_1) Z^j(\xi_1) R. \tag{2.3.16}$$

Plugging now (2.3.16) into the boundary condition (2.3.14b) and (2.3.14c), we get the following linear system

$$N(\xi_1)R = 0,$$

where $N(\xi_1)$ is the same as in (2.2.32).

Following Proposition 2.2.6, under the sufficient condition $B_u B_v > 0$, one has $\det N(\xi_1) \neq 0$. It means that the fully discrete (2.3.1) with homogeneous boundary condition does not admit unstable solution of the form (2.3.13) in $\ell^2(\mathbb{N}, \mathbb{C}^2)$ as soon as $B_u B_v > 0$.

2.3.3 Stiff stability of the fully discrete IBVP with homogeneous initial condition

For convenience in the forthcoming discussions, we recall that the fully discrete approximation of the IBVP (2.3.1) with homogeneous initial condition reads

$$\begin{cases} U_j^{n+1} - U_j^n + \dfrac{\Delta t}{2\Delta x} A\left(U_{j+1}^{n+1} - U_{j-1}^{n+1}\right) = \dfrac{\Delta t}{\varepsilon} S U_j^{n+1}, & j \geq 1,\ n \geq 0, \\ U_j^0 = 0, & j \geq 0, \\ B U_0^n = b^n, & n \geq 0, \\ \Pi_2 H_P P\left(U_0^{n+1} - U_0^n\right) + \dfrac{\Delta t}{\Delta x} \Pi_2 H_P P A\left(U_1^{n+1} - U_0^{n+1}\right) = \dfrac{\Delta t}{\varepsilon} \Pi_2 H_P P S U_0^{n+1}, & n \geq 0. \end{cases} \tag{2.3.17}$$

Under the strict dissipativity condition (2.2.8), the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ can be constructed by the method of Laplace transform. By using the Plancherel's theorem, we get the expected results of Proposition 2.3.3.

2.3.3.1 Solution by Laplace transform

To find the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ to the fully discrete IBVP (2.3.17), we need the grid vector function $(U_j^n)_{j\in\mathbb{N}}$ and the data to be defined for all t. Therefore, we defined piecewise constant functions from the discrete values

$$U_j(t) = U_j^n, \quad \text{for } t^n \leq t < t^{n+1}$$

and

$$b(t) = b^n, \quad \text{for } t^n \leq t < t^{n+1}.$$

We recall the definition of Laplace transform of $U_j(t)$ defined on $\mathbb{R}^+$

$$\widetilde{U}_j(\xi) = \mathcal{L}U_j = \int_0^{+\infty} e^{-\xi t} U_j(t)dt, \quad \operatorname{Re}(\xi) > 0.$$

It is easy to see that the Laplace transform of $U_j(t)$ is well-defined ,

$$\left|\widetilde{U}_j(\xi)\right| \leq \int_0^{+\infty} \left|e^{-\xi t}U_j(t)\right| dt \leq \sum_{n\geq 0} \left|U_j^n\right| \int_{t^n}^{t^{n+1}} \left|e^{-\xi t}\right| dt \leq +\infty, \quad \mathrm{Re}\,(\xi) > 0.$$

With $U_j^0 \equiv 0$, we have

$$\int_0^{+\infty} e^{-\xi t}U_j(t+\Delta t)dt = \int_{\Delta t}^{+\infty} e^{-\xi(t-\Delta t)}U_j(t)dt = e^{\xi\Delta t}\int_0^{+\infty} e^{-\xi t}U_j(t)dt$$

and therefore the system (2.3.17) becomes

$$\widetilde{U}_{j+1}(\xi) - \widetilde{U}_{j-1}(\xi) = 2\Delta x\varepsilon^{-1}M(\varepsilon\widetilde{\xi})\widetilde{U}_j(\xi), \quad j > 0, \tag{2.3.18a}$$

$$B\widetilde{U}_0(\xi) = \widetilde{b}(\xi), \tag{2.3.18b}$$

$$\Pi_2 H_P PA\Big(\widetilde{U}_1(\xi) - \Big(I + \Delta x\varepsilon^{-1}M(\varepsilon\widetilde{\xi})\Big)\,\widetilde{U}_0(\xi)\Big) = 0, \tag{2.3.18c}$$

where

$$\begin{aligned}\widetilde{\xi} &= \left(1 - e^{-\xi\Delta t}\right)\Delta t^{-1},\\ \widetilde{b}(\xi) &= \mathcal{L}b = \int_0^{+\infty} e^{-\xi t}b(t)dt\end{aligned} \tag{2.3.19}$$

and the matrix $M(\varepsilon\widetilde{\xi})$ is the same as in (2.2.22).

Let $\xi = \alpha + i\beta$, $\alpha > 0$, $\beta \in \mathbb{R}$ and $\Delta t > 0$. Then,

$$\mathrm{Re}\,(\widetilde{\xi}) = \widetilde{\alpha} = \Big(1 - e^{-\alpha\Delta t}\cos(-\beta\Delta t)\Big)\Delta t^{-1} > 0, \tag{2.3.20}$$

and thus,

$$\left(1 - e^{-\alpha\Delta t}\right)\Delta t^{-1} \leq \widetilde{\alpha} \leq 2\Delta t^{-1}. \tag{2.3.21}$$

According to Lemma 2.5.1 with the property $\mathrm{Re}\,(\widetilde{\xi}) > 0$, $\varepsilon > 0$, we can prove

$$\mathrm{Re}\,\Big(\mu_-(\varepsilon\widetilde{\xi})\Big) < 0 \quad \text{and} \quad \mathrm{Re}\,\Big(\mu_+(\varepsilon\widetilde{\xi})\Big) > 0$$

with $\mu_\pm(\varepsilon\widetilde{\xi})$ is defined in (2.2.23).

One can proceed as in (2.2.25)-(2.2.29) to find the solution $\widetilde{U}_j(\xi)$ of (2.3.18a) in $\ell^2(\mathbb{N}, \mathbb{C}^2)$. For some vector $R \in \mathbb{C}^2$, it takes the form

$$\widetilde{U}_j(\xi) = \overline{P}(\varepsilon\widetilde{\xi})Z^j(\varepsilon\widetilde{\xi})R.$$

The value of vector R can be determined easily from the boundary condition (2.3.18b) and (2.3.18c)

$$R = \frac{\widetilde{b}(\xi)}{\mathrm{det}N(\varepsilon\widetilde{\xi})}N_1(\varepsilon\widetilde{\xi}),$$

where the matrix $N(\varepsilon\widetilde{\xi})$ and $N_1(\varepsilon\widetilde{\xi})$ are the same as in (2.2.32) and (2.2.43), respectively.

Therefore,

$$\widetilde{U}_j(\xi) = \frac{\widetilde{b}(\xi)}{\mathrm{det}N(\varepsilon\widetilde{\xi})}\overline{P}(\varepsilon\widetilde{\xi})Z^j(\varepsilon\widetilde{\xi})N_1(\varepsilon\widetilde{\xi}).$$

With $\widetilde{U}_j(\xi)$ found, the numerical solution of (2.3.17) can be obtained by inverting the Laplace transform

$$U_j(t) = \mathcal{L}^{-1}\widetilde{U}_j(\xi) = \frac{1}{2\pi}\oint_{-\infty}^{+\infty} e^{(\alpha+i\beta)t}\widetilde{U}_j(\alpha+i\beta)d\beta, \quad \alpha > 0.$$

2.3.3.2 Stiff stability analysis

Under the strict dissipativity condition $B_u B_v > 0$, we consider Proposition 2.3.3 with homogeneous initial condition $(f_j)_{j\in\mathbb{N}} \equiv 0$ and nonzero boundary data $b^n \in \ell^2(\mathbb{N},\mathbb{R})$. By an application of the following Plancherel's theorem for Laplace transform

$$\int_0^{+\infty} e^{-2\alpha t} \left|U_j(t)\right|^2 dt = \frac{1}{2\pi} \oint_{-\infty}^{+\infty} \left|\widetilde{U}_j(\alpha + i\beta)\right|^2 d\beta, \quad \alpha > 0$$

and

$$e^{-2\alpha t^n} = \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \int_{t^n}^{t^{n+1}} e^{-2\alpha t} dt,$$

we have

$$\begin{aligned}
\sum_{n\geq 0} e^{-2\alpha t^n} \left|U_0^n\right|^2 &= \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \sum_{n\geq 0} \int_{t^n}^{t^{n+1}} e^{-2\alpha t} dt \left|U_0^n\right|^2 = \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \sum_{n\geq 0} \int_{t^n}^{t^{n+1}} e^{-2\alpha t} \left|U_0(t)\right|^2 dt \\
&= \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \int_0^{+\infty} e^{-2\alpha t} \left|U_0(t)\right|^2 dt = \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \times \frac{1}{2\pi} \oint_{-\infty}^{+\infty} \left|\widetilde{U}_0(\alpha + i\beta)\right|^2 d\beta \\
&= \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \times \frac{1}{2\pi} \oint_{-\infty}^{+\infty} \left|\widetilde{b}(\xi)\right|^2 \times \frac{\left\|N_1(\varepsilon\widetilde{\xi})\right\|^2}{\left|\mathrm{det} N(\varepsilon\widetilde{\xi})\right|^2} \times \left|\overline{P}(\varepsilon\widetilde{\xi})\right|^2 d\beta,
\end{aligned}$$

where $\xi = \alpha + i\beta$ and $\widetilde{\xi}$ is the same as in (2.3.19). We fix $\alpha > 0$ from now on.

According to Proposition 2.2.9, there exists $C_1 > 0$ such that for any $\delta = \Delta x \varepsilon^{-1} > 0$, $\widetilde{\xi} \in \mathbb{C}$, $\mathrm{Re}\,(\widetilde{\xi}) \geq 0$,

$$\frac{\|N_1(\varepsilon\widetilde{\xi})\|^2}{|\det N(\varepsilon\widetilde{\xi})|^2} \leq C_1.$$

On the other hand, since $k(\varepsilon\widetilde{\xi}) = -g(\varepsilon\widetilde{\xi})$ is uniformly bounded in $\mathrm{Re}\,(\widetilde{\xi}) \geq 0$, $\varepsilon > 0$, we obtain

$$\begin{aligned}
\sum_{n\geq 0} e^{-2\alpha t^n} \left|U_0^n\right|^2 &\lesssim \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \times \frac{1}{2\pi} \oint_{-\infty}^{+\infty} \left|\widetilde{b}(\xi)\right|^2 d\beta \\
&\lesssim \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \int_0^{+\infty} e^{-2\alpha t} \left|b(t)\right|^2 dt \\
&\lesssim \sum_{n\geq 0} e^{-2\alpha t^n} \left|b^n\right|^2 .
\end{aligned} \tag{2.3.22}$$

Similarly, by an application of the Plancherel's theorem for Laplace transform, we have

$$\begin{aligned}
\sum_{n\geq 0} \sum_{j\geq 0} e^{-2\alpha t^n} \left|U_j^n\right|^2 &= \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \sum_{j\geq 0} \int_0^{+\infty} e^{-2\alpha t} \left|U_j(t)\right|^2 dt \\
&= \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \times \frac{1}{2\pi} \sum_{j\geq 0} \oint_{-\infty}^{+\infty} \left|\widetilde{U}_j(\xi)\right|^2 d\beta.
\end{aligned}$$

Since $k(\varepsilon\widetilde{\xi}) = -g(\varepsilon\widetilde{\xi})$ is uniformly bounded in $\mathrm{Re}\,(\widetilde{\xi}) \geq 0$, $\varepsilon > 0$ and using Proposition 2.2.9, we obtain

$$\sum_{n\geq 0} \sum_{j\geq 0} e^{-2\alpha t^n} \left|U_j^n\right|^2 \lesssim \frac{2\alpha}{1 - e^{-2\alpha\Delta t}} \times \frac{1}{2\pi} \sum_{j\geq 0} \oint_{-\infty}^{+\infty} \left|\widetilde{b}(\xi)\right|^2 \left(\left|z_-(\varepsilon\widetilde{\xi})\right|^{2j} + \left|w_+(\varepsilon\widetilde{\xi})\right|^{2j} \right) d\beta,$$

where $z_-(\varepsilon\widetilde{\xi})$ and $w_+(\varepsilon\widetilde{\xi})$ are defined in (2.2.27) and (2.2.28), respectively.

On the other hand, since $\mu_-(\varepsilon\widetilde{\xi}) = -\mu_+(\varepsilon\widetilde{\xi})$ in $\mathrm{Re}\,(\widetilde{\xi})$, $\varepsilon > 0$, we get

$$\left|z_-(\varepsilon\widetilde{\xi})\right| = \left|w_+(\varepsilon\widetilde{\xi})\right|$$

Thus,

$$\sum_{n\geq 0}\sum_{j\geq 0} e^{-2\alpha t^n}\left|U_j^n\right|^2 \lesssim \frac{2\alpha}{1-e^{-2\alpha\Delta t}} \times \frac{1}{2\pi}\oint_{-\infty}^{+\infty}\sum_{j\geq 0}\left|w_+(\varepsilon\widetilde{\xi})\right|^{2j}\left|\widetilde{b}(\xi)\right|^2 d\beta. \tag{2.3.23}$$

According to (2.2.24), for all $\varepsilon > 0$, $\mathrm{Re}\,(\widetilde{\xi}) > 0$, we have

$$\mathrm{Re}\,\left(\mu_-(\varepsilon\widetilde{\xi})\right) \leq -\frac{\varepsilon \mathrm{Re}\,(\widetilde{\xi})}{\sqrt{a}} < 0.$$

Furthermore, we can prove

$$\left(\frac{\mathrm{Re}\,\left(\mu_-(\varepsilon\widetilde{\xi})\right)\Delta x}{\varepsilon} + \sqrt{\left(\frac{\mathrm{Re}\,\left(\mu_-(\varepsilon\widetilde{\xi})\right)\Delta x}{\varepsilon}\right)^2 + 1}\right)^2 \leq \left(\eta\Delta x + \sqrt{\eta^2\Delta x^2 + 1}\right)^2, \tag{2.3.24}$$

where $\eta = -\dfrac{\mathrm{Re}\,(\widetilde{\xi})}{\sqrt{a}}$. According to Lemma 2.5.2 and (2.3.24), we have now

$$\begin{aligned}\sum_{j\geq 0}|w_+(\varepsilon\widetilde{\xi})|^{2j} &= \left(1 - \left|\frac{\mu_-(\varepsilon\widetilde{\xi})\Delta x}{\varepsilon} + \sqrt{\left(\frac{\mu_-(\varepsilon\widetilde{\xi})\Delta x}{\varepsilon}\right)^2 + 1}\right|^2\right)^{-1} \\ &\leq \left(1 - \left(\eta\Delta x + \sqrt{\eta^2\Delta x^2 + 1}\right)^2\right)^{-1}.\end{aligned}$$

Since $\mathrm{Re}\,(\widetilde{\xi})$ satisfies (2.3.21), we get

$$\frac{\Delta t\sqrt{a}}{2} \leq -\eta^{-1} \leq \frac{\Delta t\sqrt{a}}{1-e^{-\alpha\Delta t}}.$$

If we assume $\Delta x \leq \dfrac{3\sqrt{a}}{8}\Delta t \leq -\dfrac{3}{4\eta}$, then

$$\left(1 - \left(\eta\Delta x + \sqrt{\eta^2\Delta x^2 + 1}\right)^2\right)^{-1} \leq -\eta^{-1}\Delta x^{-1}.$$

Thus,

$$\sum_{j\geq 0}|w_+(\varepsilon\widetilde{\xi})|^{2j} \leq -\eta^{-1}\Delta x^{-1} \leq \frac{\Delta t\sqrt{a}}{\Delta x\left(1-e^{-\alpha\Delta t}\right)}. \tag{2.3.25}$$

Substituting (2.3.25) into (2.3.23), we have

$$\begin{aligned}\frac{e^{\alpha\Delta t}-1}{e^{\alpha\Delta t}}\Delta x\sum_{n\geq 0}\sum_{j\geq 0} e^{-2\alpha t^n}\left|U_j^n\right|^2 &\lesssim \frac{2\alpha}{1-e^{-2\alpha\Delta t}} \times \frac{\Delta t}{2\pi}\oint_{-\infty}^{+\infty}\left|\widetilde{b}(\xi)\right|^2 d\beta \\ &\lesssim \Delta t\sum_{n\geq 0} e^{-2\alpha t^n}\left|b^n\right|^2.\end{aligned} \tag{2.3.26}$$

According to (2.3.22) and (2.3.26), there exists a constant $C > 0$ such that

$$\frac{e^{\alpha\Delta t}-1}{e^{\alpha\Delta t}}\Delta x \sum_{n\geq 0}\sum_{j\geq 0} e^{-2\alpha t^n}\left|U_j^n\right|^2 + \sum_{n\geq 0} e^{-2\alpha t^n}\Delta t \left|U_0^n\right|^2 \leq C\Delta t \sum_{n\geq 0} e^{-2\alpha t^n}\left|b^n\right|^2.$$

By using the power series expansion of exponential function

$$e^{\alpha\Delta t} \geq 1+\alpha\Delta t, \quad \text{for } \alpha > 0, \Delta t > 0,$$

there exists a constant $C > 0$ such that

$$\frac{\alpha}{1+\alpha\Delta t}\sum_{n\geq 0}\sum_{j\geq 0} e^{-2\alpha n\Delta t}\Delta t\Delta x\left|U_j^n\right|^2 + \sum_{n\geq 0} e^{-2\alpha n\Delta t}\Delta t\left|U_0^n\right|^2 \leq C\sum_{n\geq 0} e^{-2\alpha n\Delta t}\Delta t\left|b^n\right|^2.$$

This ends the proof of the Proposition 2.3.3.

Together with a consequence of the hyperbolicity of (2.1.1) by using the classical argument of changing the data b to zero after time T and unchanged before time T, there exists a constant $C_T > 0$ such that

$$\sum_{n=0}^{N}\sum_{j\geq 0}\Delta x\Delta t\left|U_j^n\right|^2 + \sum_{n=0}^{N}\Delta t\left|U_0^n\right|^2 \leq C_T\sum_{n=0}^{N}\Delta t\left|b^n\right|^2 \tag{2.3.27}$$

with $N := T/\Delta t$.

By linearity, we can break up the fully discrete IBVP (2.3.1) into two simpler problems, one with homogeneous initial condition and the other with homogeneous boundary condition. Finally, from (2.3.12) and (2.3.27), we get the expected result of Theorem 2.3.1.

2.4 The upwind scheme

Following Section 2.3, we focus in this section on the fully discrete approximation of the IBVP (2.1.1)-(2.1.3) with homogeneous boundary condition $b^n \equiv 0$, for any $n \in \mathbb{N}$, obtained by the upwind scheme in space and the implicit scheme in time

$$\begin{cases} U_j^{n+1} - U_j^n + \Delta t(\mathcal{Q}U)_j^{n+1} = \Delta t\varepsilon^{-1}SU_j^{n+1}, & j\geq 1, n\geq 0,\\ U_j^0 = f_j, & j\geq 0,\\ BU_0^n = 0, & n\geq 0,\\ \Pi_2 H_P P\left(U_0^{n+1}-U_0^n\right) + \Delta t\Pi_2 H_P P(\mathcal{Q}U)_0^{n+1} = \Delta t\varepsilon^{-1}\Pi_2 H_P PSU_0^{n+1}, & n\geq 0. \end{cases} \tag{2.4.1}$$

where A, S and B are the same as in (2.2.1). In the case of the upwind scheme, the considered operator reads

$$(\mathcal{Q}U)_j = \frac{1}{2\Delta x}\Big((A-\sqrt{a}I)U_{j+1} + 2\sqrt{a}U_j - (A+\sqrt{a}I)U_{j-1}\Big), \quad j\in\mathbb{N}.$$

At the boundary point $j = 0$, we use the upwind scheme at the boundary point $j = 0$ but supply another boundary condition that determines the value $(A+\sqrt{a}I)U_{-1}^n$ through the identity

$$\left(A-\sqrt{a}I\right)U_1^n - 2AU_0^n + \left(A+\sqrt{a}I\right)U_{-1}^n = 0.$$

Therefore, we propose the following numerical approximation at boundary

$$\frac{1}{\Delta t}\Pi_2 H_P P\left(U_0^{n+1}-U_0^n\right) + \frac{1}{\Delta x}\Pi_2 H_P P\left(A-\sqrt{a}I\right)\left(U_1^{n+1}-U_0^{n+1}\right) = \frac{1}{\varepsilon}\Pi_2 H_P PSU_0^{n+1}, \quad n\geq 0.$$

To summarize, along the rest of the section, we will study the following fully discrete approximation of the IBVP (2.1.1)-(2.1.3):

$$\begin{cases} \dfrac{U_j^{n+1}-U_j^n}{\Delta t} + \dfrac{1}{2\Delta x}\Big(\left(A-\sqrt{a}I\right)U_{j+1}^{n+1} + 2\sqrt{a}U_j^{n+1} - \left(A+\sqrt{a}I\right)U_{j-1}^{n+1}\Big) = \dfrac{1}{\varepsilon}SU_j^{n+1}, & j \geq 1, n \geq 0, \\ U_j^0 = f_j, & j \geq 0, \\ BU_0^n = 0, & n \geq 0, \\ \dfrac{1}{\Delta t}\Pi_2 H_P P\left(U_0^{n+1}-U_0^n\right) + \dfrac{1}{\Delta x}\Pi_2 H_P P\left(A-\sqrt{a}I\right)\left(U_1^{n+1}-U_0^{n+1}\right) = \dfrac{1}{\varepsilon}\Pi_2 H_P P S U_0^{n+1}, & n \geq 0, \end{cases} \tag{2.4.2}$$

where H_P and P are defined in (2.2.5), $\Pi_2 = \begin{pmatrix} 0 & 1 \end{pmatrix}$.

In Section 2.3, we show that the discrete strict dissipativity condition (2.2.10) is sufficient to derive uniform stability estimates of the numerical scheme (2.3.1) with $b^n \equiv 0$, for any $n \in \mathbb{N}$. Our aim is to determine a sufficient condition for the stiff stability of the above fully discrete (2.4.2), in order words the uniform stability with respect to the stiffness of the relaxation term. We state successively hereafter the following statement

Proposition 2.4.1. *Assume that the parameters $\Delta x \in \ (0,1]$, $\varepsilon > 0$ and (B_u, B_v) satisfies the discrete strict dissipativity condition (2.2.10). Then, there exists a constant $C > 0$ such that for all $\Delta t > 0$ and any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, the solution $(U_j^n)_{j\in\mathbb{N}}$ to scheme (2.4.2) satisfies*

$$\langle U^n, HU^n \rangle_{\Delta x} + C\Delta t \sum_{k=0}^{n} \left|U_0^k\right|^2 \leq \langle f, Hf \rangle_{\Delta x}, \quad n \in \mathbb{N}. \tag{2.4.3}$$

More precisely,

a) If $B_u B_v > 0$ then (2.4.3) holds uniformly, i.e. with C independent of ε and Δx.

b) If $B_u B_v < 0$ then considering some $\delta_0 > -2aB_v B_u^{-1}$, there exists $C(\delta_0) > 0$ such that (2.4.3) holds uniformly with $C = C(\delta_0)$, as soon as $\Delta x \geq \delta_0 \varepsilon$.

The Proposition 2.4.1 is studied in Section 2.4.1 by means of the discrete energy method. In order to illustrate the relevance of the discrete strict dissipativity condition (2.2.10), we present in Section 2.4.2 some numerical results, for various values of the parameters (B_u, B_v) and show that the energy $\left\langle U^n, P^T H_P P U^n \right\rangle_{\Delta x}$ increases if the condition (2.2.10) does not hold. Let us remark that we have difficulty finding a sufficient condition for the stiff stability of fully discrete (2.4.2) with nonzero boundary condition, but we postpone its possibility to a further work.

2.4.1 The energy method

One can proceed as in (2.3.5)-(2.3.6) to obtain the following equation

$$\begin{aligned} \frac{1}{\Delta t}\left\langle U^{n+1}-U^n, P^T H_P P U^{n+1} \right\rangle_{\Delta x} &= \frac{\delta_{xt}}{2}\Big(\Pi_2 H_P P\left(U_0^{n+1}-U_0^n\right)\Big)\left(\Pi_2 P U_0^{n+1}\right) \\ &\quad + \delta_{xt}\sum_{j=1}^{+\infty}\left\langle U_j^{n+1}-U_i^n, P^T H_P P U_j^{n+1}\right\rangle, \end{aligned} \tag{2.4.4}$$

where $\delta_{xt} = \Delta x \Delta t^{-1}$.

On the other hand, from the first and fourth equations in (2.4.2), we have

$$\frac{1}{\Delta t}\Pi_2 H_P P\left(U_0^{n+1}-U_0^n\right) = \frac{1}{\varepsilon}\Pi_2 H_P P S U_0^{n+1} - \frac{1}{\Delta x}\Pi_2 H_P P\left(A-\sqrt{a}I\right)\left(U_1^{n+1}-U_0^{n+1}\right)$$

and

$$\frac{1}{\Delta t}\left(U_j^{n+1}-U_j^n\right)=\frac{1}{\varepsilon}SU_j^{n+1}-\frac{1}{2\Delta x}\Big(\left(A-\sqrt{a}I\right)U_{j+1}^{n+1}+2\sqrt{a}U_j^{n+1}-\left(A+\sqrt{a}I\right)U_{j-1}^{n+1}\Big).$$

As a consequence, the equation (2.4.4) can be represented as

$$\begin{aligned}
&\frac{1}{\Delta t}\left\langle U^{n+1}-U^n,P^TH_PPU^{n+1}\right\rangle_{\Delta x}\\
=&\frac{\Delta x}{2\varepsilon}\left(\Pi_2H_PPSU_0^{n+1}\right)\left(\Pi_2PU_0^{n+1}\right)+\frac{\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left\langle SU_j^{n+1},P^TH_PPU_j^{n+1}\right\rangle\\
-&\frac{1}{2}\Big(\Pi_2H_PPA\left(U_1^{n+1}-U_0^{n+1}\right)\Big)\left(\Pi_2PU_0^{n+1}\right)-\frac{1}{2}\sum_{j=1}^{+\infty}\Big\langle A\left(U_{j+1}^{n+1}-U_{j-1}^{n+1}\right),P^TH_PPU_j^{n+1}\Big\rangle\\
+&\frac{\sqrt{a}}{2}\Big(\Pi_2H_PP\left(U_1^{n+1}-U_0^{n+1}\right)\Big)\left(\Pi_2PU_0^{n+1}\right)+\frac{\sqrt{a}}{2}\sum_{j=1}^{+\infty}\left\langle U_{j+1}^{n+1}-2U_j^{n+1}+U_{j-1}^{n+1},P^TH_PPU_j^{n+1}\right\rangle.
\end{aligned}$$

We observe now that

$$-\sum_{j=1}^{+\infty}\Big\langle A\left(U_{j+1}^{n+1}-U_{j-1}^{n+1}\right),P^TH_PPU_j^{n+1}\Big\rangle=\left(\Pi_2H_PPAU_1^{n+1}\right)\left(\Pi_2PU_0^{n+1}\right),$$

$$\sum_{j=1}^{+\infty}\left\langle U_{j+1}^{n+1},P^TH_PPU_j^{n+1}\right\rangle=-\left(\Pi_2H_PPU_1^{n+1}\right)\left(\Pi_2PU_0^{n+1}\right)+\sum_{j=1}^{+\infty}\left\langle U_{j-1}^{n+1},P^TH_PPU_j^{n+1}\right\rangle$$

and

$$\begin{aligned}
\sum_{j=1}^{+\infty}\left\langle U_j^{n+1},P^TH_PPU_j^{n+1}\right\rangle=&\frac{1}{2}\sum_{j=1}^{+\infty}\left\langle U_j^{n+1},P^TH_PPU_j^{n+1}\right\rangle+\frac{1}{2}\sum_{j=1}^{+\infty}\left\langle U_{j-1}^{n+1},P^TH_PPU_{j-1}^{n+1}\right\rangle\\
&-\frac{1}{2}\left(\Pi_2H_PPU_0^{n+1}\right)\left(\Pi_2PU_0^{n+1}\right).
\end{aligned}$$

Then,

$$\begin{aligned}
&\frac{1}{\Delta t}\left\langle U^{n+1}-U^n,P^TH_PPU^{n+1}\right\rangle_{\Delta x}\\
=&\frac{\Delta x}{2\varepsilon}\left(\Pi_2H_PPSU_0^{n+1}\right)\left(\Pi_2PU_0^{n+1}\right)+\frac{\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left\langle SU_j^{n+1},P^TH_PPU_j^{n+1}\right\rangle\\
+&\frac{1}{2}\left(\Pi_2H_PPAU_0^{n+1}\right)\left(\Pi_2PU_0^{n+1}\right)-\frac{\sqrt{a}}{2}\sum_{j=1}^{+\infty}\left\langle U_j^{n+1},P^TH_PPU_j^{n+1}\right\rangle\\
+&\sqrt{a}\sum_{j=1}^{+\infty}\left\langle U_{j-1}^{n+1},P^TH_PPU_j^{n+1}\right\rangle-\frac{\sqrt{a}}{2}\sum_{j=1}^{+\infty}\left\langle U_{j-1}^{n+1},P^TH_PPU_{j-1}^{n+1}\right\rangle.
\end{aligned}\tag{2.4.5}$$

Besides, the three last terms in (2.4.5) becomes

$$\begin{aligned}
&-\frac{\sqrt{a}}{2}\sum_{j=1}^{+\infty}\Big(\left\langle U_j^{n+1},P^TH_PPU_j^{n+1}\right\rangle-2\left\langle U_{j-1}^{n+1},P^TH_PPU_j^{n+1}\right\rangle+\left\langle U_{j-1}^{n+1},P^TH_PPU_{j-1}^{n+1}\right\rangle\Big)\\
=&-\frac{B_v^2\sqrt{a}}{2}\sum_{j=1}^{+\infty}\Big(a\left(u_j^{n+1}-u_{j-1}^{n+1}\right)^2+\left(v_j^{n+1}-v_{j-1}^{n+1}\right)^2\Big)\leq 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{1}{\Delta t}\left\langle U^{n+1}-U^{n}, P^{T}H_{P}PU^{n+1}\right\rangle_{\Delta x}\\
&\leq \frac{\Delta x}{2\varepsilon}\left(\Pi_2 H_P PSU_0^{n+1}\right)\left(\Pi_2 PU_0^{n+1}\right)+\frac{\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left\langle SU_j^{n+1}, P^T H_P PU_j^{n+1}\right\rangle\\
&+\frac{1}{2}\left(\Pi_2 H_P PAU_0^{n+1}\right)\left(\Pi_2 PU_0^{n+1}\right),
\end{aligned}$$

or

$$\frac{1}{\Delta t}\left\langle U^{n+1}-U^{n}, HU^{n+1}\right\rangle_{\Delta x}+a\frac{B_u}{B_v}\left(u_0^{n+1}\right)^2+\frac{\Delta x}{2\varepsilon}\left(v_0^{n+1}\right)^2\leq-\frac{\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left(v_j^{n+1}\right)^2, \tag{2.4.6}$$

where the symmetric positive definite H is the same as in (2.2.17).
Since H is a symmetric positive definite matrix, we can see that

$$\begin{aligned}
\left\langle U^{n+1}-U^{n}, HU^{n+1}\right\rangle_{\Delta x} &= \frac{1}{2}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x}-\left\langle U^{n}, HU^{n}\right\rangle_{\Delta x}\\
&\quad+\left\langle U^{n+1}-U^{n}, H\left(U^{n+1}-U^{n}\right)\right\rangle_{\Delta x}\Big)\\
&\geq \frac{1}{2}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x}-\left\langle U^{n}, HU^{n}\right\rangle_{\Delta x}\Big).
\end{aligned} \tag{2.4.7}$$

According to (2.4.6) and (2.4.7), one gets

$$\frac{1}{\Delta t}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x}-\left\langle U^{n}, HU^{n}\right\rangle_{\Delta x}\Big)+2a\frac{B_u}{B_v}\left(u_0^{n+1}\right)^2+\frac{\Delta x}{\varepsilon}\left(v_0^{n+1}\right)^2\leq-\frac{2\Delta x}{\varepsilon}\sum_{j=1}^{+\infty}\left(v_j^{n+1}\right)^2.$$

In order for the energy method to work, the boundary condition has to satisfy

$$2a\frac{B_u}{B_v}\left(u_0^{n+1}\right)^2+\frac{\Delta x}{\varepsilon}\left(v_0^{n+1}\right)^2\geq C\left|U_0^{n+1}\right|^2,$$

for some constant $C>0$ whenever $B_u u_0^n+B_v v_0^n=0$, for any $n\in\mathbb{N}$. This leads to the following sufficient condition

$$2a\frac{B_u}{B_v}+\frac{\Delta x}{\varepsilon}\left(\frac{B_u}{B_v}\right)^2>0,$$

under which we directly get the inequality

$$\frac{1}{\Delta t}\Big(\left\langle U^{n+1}, HU^{n+1}\right\rangle_{\Delta x}-\left\langle U^{n}, HU^{n}\right\rangle_{\Delta x}\Big)+C\left|U_0^{n+1}\right|^2\leq 0$$

and the initial data is compatible at the space-time corner $(x_j,t^n)=(0,0)$, i.e, $U_0^0=0$. Thus,

$$\left\langle U^{n}, HU^{n}\right\rangle_{\Delta x}+C\Delta t\sum_{k=0}^{n}\left|U_0^k\right|^2\leq\left\langle f, Hf\right\rangle_{\Delta x}, \quad \text{for any } n>0. \tag{2.4.8}$$

More into the details, the following cases occur:

- If $B_uB_v>0$ then there exists $C\leq 2aB_uB_v\left(B_u^2+B_v^2\right)^{-1}$ such that the inequality (2.4.8) holds uniformly.

- If $B_uB_v<0$, consider some $\delta_0>-2aB_vB_u^{-1}$. Then there exists $C(\delta_0)>0$ such that the inequality (2.4.8) holds uniformly as soon as $\Delta x\geq\delta_0\varepsilon$ with $C=C(\delta_0)$. For example, if we choose $\delta_0=-3aB_vB_u^{-1}$ then there exists $C\leq -aB_uB_v\left(B_u^2+B_v^2\right)^{-1}$ such that (2.4.8) holds uniformly.

This ends the proof of Proposition 2.4.1.

2.4.2 Numerical experiments

In this section we perform some numerical experiments and observe the effective behavior (i.e. the time evolution) of the energy

$$E(t^n) := \langle U^n, HU^n \rangle_{\Delta x},$$

according to whether or not the discrete strict dissipativity condition (2.2.10) is valid. We also have a look at the degenerate case when the UKC (2.1.5) does not hold (and thus, none of the other stability conditions). As discussed in the previous section and in the calculations of Xin and Xu [96], we expect to observe the decrease of the energy $E(t)$ as soon as $B_u B_v > 0$. What happens in the case $B_u B_v < 0$, but while the the discrete strict dissipativity condition (2.2.10) still holds, is also experimented.

As main parameters for the experiments, we fix the space step $\Delta x = 10^{-2}$, choose $a = 4$ and let ε and the boundary parameter (B_u, B_v) vary. The initial data is

$$f_j = \begin{cases} (0,0), & \text{if } x_j = 0, \\ (15,10)^T, & \text{if } 0 < x_j \leq 1/2, \\ (0,0), & \text{if } x_j > 1/2. \end{cases}$$

Following Section 2.2.1.2, we choose for our experiments the space interval $[0,1]$ and the time interval $[0,T]$ with $T = 0.2$. The most natural choice in the right boundary condition at $x = 1$ is to select the homogeneous Neumann boundary condition $U^n_{J+1} = U^n_J$ at the rightmost cell J.

Firstly, we choose a set of values (B_u, B_v) such that the discrete strict dissipativity condition (2.2.10) is satisfied with $\varepsilon = 10^{-2}$ and also with $\varepsilon = 10^2$. The Figure 2.4.1 shows the evolution of the energy $E(t^n)$ over the time interval $t^n \in [0, 0.2]$.

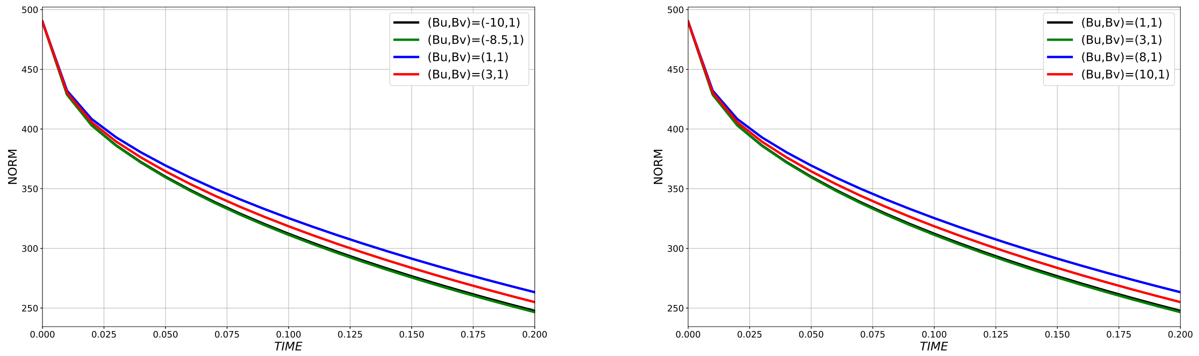

Figure 2.4.1: Energy evolution with the discrete strict dissipativity condition (2.2.10), for $\varepsilon = 10^{-2}$ (left) and $\varepsilon = 10^2$ (right).

We proved that for any $\varepsilon \in (0, +\infty)$ and (B_u, B_v) satisfying the discrete strict dissipativity condition (2.2.10), $E(t^n)$ is decreasing. This is strongly supported by the experiments. Observe also that the decrease of $E(t^n)$ is true even in the case $B_u B_v < 0$ provided the discrete strict dissipativity condition (2.2.10) is true. This is the case for example for $\varepsilon = 10^{-2}$ together with the parameters $(B_u, B_v) = (-10, 1)$.

Secondly, we choose a set of values (B_u, B_v) such that the discrete strict dissipativity condition (2.2.10) is not satisfied with $\varepsilon = 10^{-2}$ nor with $\varepsilon = 10^2$. Besides, we also present the evolution of the energy for parameters such that the Uniform Kreiss Condition (2.1.5) is wrong. The Figure 2.4.2 shows the evolution of $E(t^n)$ over the time interval $t^n \in [0, 0.2]$.

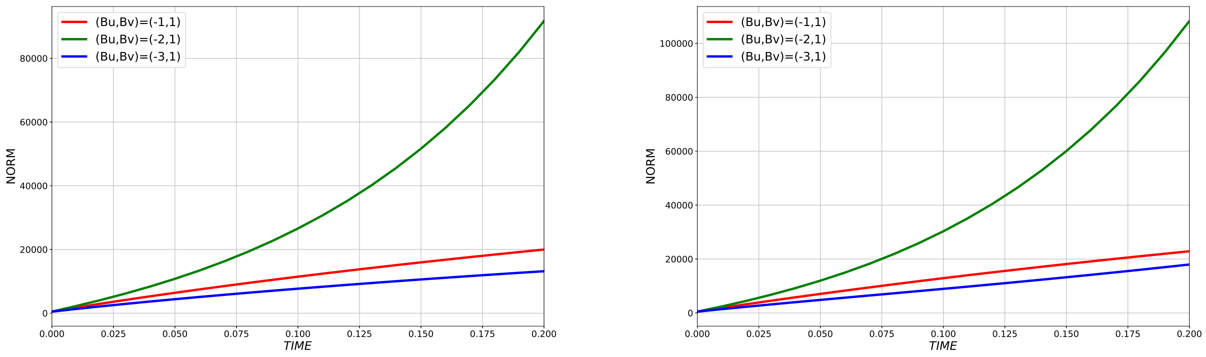

Figure 2.4.2: Energy evolution without the discrete strict dissipativity condition (2.2.10), for $\varepsilon = 10^{-2}$ (left) and $\varepsilon = 10^{2}$ (right).

- On the boundary $x_j = 0$, for all $\varepsilon > 0$, if the boundary condition (2.1.3) with homogeneous boundary condition $b^n \equiv 0$, for any $n \in \mathbb{N}$ does not satisfy the UKC, then $v_0^n = \sqrt{a}u_0^n$. Therefore, the numerical scheme of the IBVP is not stable for each fixed ε. For $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{2}$, if we choose $(B_u, B_v) = (-2, 1)$ then the values of $E(t^n)$ increase quickly.

- When the discrete strict dissipativity condition (2.2.10) fails, then we observe for any $t^n \in (0, 0.2]$ the inequality $E(t^n) > E(0)$.

Clearly, the numerical results show that the energy $E(t^n)$ increases in time as soon as the discrete strict dissipativity condition (2.2.10) does not hold. The behavior is even worse when the UKC (2.1.5) is not satisfied. It seems that the condition (2.2.10) is also necessary to ensure the non-increase of the energy.

2.5 Appendix A. Technical lemmas

Lemma 2.5.1. *Let $\zeta \in \mathbb{C}$ with* $\mathrm{Re}(\zeta) > 0$ *and* $h(\zeta) = \sqrt{\zeta(1+\zeta)}$, *then* $\mathrm{Re}(\zeta) \leq \mathrm{Re}(h(\zeta))$.

Proof. In the half plane $\{\zeta \in \mathbb{C} : \mathrm{Re}(\zeta) \geq 0\}$, the complex function $h(\zeta)$ is analytic. As usual, we take $\sqrt{\zeta}$ to be the principal branch with the branch cut along the negative real axis.

Let $\zeta = x + yi$ with $x > 0$, $y \in \mathbb{R}$ and

$$p = x(1+x) - y^2, \qquad q = (1+2x)y.$$

Then,

$$\mathrm{Re}(h(\zeta)) = \mathrm{Re}\left(\sqrt{p+qi}\right) = \sqrt{\frac{p + \sqrt{p^2+q^2}}{2}}.$$

Now, we observe that

$$\begin{aligned}\sqrt{p^2+q^2} &= \sqrt{(x(1+x)-y^2)^2 + (1+2x)^2y^2} \\ &= \sqrt{(x(1+x)+y^2)^2 + y^2} \\ &\geq x(1+x) + y^2.\end{aligned}$$

Therefore,

$$\mathrm{Re}(h(\zeta)) \geq \sqrt{x(1+x)} \geq x.$$

This ends the proof of Lemma 2.5.1. □

Lemma 2.5.2. *Let* $\zeta \in \mathbb{C}$ *with* $\mathrm{Re}(\zeta) < 0$*, then* $|\zeta + \sqrt{\zeta^2+1}| \leq \mathrm{Re}(\zeta) + \sqrt{(\mathrm{Re}(\zeta))^2+1} < 1$*.*

Proof. Assume that $\zeta = x + yi$ with $x < 0$ and $y \in \mathbb{R}$.

Case 1 Consider first the easy case $y = 0$. Then

$$\left|\zeta + \sqrt{\zeta^2+1}\right| = x + \sqrt{x^2+1},$$

but since $x < 0$, one obtains by simple considerations the inequality $x + \sqrt{x^2+1} < 1$.

Case 2 In the general case $y \neq 0$, let us begin with some notations:

$$\zeta^2 + 1 = p_1 + q_1 i, \quad \text{with } p_1 = x^2 - y^2 + 1 \text{ and } q_1 = 2xy,$$
$$\sqrt{\zeta^2+1} = a_1 + b_1 i, \quad \text{with } a_1 = \sqrt{\frac{p_1 + \sqrt{p_1^2+q_1^2}}{2}} \text{ and } b_1 = \mathrm{sgn}(q_1)\sqrt{\frac{-p_1 + \sqrt{p_1^2+q_1^2}}{2}}.$$

Together with these notations, some algebraic identities are available:

$$x^2 + b_1^2 + 1 = a_1^2 + y^2 \quad \text{and} \quad y = \frac{a_1 b_1}{x}, \tag{2.5.1}$$

Firstly, we prove the next inequality

$$a_1 x^2 + b_1^2 x + a_1 b_1^2 \geq 0. \tag{2.5.2}$$

We can see that the inequality (2.5.2) is equivalent to $a_1(x^2 + b_1^2) \geq -x b_1^2$ and since $x < 0$, the latter is now equivalent to its squared version, that reads

$$a_1^2 x^2 (x^2 + 2b_1^2) \geq b_1^4 (x^2 - a_1^2).$$

By the definition of a_1, b_1 above, the previous inequality is successively

$$\begin{aligned} &4x^2\left(p_1 + \sqrt{p_1^2+q_1^2}\right)\left(x^2 - p_1 + \sqrt{p_1^2+q_1^2}\right) \\ &\geq \left(-p_1 + \sqrt{p_1^2+q_1^2}\right)^2 \left(2x^2 - p_1 - \sqrt{p_1^2+q_1^2}\right). \end{aligned}$$

It is equivalent to

$$\begin{aligned} &4x^4\left(p_1 + \sqrt{p_1^2+q_1^2}\right) + 2x^2 q_1^2 \geq \left(p_1 - \sqrt{p_1^2+q_1^2}\right)\left(4x^2 p_1 + q_1^2\right) \\ \Leftrightarrow\ &4x^4\left(p_1 + \sqrt{p_1^2+q_1^2}\right) + 2x^2 \times 4x^2y^2 \geq \left(p_1 - \sqrt{p_1^2+q_1^2}\right)\left(4x^2(x^2-y^2+1) + 4x^2y^2\right) \\ \Leftrightarrow\ &2x^2\left(\sqrt{p_1^2+q_1^2} + y^2\right) \geq p_1 - \sqrt{p_1^2+q_1^2}. \end{aligned}$$

But, for any $p_1, q_1 \in \mathbb{R}$, this is easy to see that $p_1 - \sqrt{p_1^2+q_1^2} \leq 0$ and thus any of the previous inequalities and so the expected one (2.5.2) follow.

Now let us observe that the required inequality $|\zeta + \sqrt{\zeta^2+1}| \leq x + \sqrt{x^2+1}$ is fully equivalent to

$$(x + a_1)^2 + (y + b_1)^2 \leq (x + \sqrt{x^2+1})^2, \tag{2.5.3}$$

that we prove now. According to the algebraic identities in (2.5.1), by eliminating the occurences of y, the previous formula is equivalent to

$$a_1 x + b_1^2 + \frac{a_1 b_1^2}{x} \leq x\sqrt{x^2+1}.$$

In addition, we observe that $x^2 + 1 = x^{-2}\left(a_1^2 x^2 + a_1^2 b_1^2 - b_1^2 x^2\right)$, and thus the previous inequality is equivalent to

$$a_1 x + b_1^2 + \frac{a_1 b_1^2}{x} \leq -\sqrt{a_1^2 x^2 + a_1^2 b_1^2 - b_1^2 x^2}. \tag{2.5.4}$$

Since $x < 0$ and from the inequality (2.5.2), the formula (2.5.4) reads also

$$\begin{aligned} &\left(a_1 x^2 + b_1^2 x + a_1 b_1^2\right)^2 \geq x^2\left(a_1^2 x^2 + a_1^2 b_1^2 - b_1^2 x^2\right) \\ \Leftrightarrow\ &(x+a_1)^2 (x^2 + b_1^2) \geq 0. \end{aligned}$$

This ends the proof of the inequality (2.5.3). Now since $\mathrm{Re}(\zeta) < 0$, the analysis of the first easy case again applies to get $\mathrm{Re}(\zeta) + \sqrt{(\mathrm{Re}(\zeta))^2 + 1} < 1$.

This ends the proof of Lemma 2.5.2. □

Lemma 2.5.3. *Let $a > 0$ be fixed and consider for any $\xi \in \mathbb{C}$ with $\mathrm{Re}\,(\xi) \in [0,1]$ and $\mathrm{Im}\,(\xi) \in [-1,1]$:*

$$\mu_+(\xi) = \sqrt{\frac{\xi(1+\xi)}{a}}.$$

There exists a constant $c > 0$, independent of δ and ξ such that

$$\mathrm{Re}\left(\sqrt{1 + (\mu_+(\xi)\delta)^2}\right) \geq c.$$

Proof. Let us denote $\xi = \alpha + i\beta$ with $\alpha \in [0,1]$ and $\beta \in [-1,1]$ and introduce the notation $\overline{\delta} = \delta^2/a$. After some calculations, one obtains

$$\mathrm{Re}\left(\sqrt{1 + (\mu_+(\xi)\delta)^2}\right) = \frac{1}{\sqrt{2}}\sqrt{h(\alpha,\beta,\overline{\delta})},$$

with the function with positive real values:

$$h(\alpha,\beta,\overline{\delta}) = 1 + \overline{\delta}\left(\alpha(1+\alpha) - \beta^2\right) + \sqrt{\left(1 + \overline{\delta}\left(\alpha(1+\alpha) - \beta^2\right)\right)^2 + \overline{\delta}^2 \beta^2 (1+2\alpha)^2}.$$

Now, the required uniform lower bound will be provided directly by a uniform lower bound for the quantity $h(\alpha,\beta,\overline{\delta})$, what we are looking for now by exhaustion.

Case 1 For any $\overline{\delta} > 0$, $\alpha \in [0,1]$ and $\beta = 0$, we have the simple lower bound

$$h(\alpha,0,\overline{\delta}) = 2\left(1 + \overline{\delta}\alpha(1+\alpha)\right) \geq 2.$$

Case 2 For any $\overline{\delta} > 0$, $\alpha \in [0,1]$ and $\beta \in [-1,0) \cup (0,1]$ such that $\beta^2 \leq \alpha(1+\alpha)$, we get

$$h(\alpha,\beta,\overline{\delta}) \geq 1 + \sqrt{1 + \overline{\delta}^2 \beta^2 (1+2\alpha)^2} \geq 2.$$

Case 3 For any $\overline{\delta} > 0$, $\alpha \in [0,1]$ and $\beta \in [-1,0) \cup (0,1]$ such that $\beta^2 > \alpha(1+\alpha)$, let us introduce the quantity $\tau := \beta^2 - \alpha(1+\alpha)$, so that

$$h(\alpha,\beta,\overline{\delta}) = 1 - \tau\overline{\delta} + \sqrt{\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}^2\beta^2(1+2\alpha)^2}.$$

Note that α being nonnegative

$$0 < \tau \leq \beta^2 \leq 1. \tag{2.5.5}$$

Subcase 3.a Assume that $\overline{\delta} \leq \tau^{-1}$. Then, $1 - \tau\overline{\delta} \geq 0$ and therefore

$$h(\alpha,\beta,\overline{\delta}) \geq \sqrt{\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}^2\beta^2(1+2\alpha)^2}.$$

We then can compute

$$h^2(\alpha,\beta,\overline{\delta}) \geq \left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}^2\beta^2(1+2\alpha)^2 \geq \left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}^2\beta^2.$$

From (2.5.5), we then have successively

$$\begin{aligned} h^2(\alpha,\beta,\overline{\delta}) &\geq 1 - 2\tau\overline{\delta} + \tau^2\overline{\delta}^2 + \overline{\delta}^2\tau \geq 1 - 2\tau\overline{\delta} + \tau^2\overline{\delta}^2\left(1+\frac{1}{\tau}\right) \\ &\geq 1 - 2\tau\overline{\delta} + 2\tau^2\overline{\delta}^2 \geq \frac{1}{2}, \end{aligned}$$

where the last inequality comes from the property $\tau\overline{\delta} \in (0,1]$. Thus we get

$$h(\alpha,\beta,\overline{\delta}) \geq \frac{1}{\sqrt{2}}.$$

Subcase 3.b The last case is for $\overline{\delta} > \tau^{-1}$.Then we can rewrite

$$\begin{aligned} h(\alpha,\beta,\overline{\delta}) &= -\sqrt{\left(1-\tau\overline{\delta}\right)^2} + \sqrt{\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}^2\beta^2(1+2\alpha)^2} \\ &\geq -\sqrt{\left(1-\tau\overline{\delta}\right)^2} + \sqrt{\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}^2\beta^2}. \end{aligned}$$

From (2.5.5) and the subcase assumption, we get successively

$$\overline{\delta}\beta^2 > \frac{\beta^2}{\tau} \geq 1.$$

Thus,

$$\begin{aligned} h(\alpha,\beta,\overline{\delta}) &\geq -\sqrt{\left(1-\tau\overline{\delta}\right)^2} + \sqrt{\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}} \\ &\geq \frac{\overline{\delta}}{\sqrt{\left(1-\tau\overline{\delta}\right)^2} + \sqrt{\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}}} \\ &\geq \frac{\overline{\delta}}{2\sqrt{\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta}}}. \end{aligned}$$

On the other hand, from (2.5.5) and the subcase assumption and since $\overline{\delta}^2 \geq \overline{\delta} \geq 1$, we have successively

$$\left(1-\tau\overline{\delta}\right)^2 + \overline{\delta} \leq 1 + \tau^2\overline{\delta}^2 + \overline{\delta} \leq 1 + (\tau^2+1)\overline{\delta}^2 \leq 1 + 2\overline{\delta}^2.$$

Thus finally, and since $\overline{\delta} > 1$, we have

$$h(\alpha,\beta,\overline{\delta}) \geq \frac{\overline{\delta}}{2\sqrt{1+2\overline{\delta}^2}} \geq \frac{1}{2\sqrt{3}}.$$

□

2.6 Appendix B. Modeling an elastic string

2.6.1 Derivation of the damped wave equation

The damped wave equation in one space dimension can be derived in a variety of different physical settings. As an example of how waves occur in physical systems, we now derive the damped wave equation for a stretched string. Other physical systems, such as sound waves in air, can be analyzed in a similar way (see [70, 40]). We start by considering model the action of an elastic string over time. Consider a tiny element of the string between x and $x + \Delta x$

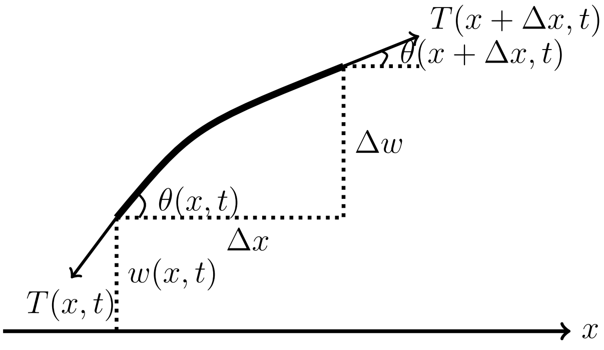

Figure 2.6.1: Modeling an Elastic String.

The following quantities are needed in our derivation (see Figure 2.6.1):

- $w(x,t)$ denotes vertical displacement of the string from the $x-$axis at position x and time t.
- $\theta(x,t)$ is an angle between the string and a horizontal line at position x and time t.
- $T(x,t)$ is a tension in the string at position x.

We can dispose of all the θ's observing from the figure that

$$\tan\theta(x,t) = \lim_{\Delta x\to 0}\frac{\Delta w}{\Delta x} = \frac{\partial w}{\partial x} : \text{slope of tangent at } (x,t) \text{ in } wx - \text{plane} \tag{2.6.1}$$

The Newton's Second Law of Motion ($F = ma$) states that

$$F = (\rho\Delta x)\frac{\partial^2 w}{\partial t^2} \tag{2.6.2}$$

where ρ is the linear density of the string and Δx is the length of the segment.

The force F comes from the tension in the string and also the damping force. The damping force acts in the opposite direction to the motion and is denoted by $-c\dfrac{\partial w}{\partial t}$ with $c > 0$. We assume for our model that there are only transverse vibrations, and so the string does not move horizontally, but only vertically. So, we know that the total horizontal force must be zero. Balancing the forces in the horizontal direction gives

$$T(x+\Delta x,t)\cos\theta(x+\Delta x,t) = T(x,t)\cos\theta(x,t) = \tau, \tag{2.6.3}$$

where τ is the constant horizontal tension. Balancing the forces in the vertical direction yields

$$\begin{aligned} F &= T(x+\Delta x,t)\sin\theta(x+\Delta x,t) - T(x,t)\sin\theta(x,t) - c\frac{\partial w}{\partial t}\Delta x \\ &= T(x+\Delta x,t)\cos\theta(x+\Delta x,t)\tan\theta(x+\Delta x,t) - T(x,t)\cos\theta(x,t)\tan\theta(x,t) - c\frac{\partial w}{\partial t}\Delta x. \end{aligned} \tag{2.6.4}$$

Substituting (2.6.3) into (2.6.4) yields,

$$F = \tau\left(\tan\theta(x+\Delta x,t) - \tan\theta(x,t)\right) - c\frac{\partial w}{\partial t}\Delta x = \tau\left(\frac{\partial w}{\partial x}(x+\Delta x,t) - \frac{\partial w}{\partial x}(x,t)\right) - c\frac{\partial w}{\partial t}\Delta x.$$

So, the vertical component of Newton's Law becomes

$$\rho\frac{\partial^2 w}{\partial t^2}(\xi,t) = \tau\frac{1}{\Delta x}\left(\frac{\partial w}{\partial x}(x+\Delta x,t) - \frac{\partial w}{\partial x}(x,t)\right) - c\frac{\partial w}{\partial t}$$

for $\xi \in [x, x+\Delta x]$. Dividing by ρ and letting Δx tends to 0 gives

$$\frac{\partial^2 w}{\partial t^2} = \frac{\tau}{\rho}\frac{\partial^2 w}{\partial x^2} - \frac{c}{\rho}\frac{\partial w}{\partial t}. \tag{2.6.5}$$

In order to guarantee that the equation (2.6.5) has a unique solution, some initial and boundary conditions have to be suitably selected: two initial conditions and boundary condition (see [70, 40]).

2.6.2 Initial conditions

The initial position of the string and its initial velocity may be written as follow

$$w(x,0) = f(x) \quad \text{and} \quad w_t(x,0) = h(x). \tag{2.6.6}$$

To see why we need two initial condition, note that the Taylor series of $w(x,t)$ about $t = 0$ is

$$w(x,t) = w(x,0) + w_t(x,0)t + w_{tt}(x,0)\frac{t^2}{2} + w_{ttt}(x,0)\frac{t^3}{3!} + ...$$

From the initial condition (2.6.6) and the PDE (2.6.5) give

$$\begin{aligned} w_{tt}(x,0) &= (\tau/\rho)w_{xx}(x,0) - (c/\rho)w_t(x,0) = (\tau/\rho)f''(x) - (c/\rho)h(x), \\ w_{ttt}(x,0) &= (\tau/\rho)w_{txx}(x,0) - (c/\rho)w_{tt}(x,0) = (\tau/\rho)h''(x) - (c\tau/\rho^2)f''(x) + (c/\rho)^2h(x). \end{aligned}$$

Higher order terms can be found similarly. Therefore, the two initial conditions for $w(x,0)$ and $w_t(x,0)$ are sufficient to determine $w(x,t)$ near $t = 0$.

2.6.3 Boundary condition

We assumed the string is connected to frictionless cylinders of mass m_1 that move vertically on tracks at $x = 0$ with an acceleration $g(t)$.

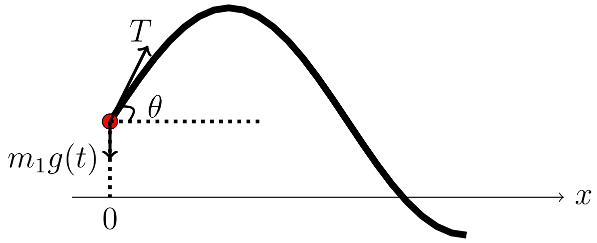

Figure 2.6.2: Boundary condition at $x = 0$.

Lemma 2.6.1. *For any $c > 0$, $\rho > 0$ and $m_1 > 0$, the boundary condition can be rewritten as follows*

$$B_u w_x(0,t) - B_v w_t(0,t) = g(t).$$

with $B_u B_v > 0$.

Proof. Performing the force balance at $x = 0$ gives

$$T \sin\theta - c\frac{\partial w}{\partial t} = m_1 g(t).$$

In other words, the vertical tension in the string balances the mass of the cylinder. However, $\tau = T\cos\theta = const$ and $\tan\theta = w_x$, so that the previous equation becomes

$$T\cos\theta\tan\theta - c\frac{\partial w}{\partial t} = m_1 g(t),$$

or also, denoting $B_u = \tau/m_1$ and $B_v = c/m_1$

$$B_u w_x(0,t) - B_v w_t(0,t) = g(t).$$

□

To summarize, the IBVP of the linear damped wave equation in one space dimension reads

$$\begin{aligned} \text{PDE } &: \frac{\partial^2 w}{\partial t^2}(x,t) = a\frac{\partial^2 w}{\partial x^2}(x,t) - \frac{1}{\varepsilon}\frac{\partial w}{\partial t}(x,t), && x > 0, t > 0, \\ \text{BC} &: B_u w_x(0,t) - B_v w_t(0,t) = g(t), && t > 0, \\ \text{IC} &: w(x,0) = f(x), \quad w_t(x,0) = h(x) && t > 0. \end{aligned}$$

where $a = \tau/\rho$, and $\varepsilon = \rho/c$. Let now denote $u^\varepsilon(x,t) = w_x(x,t)$ and $v^\varepsilon(x,t) = -w_t(x,t)$. The previous IBVP can be represented as

$$\begin{aligned} &\partial_t u^\varepsilon(x,t) + \partial_x v^\varepsilon(x,t) = 0, \\ &\partial_t v^\varepsilon(x,t) + a\partial_x u^\varepsilon(x,t) = -\frac{1}{\varepsilon}v^\varepsilon(x,t), \end{aligned} \tag{2.6.7}$$

with the initial data

$$u^\varepsilon(x,0) = f'(x), \quad v^\varepsilon(x,0) = -h(x),$$

and the linear boundary condition

$$B_u u^\varepsilon(0,t) + B_v v^\varepsilon(0,t) = g(t).$$

Remark 2.6.2. *The boundary condition $B_u B_v > 0$ corresponds to stability condition (2.2.8) of the linear damped wave equation (2.6.7).*

Chapter 3

A stiffly stable fully discrete scheme for the damped wave equation using discrete transparent boundary condition

We study the stability analysis of the time-implicit central differencing scheme for the linear damped wave equation with boundary. In [96], Xin and Xu prove that the initial-boundary value problem (IBVP) for this model is well-posed, uniformly with respect to the stiffness of the damping, under the so-called stiff Kreiss condition (SKC) on the boundary condition. We show here that the (SKC) is also a sufficient condition to guarantee the uniform stability of the discrete IBVP for the relaxation system independently of the stiffness of the source term, of the space step and of the time step. The boundary is approximated using discrete transparent boundary conditions and the stiff stability is proved using energy estimates and the $\mathcal{Z}-$ transform.

3.1 Description of the numerical scheme

Let $\Delta t > 0$ being the time step. The space step $\Delta x > 0$ will always be chosen so that the parameter $\lambda_{xt} = \Delta x \Delta t^{-1}$ is kept fixed. Letting now $U_j^n = (u_j^n, v_j^n)^T$ denotes the approximation of the exact solution to (2.1.1)-(2.1.3) at the grid point $(x_j, t^n) = (j\Delta x, n\Delta t)$, for any $(j, n) \in \mathbb{N} \times \mathbb{N}$ (where we omit the explicit dependence on ε). We focus in this chapter on the fully discrete approximation of the IBVP (2.1.1)-(2.1.3) obtained by the central differencing scheme in space and the implicit scheme in time.

A first step towards the fully discrete approximation of the IBVP (2.1.1)-(2.1.3) is the following system

$$\begin{cases} \dfrac{U_j^{n+1} - U_j^n}{\Delta t} + \dfrac{1}{2\Delta x} A \left(U_{j+1}^{n+1} - U_{j-1}^{n+1} \right) = \dfrac{1}{\varepsilon} S U_j^{n+1}, & j \geq 1, n \geq 0, \\ U_j^0 = f_j, & j \geq 0, \\ B U_0^n = b^n, & n \geq 0, \end{cases} \tag{3.1.1}$$

where the approximations of the initial condition f_j and of the boundary data b^n are defined for example by setting $f_j = f(j\Delta x)$ for $j \geq 0$ and $b^n = b(n\Delta t)$ for $n \geq 0$.

Let us emphasize that the numerical scheme (3.1.1) still needs one more scalar equation at the boundary point $j = 0$ so as to be fully defined, due to the fact that the matrix B has rank one only. This is actually a discrete feature only, since in the continuous case this single equation is exactly complemented by the only incoming characteristic (at least under UKC). An additional relation to define U_0^{n+1} is thus needed. We want to use the central scheme at the boundary point j $=0$, so that the modification of the ghost value U_{-1}^{n+1} can also be interpreted as the use of an extra boundary condition. From a mathematical point of view, the problem is set, in both cases, as follows: given an initial data compactly supported, one can construct boundary condition at $j=0$ with the objective to approximate the exact solution of the whole space problem $\{j \in \mathbb{Z}\}$, restricted to $\{j \in \mathbb{N}\}$. If the approximate solution on $\{j \in \mathbb{N}\}$ coincides with the exact solution, one refers to these boundary conditions as transparent boundary conditions. Of course, these boundary condition should lead to a well-posed initial boundary value problem. It means that we use the discrete transparent boundary condition at $j=0$ that determines a ghost value U_{-1}^{n+1} through the identity

$$U_{-1}^{n+1} = \sum_{k=0}^{n+1} \mathcal{C}_{n+1-k} U_0^k,$$

where the coefficients $\mathcal{C}_k$ will be precised explicitly in the forthcoming Definition 3.2.4. The extra boundary condition determines U_{-1}^{n+1} as a linear function of U_0^k for past step times only: $0 \leq k \leq n+1$. We propose the following numerical approximation at the boundary:

$$\frac{1}{\Delta t}\Gamma\left(U_0^{n+1} - U_0^n\right) + \frac{1}{2\Delta x}\Gamma A\left(U_1^{n+1} - \sum_{k=0}^{n+1}\mathcal{C}_{n+1-k}U_0^k\right) = \frac{1}{\varepsilon}\Gamma S U_0^{n+1}$$

with the matrix $\Gamma = (-aB_v \quad B_u)$. Under the SKC, this choice for the matrix Γ will be useful to construct the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ in the Propositions 3.2.3 and 3.3.1.

To summarize, we study all along this chapter the following fully discrete approximation of the IBVP (2.1.1)-(2.1.3):

$$\begin{cases} \dfrac{U_j^{n+1} - U_j^n}{\Delta t} + \dfrac{1}{2\Delta x} A\left(U_{j+1}^{n+1} - U_{j-1}^{n+1}\right) = \dfrac{1}{\varepsilon} S U_j^{n+1}, & j \geq 1,\ n \geq 0, \\ U_j^0 = f_j, & j \geq 0, \\ BU_0^n = b^n, & n \geq 0, \\ \dfrac{1}{\Delta t}\Gamma\left(U_0^{n+1} - U_0^n\right) + \dfrac{1}{2\Delta x}\Gamma A\left(U_1^{n+1} - \displaystyle\sum_{k=0}^{n+1}\mathcal{C}_{n+1-k}U_0^k\right) = \dfrac{1}{\varepsilon}\Gamma S U_0^{n+1}, & n \geq 0. \end{cases} \tag{3.1.2}$$

Main result: Dealing with the continuous IBVP (2.1.1)-(2.1.3), the UKC (2.1.5) is not enough and a more stringent restriction has to be imposed. Our aim is to prove that the SKC derived in [96] is then a sufficient condition for the stiff stability of the fully discrete IBVP (3.1.2), in other words the uniform stability with respect to the stiffness of the relaxation term.

Theorem 3.1.1 (Main result)**.** *Assume that $(B_u, B_v) \in \mathbb{R}^2$ satisfies the SKC*

$$B_v = 0 \quad or \quad \frac{B_u}{B_v} \notin \left[-\sqrt{a}, 0\right]. \tag{3.1.3}$$

Let $\lambda_{xt} \leq 3\sqrt{a}/8$ be a positive number. For any $T > 0$, there exists a constant $C_T > 0$ such that for all $\Delta t > 0$ and $\Delta x = \lambda_{xt}\Delta t$, any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R})$, the solution $(U_j^n)_{j\in\mathbb{N}}$ to the scheme (3.1.2) satisfies

$$\sum_{n=0}^{N}\sum_{j\geq 0}\Delta x \Delta t |U_j^n|^2 + \sum_{n=0}^{N}\Delta t |U_0^n|^2 \leq C_T\left(\sum_{j\geq 0}\Delta x |f_j|^2 + \sum_{n=0}^{N}\Delta t |b^n|^2\right), \tag{3.1.4}$$

where $N := T/\Delta t$ and C_T is independent of $\varepsilon \in (0, +\infty)$.

In [96], Xin and Xu considered the IBVP for the Jin-Xin relaxation model [51] and derived the SKC (3.1.3) to characterize its stiff well-posedness. They show in particular that the IBVP (2.1.1)-(2.1.3) is well-posed if and only if (3.1.3) holds. In the discrete IBVP (3.1.2), it seems that the SKC is also sufficient to derive uniform stability estimates. Besides, by linearity, the numerical scheme of the IBVP (3.1.2) can be broken up into two simpler problems, one with homogeneous initial condition $(f_j)_{j\in\mathbb{N}} \equiv 0$ and the other with homogeneous boundary $b^n \equiv 0$, for any $n \in \mathbb{N}$. The proof of Theorem 3.1.1 is based on two main ingredients, by assembling a result for the case of the following Cauchy problem

$$\begin{cases} \dfrac{\left(U_j^I\right)^{n+1} - \left(U_j^I\right)^n}{\Delta t} + \dfrac{1}{2\Delta x} A\Big(\left(U_{j+1}^I\right)^{n+1} - \left(U_{j-1}^I\right)^{n+1} \Big) = \dfrac{1}{\varepsilon} S \left(U_j^I\right)^{n+1}, & j \in \mathbb{Z},\ n \geq 0, \\ \left(U_j^I\right)^0 = f_j, & j \in \mathbb{Z}. \end{cases} \tag{3.1.5}$$

and another one for the problem (3.1.2) with homogeneous initial data. We state hereafter these two statements.

Proposition 3.1.2 (Cauchy problem)**.** *For any $T > 0$, there exists $C_T > 0$ such that for all $\Delta t > 0$, any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, the solution $(U_j^I)_{j\in\mathbb{Z}}^n$ to (3.1.5) satisfies*

$$\sum_{j\in\mathbb{Z}} \Delta x |(U_j^I)^n|^2 \leq C_T \sum_{j\in\mathbb{Z}} \Delta x |f_j|^2, \qquad n \in \mathbb{N}, \tag{3.1.6}$$

where C_T is independent of $\varepsilon \in (0, +\infty)$ and $\Delta x = \lambda_{xt}\Delta t$.

Proposition 3.1.3 (Homogeneous initial condition)**.** *Assume that the SKC (3.1.3) is satisfied. Then, there exists a constant $C > 0$ such that for any $\gamma > 0$ and any positive constant $\lambda_{xt} \leq 3\sqrt{a}/8$, the following property holds. For any $\Delta t > 0$ together with $\Delta x = \lambda_{xt}\Delta t$ and any boundary data $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R})$, the solution $(U_j^n)_{j\in\mathbb{N}}$ to (3.1.2) with $(f_j)_{j\in\mathbb{N}} \equiv 0$ satisfies*

$$\frac{\gamma}{\gamma\Delta t + 1} \sum_{n\geq 0} \sum_{j\geq 0} e^{-2\gamma n\Delta t} \Delta t \Delta x |U_j^n|^2 + \sum_{n\geq 0} e^{-2\gamma n\Delta t} \Delta t |U_0^n|^2 \leq C \sum_{n\geq 0} e^{-2\gamma n\Delta t} \Delta t |b^n|^2, \tag{3.1.7}$$

where C is independent of $\varepsilon \in (0, +\infty)$.

To isolate the effects of a possible boundary layer and avoid the complicated interaction of boundary and initial layers, in Section 3.2, we consider the IBVP (3.1.2) with homogeneous initial data and nonzero boundary data b^n, for any $n \geq 0$. The numerical solution $(U_j^n)_{j\in\mathbb{N}}$ is constructed in Section 3.2.2 thanks to the $\mathcal{Z}-$transform [52, 75]. Furthermore, we follow the discrete transparent boundary condition at $j = 0$ as proposed in [2, 5, 54] to find the explicit formula of the sequence $(\mathcal{C}_m)_{m\geq 0}$. By using the Plancherel's theorem, under the SKC, the Proposition 3.1.3 is proved in Section 3.2.3. In order to illustrate the relevance of the SKC (3.1.3), we present in Section 3.2.4 some numerical results, for various values of the parameters (B_u, B_v) and show that the numerical solution at the boundary $x = 0$ increases quickly if the SKC (3.1.3) does not hold. Besides, by the decrease of the error $\|U(.,t^n) - U^n\|^2_{\ell^2(\mathbb{N},\mathbb{R}^2)}$, we can observe the convergence of the discrete solution U_j^n to the exact one $U(x_j, t^n)$. After that, we observe the behavior of the energy terms $\|U\|_{\ell^2(\mathbb{N}\times[0,T),\mathbb{R}^2)}$ and $\|U\|_{\ell^2(\{0\}\times[0,T),\mathbb{R}^2)}$ corresponding to whether or not the SKC (3.1.3) is valid. The nonzero initial data case is much more difficult with the sufficiency proof. This is due to the complicated interactions between the initial data, the boundary condition and the stiff relaxation term. Under the SKC, the numerical solution is again described by means of the $\mathcal{Z}-$transform in Section 3.3.1. It is decomposed into three parts, by assembling the solution for the case of Cauchy problem (3.1.5), a numerical error term $(U_j^{II})_{j\in\mathbb{N}}^n$ and the solution for the case IBVP (3.1.2) with homogeneous initial data. Since the coefficients for computing the boundary value U_{-1}^{n+1} are defined for homogeneous initial data, this numerical error $(U_j^{II})_{j\in\mathbb{N}}^n$ is due to the interaction between the Cauchy problem and the IBVP with zero initial data. For the Cauchy problem, the Proposition 3.1.2 is studied in Section 3.3.2 by means of the discrete energy method. By an application of the Plancherel's theorem for $\mathcal{Z}$-transform, the numerical error term $(U_j^{II})_{j\in\mathbb{N}}^n$ will be estimated in Section 3.3.3. In Section 3.3.4, we get the expected result of the Theorem 3.1.1 in the case IBVP with nonzero initial condition. In Section 3.3.5, we also look at the behavior of the numerical solution $(U_j^n)_{j\in\mathbb{Z}}$ and the energy terms $\|U\|_{\ell^2(\mathbb{N}\times[0,T),\mathbb{R}^2)}$ and $\|U\|_{\ell^2(\{0\}\times[0,T),\mathbb{R}^2)}$ corresponding to whether or not the SKC (3.1.3) is valid. It seems that the SKC (3.1.3) is also necessary condition to guarantee the uniform stability of the IBVP (3.1.2) independent of the effect of the relaxation source term and the boundary dissipation.

3.2 Stiff stability of the IBVP with homogeneous initial condition

In this section, we consider the discrete IBVP (3.1.2) with nonzero boundary condition $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N},\mathbb{R})$ and homogeneous Cauchy data $(f_j)_{j\in\mathbb{N}} \equiv 0$. Assuming that the SKC is satisfied, the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ is obtained by using the $\mathcal{Z}$-transform [52, 75]. Thanks to the Plancherel's theorem, we then are able to get the expected result of the Proposition 3.1.3.

3.2.1 Notations and preliminary results

Before we enter the important proofs, let us introduce some notations and preliminary results. All along this chapter, the complex values z and ξ are related through the formula

$$\xi = \left(1 - z^{-1}\right)\Delta t^{-1}, \quad z = R\mathrm{e}^{i\theta}, \quad \text{with } R > 1,\ \theta \in (-\pi, \pi].$$

Then, ξ obeys the inequalities

$$\left(1 - R^{-1}\right)\Delta t^{-1} \leq \operatorname{Re}\xi \leq 2\Delta t^{-1}. \tag{3.2.1}$$

Besides, one also introduces the following matrix, already concerned with the continuous case [96]:

$$M(\varepsilon\xi) = A^{-1}(S - \varepsilon\xi I) = \frac{1}{a}\begin{pmatrix} 0 & -(1+\varepsilon\xi) \\ -a\varepsilon\xi & 0 \end{pmatrix},$$

We recall that the eigenvalues and eigenvectors of $M(\varepsilon\xi)$ can be easily found to be respectively

$$\mu_\pm(\varepsilon\xi) = \pm\sqrt{\frac{\varepsilon\xi(1+\varepsilon\xi)}{a}}, \qquad r_\pm(\varepsilon\xi) = \begin{pmatrix} 1 \\ \dfrac{a\mu_\mp(\varepsilon\xi)}{1+\varepsilon\xi} \end{pmatrix}.$$

In the above formula and all along this chapter, the complex square root is defined with the branch cut along the negative real axis. Applying Lemma 2.5.1 with the property $\operatorname{Re}\xi > 0$ and $\varepsilon > 0$, we can prove

$$\operatorname{Re}\left(\mu_-(\varepsilon\xi)\right) \leq -\frac{\varepsilon\operatorname{Re}\xi}{\sqrt{a}} < 0, \tag{3.2.2}$$

while, as a consequence,

$$\operatorname{Re}\left(\mu_+(\varepsilon\xi)\right) \geq \frac{\varepsilon\operatorname{Re}\xi}{\sqrt{a}} > 0.$$

Let us introduce

$$\kappa_\pm(\varepsilon\xi) = \mu_\pm(\varepsilon\xi)\lambda_{x\varepsilon} + \sqrt{(\mu_+(\varepsilon\xi)\lambda_{x\varepsilon})^2 + 1}, \tag{3.2.3}$$

with the notation $\lambda_{x\varepsilon} = \Delta x/\varepsilon$. According to Lemma 2.5.2 together with the properties $\mathrm{Re}\,(\mu_-(\varepsilon\xi)) < 0$ for $\varepsilon > 0$ and $\mathrm{Re}\,\xi > 0$, we can prove $|\kappa_-(\varepsilon\xi)| < 1$. Besides, since $\mu_-(\varepsilon\xi) = -\mu_+(\varepsilon\xi)$, we get $\kappa_+(\varepsilon\xi)\kappa_-(\varepsilon\xi) = 1$. As a consequence, for any $\varepsilon > 0$ and $\mathrm{Re}\,\xi > 0$, one has the separation property $|\kappa_+(\varepsilon\xi)| > 1$.

We further define the following spectral projections

$$\begin{aligned} \Phi_+(\varepsilon\xi) &= \frac{1}{2g(\varepsilon\xi)} \begin{pmatrix} 1 \\ -g(\varepsilon\xi) \end{pmatrix} \begin{pmatrix} g(\varepsilon\xi) & -1 \end{pmatrix}, \\ \Phi_-(\varepsilon\xi) &= \frac{1}{2g(\varepsilon\xi)} \begin{pmatrix} 1 \\ g(\varepsilon\xi) \end{pmatrix} \begin{pmatrix} g(\varepsilon\xi) & 1 \end{pmatrix}, \end{aligned} \tag{3.2.4}$$

where we set

$$g(\varepsilon\xi) = \frac{a\mu_+(\varepsilon\xi)}{1+\varepsilon\xi}. \tag{3.2.5}$$

We also set $\Phi(\varepsilon\xi)$ the 2×2 matrix whose columns are composed by the components of the eigenvectors of the matrix $M(\varepsilon\xi)$. We recall these matrices thus satisfy the following usefull identities

$$\Phi_+(\varepsilon\xi) = \Phi(\varepsilon\xi) \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \Phi^{-1}(\varepsilon\xi) \quad \text{and} \quad \Phi_-(\varepsilon\xi) = \Phi(\varepsilon\xi) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \Phi^{-1}(\varepsilon\xi) \tag{3.2.6}$$

and

$$\Phi_+^2(\varepsilon\xi) = \Phi_+(\varepsilon\xi), \qquad \Phi_-^2(\varepsilon\xi) = \Phi_-(\varepsilon\xi), \qquad \Phi_+(\varepsilon\xi)\Phi_-(\varepsilon\xi) = \Phi_-(\varepsilon\xi)\Phi_+(\varepsilon\xi) = 0. \tag{3.2.7}$$

In order later on to construct and estimate the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ by the $\mathcal{Z}-$transform, the following lemmas are usefull:

Lemma 3.2.1. *[From [96]] Consider $\mathbb{C}_+ = \{\zeta \in \mathbb{C},\ \mathrm{Re}\,\zeta \geq 0\}$ the closed complex right half-plane. Under the SKC (3.1.3), the quantity $g(\zeta)$ is uniformly bounded in $\mathbb{C}_+$ and the quantity $B_u + g(\zeta)B_v$ is uniformly bounded away from 0 in $\mathbb{C}_+$.*

We omit the proof that the reader can find in the work by Xin and Xu [96].

Lemma 3.2.2. *Let us consider the 4×4 matrix*

$$M_1(\varepsilon\xi) = \begin{pmatrix} 2\lambda_{x\varepsilon} M(\varepsilon\xi) & I \\ I & 0 \end{pmatrix}. \tag{3.2.8}$$

Then, the k-th power of $M_1(\varepsilon\xi)$ reads also

$$M_1^k(\varepsilon\xi) = -\frac{1}{\kappa_+(\varepsilon\xi)+\kappa_-(\varepsilon\xi)} \begin{pmatrix} \widehat{\kappa}_{k+1}(\varepsilon\xi)\widehat{\Psi}_k(\varepsilon\xi) & \widehat{\kappa}_k(\varepsilon\xi)\widehat{\Psi}_{k+1}(\varepsilon\xi) \\ \widehat{\kappa}_k(\varepsilon\xi)\widehat{\Psi}_{k+1}(\varepsilon\xi) & \widehat{\kappa}_{k-1}(\varepsilon\xi)\widehat{\Psi}_k(\varepsilon\xi) \end{pmatrix}, \tag{3.2.9}$$

where

$$\begin{aligned} \widehat{\kappa}_k(\varepsilon\xi) &= (-1)^k\kappa_+^k(\varepsilon\xi) - \kappa_-^k(\varepsilon\xi), \\ \widehat{\Psi}_k(\varepsilon\xi) &= \Phi_-(\varepsilon\xi) + (-1)^k\Phi_+(\varepsilon\xi). \end{aligned} \tag{3.2.10}$$

Proof. In this algebraic proof, we skip the dependence on $\varepsilon\xi$. Since the columns of the matrix Φ are composed by the components of the eigenvectors of the matrix M, the considered matrix M_1^k can be reformulated simply as

$$M_1^k = \widehat{\Phi} M_2^k \widehat{\Phi}^{-1}, \tag{3.2.11}$$

where

$$\widehat{\Phi} = \begin{pmatrix} \Phi & 0 \\ 0 & \Phi \end{pmatrix}, \qquad M_2 = \begin{pmatrix} D_1 & I \\ I & 0 \end{pmatrix}, \qquad D_1 = 2\lambda_{x\varepsilon} \mathrm{diag}\left(\mu_-, \mu_+\right).$$

Let Ψ is the 4×4 matrix whose columns are composed by the components of the eigenvectors of the matrix M_2

$$\Psi = \begin{pmatrix} -\kappa_+ & \kappa_- & 0 & 0 \\ 0 & 0 & -\kappa_- & \kappa_+ \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

so that $M_2^k = \Psi D_2^k \Psi^{-1}$ with $D_2 = \mathrm{diag}\left(-\kappa_+, \kappa_-, -\kappa_-, \kappa_+\right)$. Therefore, the formula M_1^k in (3.2.11) reads

$$M_1^k = \widehat{\Phi} \Psi D_2^k \Psi^{-1} \widehat{\Phi}^{-1}. \tag{3.2.12}$$

By using the properties $\Phi_\pm$ in (3.2.6) and (3.2.12), one obtains

$$M_1^k = -\frac{1}{\kappa_+ + \kappa_-} \begin{pmatrix} \widehat{\kappa}_{k+1} \widehat{\Psi}_k & \widehat{\kappa}_k \widehat{\Psi}_{k+1} \\ \widehat{\kappa}_k \widehat{\Psi}_{k+1} & \widehat{\kappa}_{k-1} \widehat{\Psi}_k \end{pmatrix}$$

with $\widehat{\kappa}_k$ and $\widehat{\Psi}_k$ are the same as in (3.2.10). □

3.2.2 Solution by $\mathcal{Z}-$transform

Firstly, we apply the $\mathcal{Z}-$transform with respect to time index $n \in \mathbb{N}$, which is discrete analogue of the Laplace transform in time $t \in \mathbb{R}_+$. This method enables the representation and estimations of the numerical solution $(U_j^n)_{j \in \mathbb{N}}$. The definition reads as follows (see [52, 75] for more details)

$$\widehat{U}_j(z) = \mathcal{Z}\{U_j^n\}(z) = \sum_{n \geq 0} U_j^n z^{-n}, \qquad |z| > 1.$$

Since we assume $(U_j^0)_{j \in \mathbb{N}} \equiv 0$, observe that the $\mathcal{Z}-$transform of the time-shifted numerical solution reads

$$\sum_{n \geq 0} U_j^{n+1} z^{-n} = z\widehat{U}_j(z) - zU_j^0 = z\widehat{U}_j(z).$$

Therefore, the IBVP (3.1.2) with zero initial data becomes

$$\begin{cases} \widehat{U}_{j+1}(z) - \widehat{U}_{j-1}(z) = 2\lambda_{x\varepsilon} M(\varepsilon\xi) \widehat{U}_j(z), & j \geq 1, \qquad (3.2.13a) \\ B\widehat{U}_0(z) = \widehat{b}(z), & \qquad (3.2.13b) \\ \Gamma A \left(\widehat{U}_1(z) - z^{-1} \Upsilon(\widehat{U}_0(z)) - 2\lambda_{x\varepsilon} M(\varepsilon\xi) \widehat{U}_0(z) \right) = 0, & \qquad (3.2.13c) \end{cases}$$

where $\Upsilon(\widehat{U}_0(z))$ is the $\mathcal{Z}-$transform of the sequence $\left\{\sum_{k=0}^{n+1} \mathcal{C}_{n+1-k} U_0^k\right\}_{n\geq 0}$ and $\widehat{b}$ stands for the $\mathcal{Z}$-transform of the scalar boundary data: $\widehat{b}(z) = \mathcal{Z}\{b^n\}(z) = \sum_{n\geq 0} b^n z^{-n}$.

Secondly, we look at the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.2.13a)-(3.2.13c). This is the object of the following proposition:

Proposition 3.2.3. *Assume that the SKC (2.1.6) is satisfied. Assume that Γ and Υ in the boundary condition (3.2.13c) are defined by*

$$\Gamma = \begin{pmatrix} -aB_v & B_u \end{pmatrix}, \quad \Upsilon(\widehat{U}_0(z)) = \kappa_+(\varepsilon\xi) z \widehat{U}_0(z). \tag{3.2.14}$$

Then the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z) \in \ell^2(\mathbb{N}, \mathbb{C}^2)$ to (3.2.13a)-(3.2.13c) takes the form

$$\widehat{U}_j(z) = \frac{\widehat{b}(z)}{B_u + g(\varepsilon\xi) B_v} \kappa_-^j(\varepsilon\xi) r_-(\varepsilon\xi). \tag{3.2.15}$$

Proof. Before we prove the above result, let us notice that we omit the explicit dependence in $\varepsilon\xi$. Firstly, we look at the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.2.13a) and consider the two-dimensional problem (3.2.13a) under the following one-step recurrence form

$$W_{j+1}(z) = M_1 W_j(z), \tag{3.2.16}$$

where M_1 is given by (3.2.8) and

$$W_j(z) = \begin{pmatrix} \widehat{U}_j(z) \\ \widehat{U}_{j-1}(z) \end{pmatrix}. \tag{3.2.17}$$

The solution $(W_j)_{j\in\mathbb{N}}(z)$ to (3.2.16) is simply $W_j(z) = M_1^j W_0(z)$. Together with the the explicit formula of M_1^j in Lemma 3.2.2, the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.2.13a) is therefore given by

$$\widehat{U}_j(z) = -\frac{1}{\kappa_+ + \kappa_-} \times \left(\widehat{\kappa}_{j+1} \widehat{\Psi}_j \widehat{U}_0(z) + \widehat{\kappa}_j \widehat{\Psi}_{j+1} \widehat{U}_{-1}(z)\right).$$

By using the definition of $\widehat{\kappa}_k$ and $\widehat{\Psi}_k$ in (3.2.10), the above formula is now equivalent to

$$\begin{aligned} \widehat{U}_j(z) = &-\frac{(-1)^j \kappa_+^j}{\kappa_+ + \kappa_-} \times \left[\Phi_-\Big(-\kappa_+ \widehat{U}_0(z) + \widehat{U}_{-1}(z)\Big) + (-1)^{j+1} \Phi_+ \Big(\kappa_+ \widehat{U}_0(z) + \widehat{U}_{-1}(z)\Big)\right] \\ &+ \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \left[\kappa_- \Big(\Phi_- + (-1)^j \Phi_+\Big) \widehat{U}_0(z) + \Big(\Phi_- + (-1)^{j+1} \Phi_+\Big) \widehat{U}_{-1}(z)\right]. \end{aligned} \tag{3.2.18}$$

Since we expect $(\widehat{U}_j)_{j\in\mathbb{N}}(z) \in \ell^2(\mathbb{N}, \mathbb{C}^2)$, we need a natural boundary condition at $x = +\infty$. Besides, one gets $|\kappa_+| > 1$ and $|\kappa_-| < 1$. Thus, the natural boundary condition takes the form

$$\begin{cases} \Phi_- \left(-\kappa_+ \widehat{U}_0(z) + \widehat{U}_{-1}(z)\right) = 0, \\ \Phi_+ \left(\kappa_+ \widehat{U}_0(z) + \widehat{U}_{-1}(z)\right) = 0. \end{cases} \tag{3.2.19}$$

By the definition of $\Phi_\pm$ in (3.2.4), the system (3.2.19) is equivalent to

$$\begin{cases} (g, 1) \left(-\kappa_+ \widehat{U}_0(z) + \widehat{U}_{-1}(z)\right) = 0, \\ (g, -1) \left(\kappa_+ \widehat{U}_0(z) + \widehat{U}_{-1}(z)\right) = 0. \end{cases}$$

Then, we have

$$\widehat{U}_{-1}(z) = \frac{\kappa_+}{g} \times \begin{pmatrix} 0 & 1 \\ g^2 & 0 \end{pmatrix} \widehat{U}_0(z).$$

Furthermore, we can see that

$$\Phi_- - \Phi_+ = \frac{1}{g} \begin{pmatrix} 0 & 1 \\ g^2 & 0 \end{pmatrix}.$$

Thus,

$$\widehat{U}_{-1}(z) = \kappa_+ (\Phi_- - \Phi_+) \widehat{U}_0(z). \tag{3.2.20}$$

Plugging (3.2.20) into (3.2.18), we have

$$\widehat{U}_j(z) = \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \left[\kappa_- \Big(\Phi_- + (-1)^j \Phi_+ \Big) + \kappa_+ \Big(\Phi_- + (-1)^{j+1} \Phi_+ \Big) \Big(\Phi_- - \Phi_+ \Big) \right] \widehat{U}_0(z).$$

Under the properties of $\Phi_\pm$ in (3.2.7), the above formula becomes

$$\widehat{U}_j(z) = \kappa_-^j \Big(\Phi_- + (-1)^j \Phi_+ \Big) \widehat{U}_0(z). \tag{3.2.21}$$

Secondly, we look at the boundary condition (3.2.13b) and (3.2.13c). Under the choice $\Upsilon(\widehat{U}_0(z))$ in (3.2.14), the boundary condition (3.2.13c) becomes

$$\Gamma A \Big(\widehat{U}_1(z) - (\kappa_+ I + 2\lambda_{x\varepsilon} M)\, \widehat{U}_0(z) \Big) = 0. \tag{3.2.22}$$

Indeed, we can compute separately

$$\begin{gathered} \kappa_+ I + 2\lambda_{x\varepsilon} M = \kappa_- \Phi_- + (\kappa_+ + 2\lambda_{x\varepsilon}\mu_+)\, \Phi_+, \\ \widehat{U}_1(z) = \kappa_- (\Phi_- - \Phi_+)\, \widehat{U}_0(z). \end{gathered} \tag{3.2.23}$$

Substituting (3.2.23) into (3.2.22), one obtains

$$\Gamma A \Phi_+ \widehat{U}_0(z) = 0. \tag{3.2.24}$$

Under the choice Γ in (3.2.14), we have

$$\Gamma A \Phi_+ = \frac{a}{2g} \times (B_u + g B_v) \times (g, -1).$$

Thus,

$$(B_u + g B_v) \times (g, -1) \widehat{U}_0(z) = 0. \tag{3.2.25}$$

From the Lemma 3.2.1, the equation (3.2.25) is equivalent under the SKC to

$$(g, -1) \widehat{U}_0(z) = 0. \tag{3.2.26}$$

Together with the boundary condition (3.2.13b), the value of $\widehat{U}_0(z)$ has to satisfy

$$\begin{pmatrix} B_u & B_v \\ g & -1 \end{pmatrix} \widehat{U}_0(z) = \widehat{b}(z) \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Then, again under the SKC, we have

$$\widehat{U}_0(z) = \frac{\widehat{b}(z)}{B_u + g(\varepsilon\xi)B_v} r_-(\varepsilon\xi). \tag{3.2.27}$$

Plugging the value of $\widehat{U}_0(z)$ in (3.2.27) into (3.2.21), the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.2.13a)-(3.2.13c) is given by

$$\widehat{U}_j(z) = \frac{\widehat{b}(z)}{B_u + gB_v} \kappa_-^j \Big(\Phi_- + (-1)^j \Phi_+ \Big) r_-(\varepsilon\xi).$$

Since $\Phi_+ r_- = 0$ and $\Phi_- r_- = r_-$, the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ finally is

$$\widehat{U}_j(z) = \frac{\widehat{b}(z)}{B_u + gB_v} \kappa_-^j r_-.$$

This ends the proof of Proposition 3.2.3. □

With $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ found in (3.2.15), the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ to the IBVP (3.1.2) with nonzero boundary condition $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N},\mathbb{R})$ and homogeneous Cauchy data $(f_j)_{j\in\mathbb{N}} \equiv 0$ can be obtained by inverting the $\mathcal{Z}$-transform [52, 75]

$$U_j^n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{U}_j \left(R\mathrm{e}^{i\theta} \right) R^n \mathrm{e}^{in\theta} d\theta, \quad R > 1.$$

Let us remark that an important assumption of Proposition 3.2.3 is $\Upsilon(\widehat{U}_0(z)) = \kappa_+(\varepsilon\xi) z \widehat{U}_0(z)$. We now follow the discrete transparent boundary as proposed in [2, 5, 54] to find the explicit formula for the sequence $(\mathcal{C}_m)_{m\geq 0}$.

Definition 3.2.4. *Let $\varepsilon > 0$, $R > 1$, $\theta \in (-\pi, \pi]$ and then $\kappa_+(\varepsilon\xi)$ be given by (3.2.3). The value of $(\mathcal{C}_m)_{m\geq 0}$ is defined as follows:*

$$\mathcal{C}_m = \frac{1}{\pi} \int_0^{\pi} \mathrm{Re} \left(\kappa_+(\varepsilon\xi) R^m \mathrm{e}^{im\theta} \right) d\theta. \tag{3.2.28}$$

Let us mention at this step that the values $\mathcal{C}_m$ above are designed in the case of homogeneous initial data and are kept unchanged in the case of nonzero initial data in forthcoming Section 3.3. Thanks to the convolution property and inverting $\mathcal{Z}-$transform, we now show that the definition of $(\mathcal{C}_m)_{m\geq 0}$ in (3.2.28) is the suitable choice to get the required identity $\Upsilon(\widehat{U}_0(z)) = \kappa_+(\varepsilon\xi) z \widehat{U}_0(z)$ in Proposition 3.2.3. This is the object of the next lemma:

Lemma 3.2.5. *Let $(\mathcal{C}_m)_{m\geq 0}$ be defined from Definition 3.2.4, then*

$$\Upsilon(\widehat{U}_0(z)) = \mathcal{Z} \left\{ \sum_{k=0}^{n+1} \mathcal{C}_{n+1-k} U_0^k \right\} (z) = \kappa_+(\varepsilon\xi) z \widehat{U}_0(z).$$

Proof. Since $\mu_+(\varepsilon\overline{\xi}) = \overline{\mu_+(\varepsilon\xi)}$, one obtains $\kappa_+(\varepsilon\overline{\xi}) = \overline{\kappa_+(\varepsilon\xi)}$. Then,

$$\begin{aligned} \mathrm{Re} \left(\kappa_+(\varepsilon\xi) R^m \mathrm{e}^{im\theta} \right) &= \frac{1}{2} \Big(\kappa_+(\varepsilon\xi) R^m \mathrm{e}^{im\theta} + \overline{\kappa_+(\varepsilon\xi)} R^m \mathrm{e}^{-im\theta} \Big) \\ &= \frac{1}{2} \Big(\kappa_+(\varepsilon\xi) R^m \mathrm{e}^{im\theta} + \kappa_+(\varepsilon\overline{\xi}) R^m \mathrm{e}^{-im\theta} \Big). \end{aligned}$$

Thus, the value of $(\mathcal{C}_m)_{m\geq 0}$ in (3.2.28) can be reformulated as

$$\begin{aligned}\mathcal{C}_m &= \frac{1}{2\pi}\left(\int_0^{\pi} \kappa_+(\varepsilon\xi)R^m \mathsf{e}^{im\theta}d\theta + \int_0^{\pi} \kappa_+(\varepsilon\bar{\xi})R^m \mathsf{e}^{-im\theta}d\theta\right)\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \kappa_+(\varepsilon\xi)R^m \mathsf{e}^{im\theta}d\theta\\ &= \mathcal{Z}^{-1}\left(\kappa_+(\varepsilon\xi)\right)(m).\end{aligned}$$

By the convolution property and inverting $\mathcal{Z}-$transform, we can conclude that

$$\Upsilon(\widehat{U}_0(z)) = \mathcal{Z}\left\{\sum_{k=0}^{n+1}\mathcal{C}_{n+1-k}U_0^k\right\}(z) = \kappa_+(\varepsilon\xi)z\widehat{U}_0(z),$$

This ends the proof of Lemma 3.2.5. □

3.2.3 Stiff stability analysis

Under the SKC, we now consider the Proposition 3.1.3 with nonzero boundary condition $(b^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N},\mathbb{R})$ and homogeneous Cauchy data $(f_j)_{j\in\mathbb{N}} \equiv 0$. In order to get the uniform estimate on $(U_j^n)_{j\in\mathbb{N}}$, firstly, we prove the following lemma:

Lemma 3.2.6. *Assume that the parameters $a, \Delta x, \Delta t > 0$ satisfy*

$$\Delta x \leq \frac{3\sqrt{a}}{8}\Delta t. \qquad (3.2.29)$$

Let $\varepsilon > 0$, $R > 1$, $\theta \in (-\pi, \pi]$ and then $\kappa_-(\varepsilon\xi)$ be given by (3.2.3). Then the following property holds

$$\sum_{j\geq 0}|\kappa_-(\varepsilon\xi)|^{2j} \leq \frac{\Delta t\sqrt{a}}{\Delta x(1-R^{-1})}. \qquad (3.2.30)$$

Proof. Since the property of $\mathrm{Re}\,(\mu_-(\varepsilon\xi))$ in (3.2.2), we can prove

$$\left(\mathrm{Re}\,\left(\mu_-(\varepsilon\xi)\right)\lambda_{x\varepsilon} + \sqrt{\left(\mathrm{Re}\,\left(\mu_-(\varepsilon\xi)\right)\lambda_{x\varepsilon}\right)^2+1}\right)^2 \leq \left(\eta\Delta x + \sqrt{\eta^2\Delta x^2+1}\right)^2, \qquad (3.2.31)$$

where $\eta = -a^{-1/2}\mathrm{Re}\,\xi$. According to Lemma 2.5.2 and the inequality (3.2.31), we have

$$|\kappa_-(\varepsilon\xi)|^2 = \left|\mu_-(\varepsilon\xi)\lambda_{x\varepsilon} + \sqrt{\left(\mu_-(\varepsilon\xi)\lambda_{x\varepsilon}\right)^2+1}\right|^2 \leq \left(\eta\Delta x + \sqrt{\eta^2\Delta x^2+1}\right)^2.$$

Then, we obtain the following estimate

$$\sum_{j\geq 0}|\kappa_-(\varepsilon\xi)|^{2j} = \left(1-|\kappa_-(\varepsilon\xi)|^2\right)^{-1} \leq \left(1-\left(\eta\Delta x + \sqrt{\eta^2\Delta x^2+1}\right)^2\right)^{-1}.$$

Since $\mathrm{Re}\,\xi$ satisfies the property (3.2.1), we get

$$\frac{\Delta t\sqrt{a}}{2} \leq -\frac{1}{\eta} \leq \frac{\Delta t\sqrt{a}}{1-R^{-1}}.$$

If we assume now $\Delta x \leq \frac{3\sqrt{a}}{8}\Delta t \leq -\frac{3}{4\eta}$ then we have

$$\left(1-\left(\eta\Delta x+\sqrt{\eta^2\Delta x^2+1}\right)^2\right)^{-1} \leq -\eta^{-1}\Delta x^{-1}.$$

Thus, we conclude that

$$\sum_{j\geq 0}|\kappa_-(\varepsilon\xi)|^{2j} \leq -\eta^{-1}\Delta x^{-1} \leq \frac{\Delta t\sqrt{a}}{\Delta x(1-R^{-1})}.$$

This ends the proof of Lemma 3.2.6. □

Secondly, by an application of the following Plancherel's theorem for $\mathcal{Z}$-transform

$$\sum_{n\geq 0} R^{-2n}|U_j^n|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}|\widehat{U}_j(Re^{i\theta})|^2 d\theta, \qquad R>1,$$

we have

$$\begin{aligned}\sum_{n\geq 0} R^{-2n}|U_0^n|^2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}|\widehat{U}_0(Re^{i\theta})|^2 d\theta\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\widehat{b}(z)}{B_u+g(\varepsilon\xi)B_v}\right|^2\left(1+|g(\varepsilon\xi)|^2\right)d\theta.\end{aligned}$$

From the Lemma 3.2.1, under the SKC, we then obtain

$$\sum_{n\geq 0} R^{-2n}|U_0^n|^2 \lesssim \frac{1}{2\pi}\int_{-\pi}^{\pi}|\widehat{b}(Re^{i\theta})|^2 d\theta \lesssim \sum_{n\geq 0} R^{-2n}|b^n|^2. \tag{3.2.32}$$

Similarly, by an application of the Plancherel's theorem for $\mathcal{Z}$-transform, we have

$$\begin{aligned}\sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}|U_j^n|^2 &= \frac{1}{2\pi}\sum_{j\geq 0}\int_{-\pi}^{\pi}|\widehat{U}_j(Re^{i\theta})|^2 d\theta\\ &= \frac{1}{2\pi}\sum_{j\geq 0}\int_{-\pi}^{\pi}\left|\frac{\widehat{b}(z)}{B_u+g(\varepsilon\xi)B_v}\right|^2|\kappa_-(\varepsilon\xi)|^{2j}\left(1+|g(\varepsilon\xi)|^2\right)d\theta.\end{aligned}$$

Again, under the SKC, we get from Lemma 3.2.1

$$\sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}|U_j^n|^2 \lesssim \frac{1}{2\pi}\sum_{j\geq 0}\int_{-\pi}^{\pi}|\widehat{b}(z)|^2|\kappa_-(\varepsilon\xi)|^{2j}d\theta. \tag{3.2.33}$$

Following Lemma 3.2.6, if we assume (3.2.29) holds, then the inequality (3.2.33) becomes

$$\frac{R-1}{R}\Delta x\sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}|U_j^n|^2 \lesssim \frac{\Delta t}{2\pi}\int_{-\pi}^{\pi}|\widehat{b}(Re^{i\theta})|^2 d\theta \lesssim \Delta t\sum_{n\geq 0} R^{-2n}|b^n|^2. \tag{3.2.34}$$

According to (3.2.32) and (3.2.34), there exists a constant $C>0$ such that

$$\frac{R-1}{R}\sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}\Delta x|U_j^n|^2 + \sum_{n\geq 0} R^{-2n}\Delta t|U_0^n|^2 \leq C\sum_{n\geq 0} R^{-2n}\Delta t|b^n|^2.$$

By setting in the above formula $R = \mathsf{e}^{\gamma\Delta t}$ for $\gamma > 0$ and $\Delta t > 0$, and using the classical lower bound $\mathsf{e}^{\gamma\Delta t} \geq 1 + \gamma\Delta t$, we obtain that there exists a constant $c > 0$ such that

$$\frac{\gamma}{\gamma\Delta t + 1}\sum_{n\geq 0}\sum_{j\geq 0}\mathsf{e}^{-2\gamma n\Delta t}\Delta t\Delta x|U_j^n|^2 + \sum_{n\geq 0}\mathsf{e}^{-2\gamma n\Delta t}\Delta t|U_0^n|^2 \leq c\sum_{n\geq 0}\mathsf{e}^{-2\gamma n\Delta t}\Delta t|b^n|^2.$$

This ends the proof of the Proposition 3.1.3.

Let us observe that the scheme (3.1.2) together also with its boundary condition is closed to be forward-in-time, except it is one-step implicit. By this property, changing the data b to zero after some time T and unchanged before that time T, the discrete solution U_j^n is the same for $n\Delta t < T$. Therefore, there exists a constant $C_T > 0$ such that

$$\sum_{n=0}^{N}\sum_{j\geq 0}\Delta x\Delta t\left|U_j^n\right|^2 + \sum_{n=0}^{N}\Delta t\left|U_0^n\right|^2 \leq C_T\sum_{n=0}^{N}\Delta t\left|b^n\right|^2, \tag{3.2.35}$$

with $N := T/\Delta t$. This will be useful to prove the Theorem 3.1.1.

3.2.4 Numerical experiments

In this paragraph, we first provide the behavior of the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ according to whether or not the SKC (3.1.3) is valid. We also look at the degenerate case when the UKC (2.1.5) does not hold (and thus, none of the other stability conditions). Following the continuous case studied by Xin and Xu in [96], the solution of the IBVP (2.1.1)-(2.1.3) with homogeneous initial condition can be constructed by the method of Laplace transform. By inverting the Laplace transform, the solution $U(x,t)$ has form

$$U(x,t) = \mathcal{L}^{-1}\widetilde{U} = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\mathsf{e}^{\zeta t}\frac{\widetilde{b}(\zeta)}{B_u + g(\varepsilon\zeta)B_v}\mathsf{e}^{\mu_-(\varepsilon\zeta)x/\varepsilon}r_-(\varepsilon\zeta)d\beta,$$

where $\zeta = \alpha + i\beta$, $\alpha > 0$. We make use the function *mpmath.invertlaplace* in Python to compute the inverse Laplace transform for the exact solution $U(x,t)$. Then, we observe the error between the exact solution $U(x_j,t^n)$ and the numerical solution U_j^n of the numerical scheme (3.1.2) with homogeneous initial data at the grid point $(x_j,t^n) = (j\Delta x, n\Delta t)$. After that, we present some numerical experiments and observe the effective behavior of the energy terms $\|U\|_{\ell^2(\mathbb{N}\times[0,T),\mathbb{R}^2)}$ and $\|U\|_{\ell^2(\{0\}\times[0,T),\mathbb{R}^2)}$ corresponding to whether or not the SKC (3.1.3) is valid.

As main parameters for the experiments, we choose $a = 1$, $B_v = 1$, $\lambda_{xt} = 1/3$ and let the relaxation rate ε and the boundary data B_u vary. The test case we consider concerns the following data. The initial data is the homogeneous one $(f_j)_{j\in\mathbb{N}} \equiv 0$. The boundary data is

$$b(t) = \frac{t}{2}\sin(t).$$

Let us observe that these data are compatible in the corner $(x,t) = (0,0)$ in the sense that $Bf(0) = b(0)$. Moreover, the Laplace transform of $b(t)$ is

$$\widetilde{b}(\zeta) = \frac{\zeta}{(\zeta^2+1)^2}.$$

3.2.4.1 The behavior of the numerical solution

Let the space step $\Delta x = 10^{-2}$ and the time step $\Delta t = \lambda_{xt}^{-1}\Delta x$. Firstly, we choose the value of B_u such that the SKC (3.1.3) is satisfied with $\varepsilon = 10^{-2}$ and also with $\varepsilon = 10^2$. The Figures 3.2.1 and 3.2.2 show the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ over the time interval $t \in [0, 1.2)$.

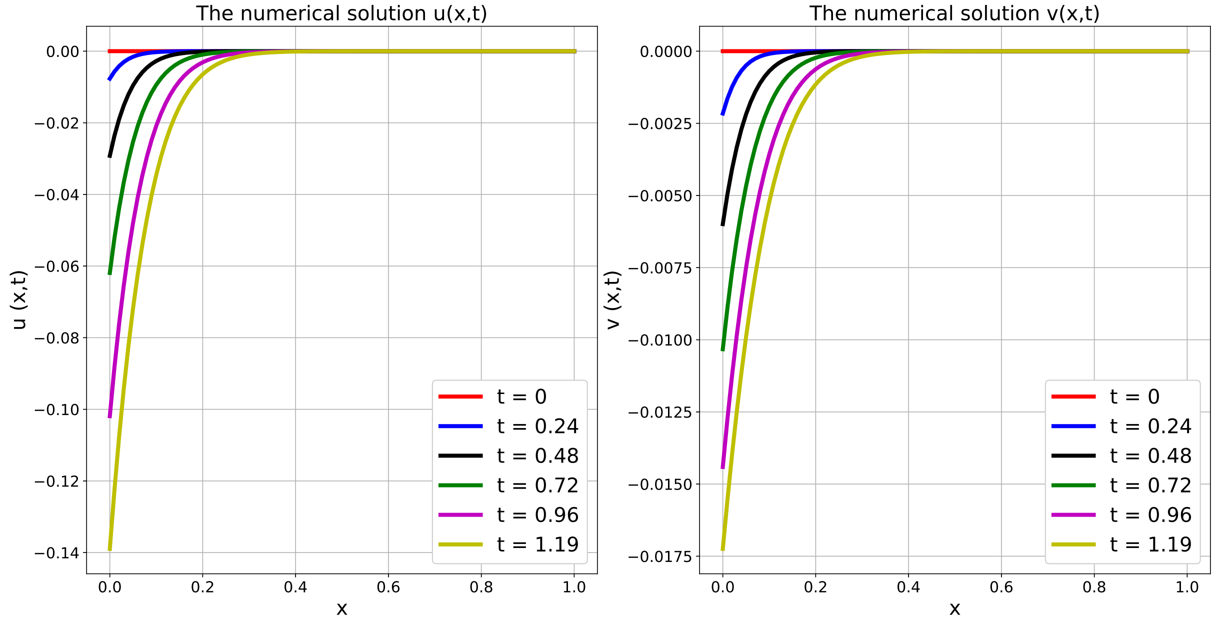

Figure 3.2.1: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^{-2}$. The SKC (3.1.3) holds with $B_u = -4$.

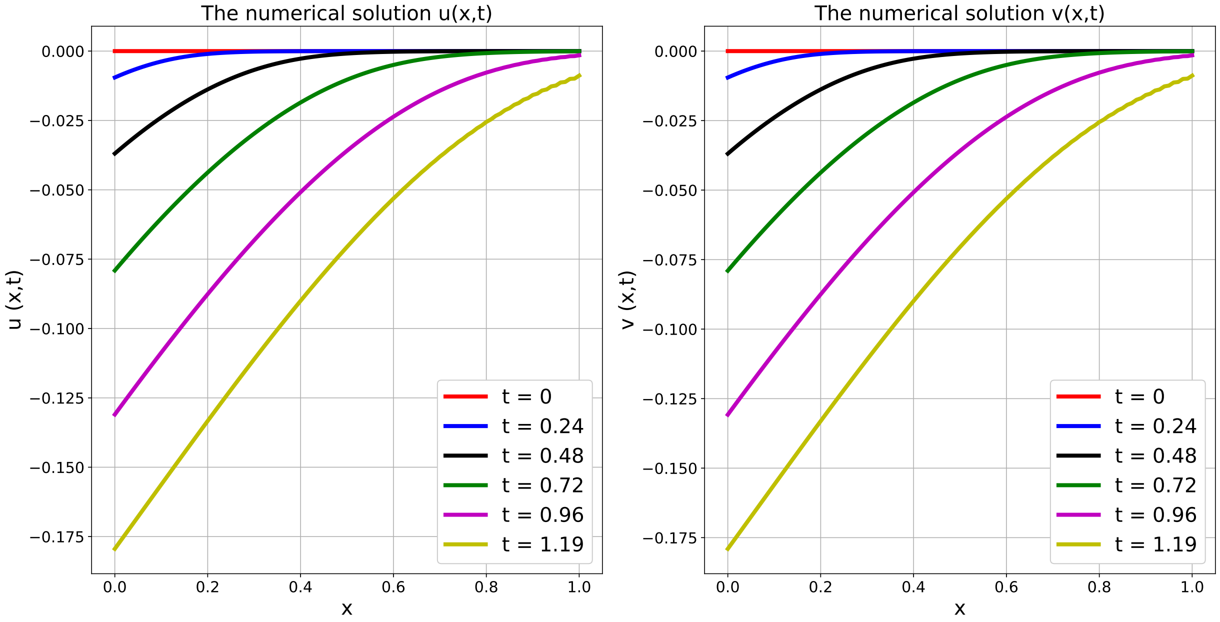

Figure 3.2.2: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^{2}$. The SKC (3.1.3) holds with $B_u = -4$.

In the first case, $\varepsilon = 10^{-2}$, the incoming solution at the boundary $x = 0$ go slowly. This is due to the initial relaxation of solution to the equilibrium system. In the case $\varepsilon = 10^{2}$, its solution seems to be faster. It is not so much influenced by relaxation source term but more by the boundary dissipation.

Secondly, we choose the value of B_u such that the SKC (3.1.3) is not satisfied. Besides, we also present the numerical solution when the Uniform Kreiss Condition (2.1.5) is wrong. The Figures 3.2.3 and 3.2.4 show the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ over the time interval $t \in [0, 0.5)$.

When the SKC (3.1.3) fails, we observe that the numerical solution at the boundary rise gradually. This is the case for example for $\varepsilon = 10^{-2}$ together with the parameters $(B_u, B_v) = (-1/2, 1)$. The behavior is even worse when the UKC (2.1.5) is not satisfied (see Figure 3.2.4).

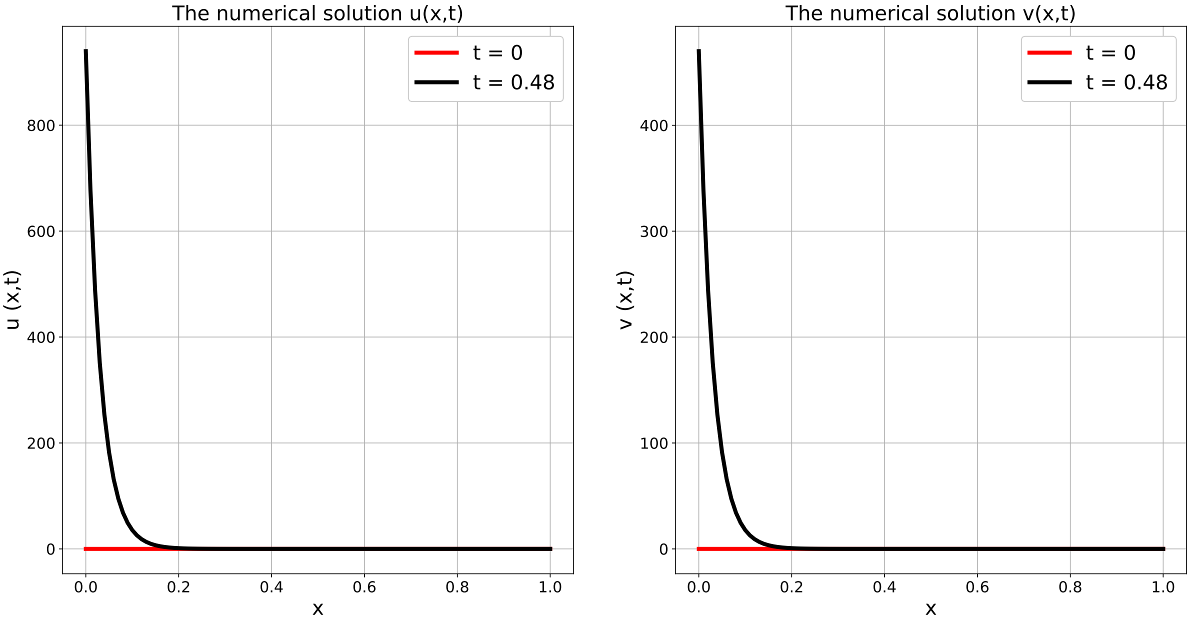

Figure 3.2.3: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^{-2}$. The SKC (3.1.3) does not hold with $B_u = -1/2$.

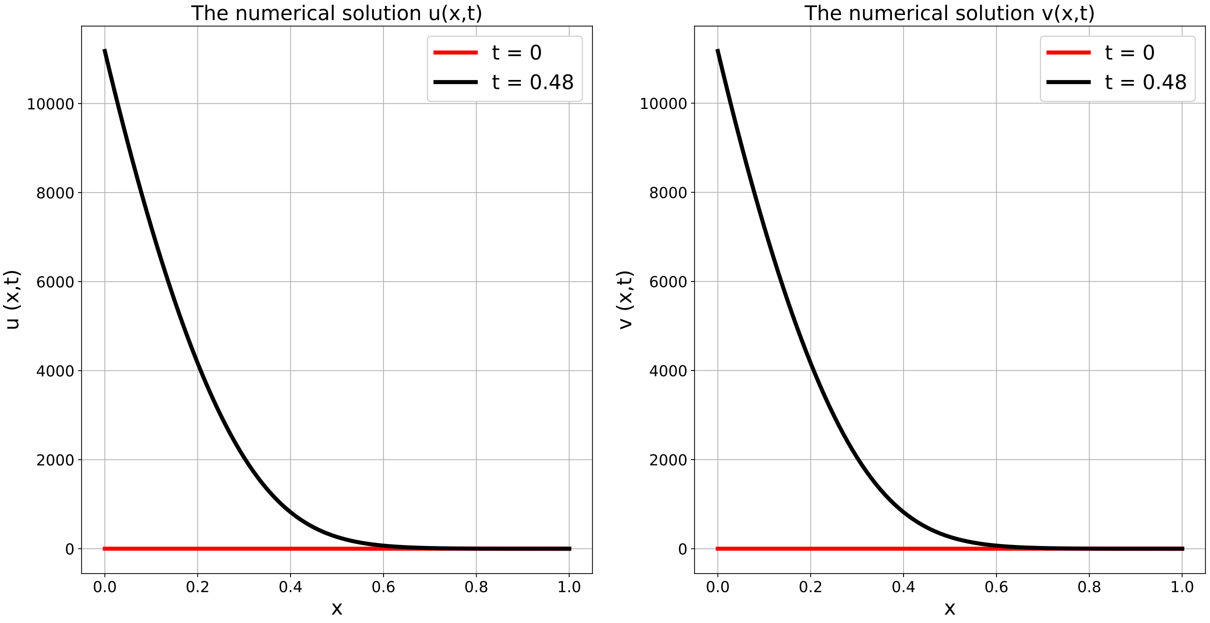

Figure 3.2.4: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^{2}$. The UKC (2.1.5) is wrong with $B_u = -1$.

3.2.4.2 The error between the exact solution and the numerical solution

Let us begin with the notation

$$E(t^n) := \left(\Delta x \sum_{j \geq 0} |U(x_j, t^n) - U_j^n|^2 \right)^{1/2}. \tag{3.2.36}$$

We choose a set of values B_u such that the SKC (3.1.3) is satisfied with the space step Δx and the relation rate ε vary. The error, as measured in (3.2.36), are reported in the Tables 3.1 and 3.2.

According to the experiments in Tables 3.1 and 3.2, for some $\varepsilon \in (0, +\infty)$ and (B_u, B_v)

Δx	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-1}$	$\varepsilon = 1$	$\varepsilon = 10$	$\varepsilon = 10^{2}$
5×10^{-2}	6.8×10^{-3}	1.2×10^{-2}	2.6×10^{-2}	3.1×10^{-2}	3.2×10^{-2}
25×10^{-3}	3×10^{-3}	5.9×10^{-3}	1.3×10^{-2}	1.6×10^{-2}	1.6×10^{-2}
125×10^{-4}	1.5×10^{-3}	2.9×10^{-3}	6.8×10^{-3}	8.2×10^{-3}	8.3×10^{-3}
625×10^{-5}	7.2×10^{-4}	1.5×10^{-3}	3.4×10^{-3}	4.1×10^{-3}	4.2×10^{-3}

Table 3.1: The error $E(1.2)$ for $B_u = -4$.

Δx	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-1}$	$\varepsilon = 1$	$\varepsilon = 10$	$\varepsilon = 10^{2}$
5×10^{-2}	8.5×10^{-3}	1.3×10^{-2}	2.1×10^{-2}	2.3×10^{-2}	2.4×10^{-2}
25×10^{-3}	3.8×10^{-3}	6.4×10^{-3}	1.1×10^{-2}	1.2×10^{-2}	1.2×10^{-2}
125×10^{-4}	1.8×10^{-3}	3.2×10^{-3}	5.4×10^{-3}	6.1×10^{-3}	6.2×10^{-3}
625×10^{-5}	9.1×10^{-4}	1.6×10^{-3}	2.7×10^{-3}	3.1×10^{-3}	3.1×10^{-3}

Table 3.2: The error $E(1.2)$ for $B_u = 3$.

satisfying the SKC (3.1.3), the observed convergence rate is 1 since going down Δx by a factor 2 decreases the error of the same factor 2. It means that the behavior of the numerical solution U_j^n is the same as the evolution of the exact solution $U(x_j, t^n)$. This is the case for example for $\varepsilon = 10^{-2}$ together with the parameters $(B_u, B_v) = (-4, 1)$.

3.2.4.3 The effective behavior of the energy terms

Let the space step $\Delta x = 10^{-2}$ and the time step $\Delta t = \lambda_{xt}^{-1} \Delta x$. We present hereafter the behavior of the following energy terms for $\varepsilon \in (0, +\infty)$, $T = 1.2$ and $N = T/\Delta t$. The first one corresponds to the ℓ^2 in time and space energy of the discrete solution and the second to the ℓ^2 in time energy of the numerical trace at the boundary:

$$\begin{aligned} E_1 &:= \|U\|^2_{\ell^2(\mathbb{N}\times[0,T),\mathbb{R}^2)} = \sum_{n=0}^{N} \sum_{j\geq 0} \Delta x \Delta t |U_j^n|^2, \\ E_2 &:= \|U\|^2_{\ell^2(\{0\}\times[0,T),\mathbb{R}^2)} = \sum_{n=0}^{N} \Delta t |U_0^n|^2, \end{aligned} \tag{3.2.37}$$

which are shown in the Table 3.3 and Figures 3.2.5, 3.2.6.

B_u	$\varepsilon = 10^{-2}$	$\varepsilon = 1$	$\varepsilon = 10^{2}$
-4	35×10^{-5}	77×10^{-5}	47×10^{-4}
-2	15×10^{-4}	39×10^{-4}	43×10^{-3}
-1	2.6×10^{17}	1.69×10^{31}	3.17×10^{56}
-0.5	44047.9	418525.12	2837033.2
1	34×10^{-4}	5×10^{-3}	10^{-2}
3	48×10^{-5}	87×10^{-5}	2×10^{-2}

B_u	$\varepsilon = 10^{-2}$	$\varepsilon = 1$	$\varepsilon = 10^{2}$
-4	66×10^{-4}	8×10^{-3}	2×10^{-2}
-2	29×10^{-3}	4×10^{-2}	18×10^{-2}
-1	1.24×10^{19}	7.17×10^{32}	1.89×10^{58}
-0.5	191420.5	9910910.98	106714237.85
1	46×10^{-3}	55×10^{-3}	67×10^{-3}
3	93×10^{-4}	94×10^{-4}	12×10^{-3}

Table 3.3: The energy terms E_1 (left) and E_2(right).

- For some $\varepsilon \in (0, +\infty)$, the values of E_1 and E_2 rise gradually when the SKC (3.1.3) is not satisfied. This is the case for example for $\varepsilon = 10^2$ together with the parameters $(B_u, B_v) = (-1/2, 1)$. The behavior is even worse when the UKC (2.1.5) is not hold.
- In the case $\varepsilon = 10^{-2}$, the energy term E_1 and E_2 increase slowly. This is due to the effect of incoming solution at the boundary when the initial relaxation of solution tends to the

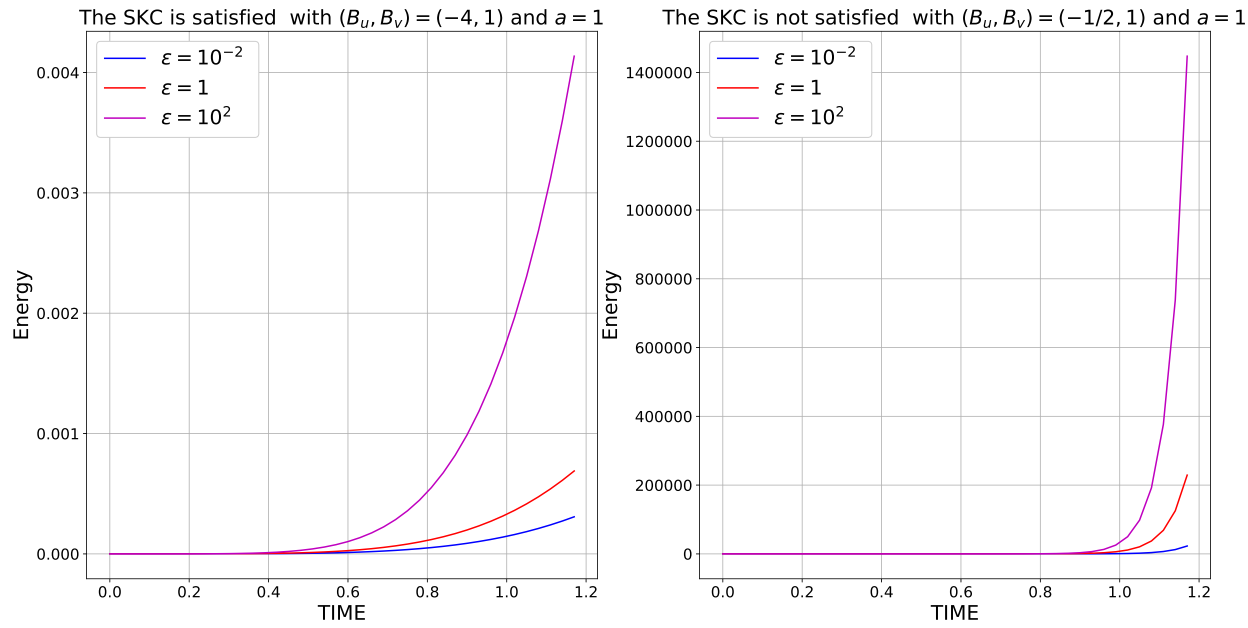

Figure 3.2.5: Energy evolution E_1 for $B_u = -4$ (left) and $B_u = -0.5$ (right).

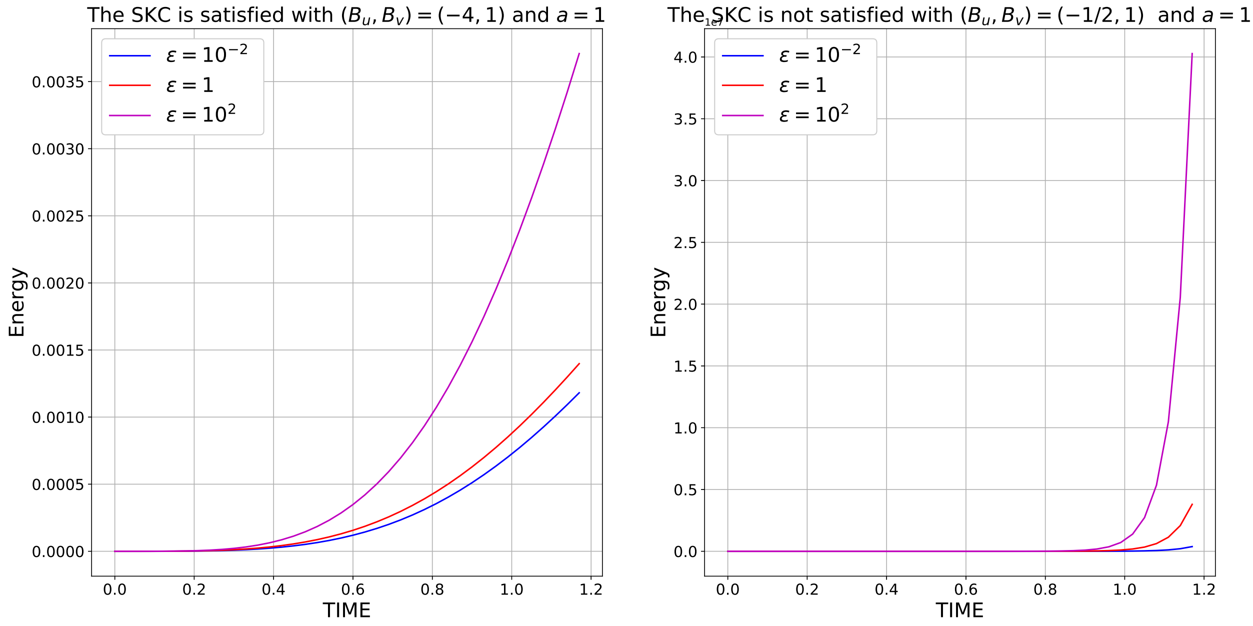

Figure 3.2.6: Energy evolution E_2 for $B_u = -4$ (left) and $B_u = -0.5$ (right).

equilibrium system. In the case $\varepsilon = 10^2$, those values increase fairly rapidly. It is not so much influenced by relaxation source term but more by the boundary dissipation.

Clearly, the numerical results show that the numerical solution at the boundary $x = 0$ increase quickly as soon as the SKC (3.1.3) does not hold. If the UKC (2.1.5) is not satisfied, the behavior of numerical solution is even worse. Besides, by the decrease of the error $E(t^n)$, we can see that the values of U_j^n tend to the exact solution $U(x_j, t^n)$. Indeed, it seems that the SKC (3.1.3) is also necessary to ensure the non-increase rapidly of the energy terms E_1 and E_2 under the effect of the relaxation source term and the boundary dissipation.

3.3 Stiff stability of the IBVP with homogeneous boundary condition

For convenience in the forthcoming discussions, we recall that the IBVP (3.1.2) with homogeneous boundary condition writes

$$\begin{cases} \dfrac{U_j^{n+1}-U_j^n}{\Delta t}+\dfrac{1}{2\Delta x}A\left(U_{j+1}^{n+1}-U_{j-1}^{n+1}\right)=\dfrac{1}{\varepsilon}SU_j^{n+1}, & j\geq 1, n\geq 0,\\ U_j^0=f_j, & j\geq 0,\\ BU_0^n=0, & n\geq 0,\\ \dfrac{1}{\Delta t}\Gamma\left(U_0^{n+1}-U_0^n\right)+\dfrac{1}{2\Delta x}\Gamma A\left(U_1^{n+1}-\displaystyle\sum_{k=0}^{n+1}\mathcal{C}_{n+1-k}U_0^k\right)=\dfrac{1}{\varepsilon}\Gamma SU_0^{n+1}, & n\geq 0.\end{cases} \tag{3.3.1}$$

In [96, Section 5], under the SKC, Xin and Xu find explicitly the solution $U(x,t)$ of the IBVP (2.1.1)-(2.1.3) with homogeneous boundary data by the method of Laplace transform. The solution is decomposed into two ingredients, by assembling a solution for the case of the Cauchy problem and another one for the case of the IBVP with homogeneous initial condition. In our case, assuming the SKC to hold, the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ can be constructed by the method of $\mathcal{Z}-$transform. Since the coefficients $(\mathcal{C}_m)_{m\geq 0}$ are defined for homogeneous initial data, the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ of the IBVP (3.3.1) consists of not only the solutions for case of the Cauchy problem and for the IBVP with zero initial data but also another numerical error term $(U_j^{II})_{j\in\mathbb{N}}^n$. To complete the proof of the Theorem 3.1.1 with homogeneous boundary condition, we first use the means of discrete energy method in order to prove the Proposition 3.1.2. By an application of the Plancherel's theorem for $\mathcal{Z}$-transform [52, 75], the numerical error term of $(U_j^{II})_{j\in\mathbb{N}}^n$ will be estimated in Section 3.3.3. After that, we get the expected result of the case IBVP with homogeneous initial condition.

3.3.1 Solution by $\mathcal{Z}-$transform

Again, we follow the explicit solving of the IBVP (3.3.1) by using the $\mathcal{Z}-$transform. With

$$\widehat{U}_j(z)=\mathcal{Z}\{U_j^n\}(z)=\sum_{n\geq 0}U_j^n z^{-n}, \qquad |z|>1.$$

Importantly, we now have $(f_j)_{j\in\mathbb{N}}\neq 0$, and thus we get

$$\sum_{n\geq 0}U_j^{n+1}z^{-n}=z\widehat{U}_j(z)-zU_j^0=z\widehat{U}_j(z)-zf_j.$$

Therefore, (3.3.1) becomes

$$\begin{cases} \widehat{U}_{j+1}(z)-\widehat{U}_{j-1}(z)=2\lambda_{x\varepsilon}M(\varepsilon\xi)\widehat{U}_j(z)+f_{j+1}-2\lambda_{x\varepsilon}M(\varepsilon\Delta t^{-1})f_j-f_{j-1},\ j\geq 1, & \text{(3.3.2a)}\\ B\widehat{U}_0(z)=0, & \text{(3.3.2b)}\\ \Gamma A\left[\widehat{U}_1(z)-\Big(\kappa_+(\varepsilon\xi)I+2\lambda_{x\varepsilon}M(\varepsilon\xi)\Big)\widehat{U}_0(z)-f_1\right]=0. & \text{(3.3.2c)}\end{cases}$$

Let us recall that the $\mathcal{Z}-$transform of $\sum_{k=0}^{n+1}\mathcal{C}_{n+1-k}U_0^k$ is given by $\kappa_+(\varepsilon\xi)z\widehat{U}_0(z)$. Firstly, we look at the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.3.2a)-(3.3.2c). This is the object of the following proposition:

Proposition 3.3.1. *Assume that the SKC (3.1.3) is satisfied. Let* $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N},\mathbb{R}^2)$ *and denote* V_k^I, $w^I(\varepsilon\xi)$ *and* $w^{II}(\varepsilon\xi)$ *as follows:*

$$\begin{aligned} V_k^I &= f_{k+1} - 2\lambda_{x\varepsilon} M(\varepsilon\Delta t^{-1}) f_k - f_{k-1}, \\ w^I(\varepsilon\xi) &= \sum_{k=0}^{+\infty} (-1)^{-k}\kappa_+^{-k}(\varepsilon\xi) V_k^I, \\ w^{II}(\varepsilon\xi) &= -\sum_{k=0}^{+\infty} \kappa_+^{-k}(\varepsilon\xi) V_k^I. \end{aligned} \tag{3.3.3}$$

Then, the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z) \in \ell^2(\mathbb{N},\mathbb{C}^2)$ *to (3.3.2a)-(3.3.2c) takes the form*

$$\begin{aligned} \widehat{U}_j(z) = {} & \frac{\kappa_-^{j+1}(\varepsilon\xi)}{4g(\varepsilon\xi)} \times \left(\frac{B_u - g(\varepsilon\xi)B_v}{B_u + g(\varepsilon\xi)B_v} \times (g(\varepsilon\xi),1)\, w^I(\varepsilon\xi) - (g(\varepsilon\xi),-1)\, w^{II}(\varepsilon\xi) \right) \\ & \times \left(\frac{B_u - g(\varepsilon\xi)B_v}{B_u + g(\varepsilon\xi)B_v} \times r_-(\varepsilon\xi) + (-1)^{j+1} r_+(\varepsilon\xi) \right) \\ & - \frac{\kappa_-^j(\varepsilon\xi)}{\kappa_+(\varepsilon\xi) + \kappa_-(\varepsilon\xi)} \times \Big(\Phi_-(\varepsilon\xi) w^I(\varepsilon\xi) + (-1)^j \Phi_+(\varepsilon\xi) w^{II}(\varepsilon\xi) \Big) \\ & + \frac{1}{\kappa_+(\varepsilon\xi) + \kappa_-(\varepsilon\xi)} \times \Bigg(\sum_{k=0}^{j-1} \kappa_-^{j-k}(\varepsilon\xi) \Big(\Phi_-(\varepsilon\xi) + (-1)^{j-k-1}\Phi_+(\varepsilon\xi) \Big) V_k^I \\ & \qquad + \sum_{k=j}^{+\infty} (-1)^{j-k} \kappa_+^{j-k}(\varepsilon\xi) \Big(\Phi_-(\varepsilon\xi) + (-1)^{j-k-1}\Phi_+(\varepsilon\xi) \Big) V_k^I \Bigg). \end{aligned} \tag{3.3.4}$$

Proof. Before we prove the above result, let us notice that we omit the explicit dependence in $\varepsilon\xi$. Firstly, we look at the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.3.2a) and consider the two-dimensional problem (3.3.2a) under the following nonhomogeneous one-step recurrence form

$$W_{j+1}(z) = M_1 W_j(z) + V_j, \tag{3.3.5}$$

where W_j is the same as in (3.2.17) and

$$V_j = \begin{pmatrix} V_j^I \\ 0 \end{pmatrix}.$$

The solution $(W_j)_{j\in\mathbb{N}}(z)$ to (3.3.5) is thus given by

$$W_j(z) = M_1^j W_0(z) + \sum_{k=0}^{j-1} M_1^{j-1-k} V_k. \tag{3.3.6}$$

Together with the the explicit formula of M_1^j in Lemma 3.2.2, the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.3.2a) is given by

$$\widehat{U}_j(z) = -\frac{1}{\kappa_+ + \kappa_-} \times \left[\widehat{\kappa}_{j+1}\widehat{\Psi}_j \widehat{U}_0(z) + \widehat{\kappa}_j \widehat{\Psi}_{j+1} \widehat{U}_{-1} + \sum_{k=0}^{j-1} \widehat{\kappa}_{j-k} \widehat{\Psi}_{j-k-1} V_k^I \right].$$

By the definition of $\widehat{\kappa}_k$ and $\widehat{\Psi}_k$ in (3.2.10), the above formula is equivalent to

$$\begin{aligned}\widehat{U}_j(z) = -\frac{(-1)^j\kappa_+^j}{\kappa_+ + \kappa_-} \times \Bigg[&-\kappa_+\Big(\Phi_- + (-1)^j\Phi_+\Big)\widehat{U}_0(z) + \Big(\Phi_- + (-1)^{j+1}\Phi_+\Big)\widehat{U}_{-1}(z) \\ &+ \sum_{k=0}^{j-1}(-\kappa_+)^{-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I\Bigg] \\ + \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \Bigg[&\kappa_-\Big(\Phi_- + (-1)^j\Phi_+\Big)\widehat{U}_0(z) + \Big(\Phi_- + (-1)^{j+1}\Phi_+\Big)\widehat{U}_{-1}(z) \\ &+ \sum_{k=0}^{j-1}\kappa_-^{-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I\Bigg].\end{aligned} \tag{3.3.7}$$

Thanks to the definition of w^I and w^{II} in (3.3.3), one has

$$\begin{aligned}&\sum_{k=0}^{j-1}(-\kappa_+)^{-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I \\ &\quad = \Phi_- w^I + (-1)^j\Phi_+ w^{II} - \sum_{k=j}^{+\infty}(-\kappa_+)^{-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I.\end{aligned} \tag{3.3.8}$$

Substituting (3.3.8) into (3.3.7), we have

$$\begin{aligned}\widehat{U}_j(z) = &-\frac{(-1)^j\kappa_+^j}{\kappa_+ + \kappa_-} \times \Bigg[\Phi_-\Big(-\kappa_+\widehat{U}_0(z) + \widehat{U}_{-1}(z) + w^I\Big) + (-1)^{j+1}\Phi_+\Big(\kappa_+\widehat{U}_0(z) + \widehat{U}_{-1}(z) - w^{II}\Big)\Bigg] \\ &+ \frac{1}{\kappa_+ + \kappa_-} \times \sum_{k=j}^{+\infty}(-1)^{j-k}\kappa_+^{j-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I \\ &+ \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \Bigg[\kappa_-\Big(\Phi_- + (-1)^j\Phi_+\Big)\widehat{U}_0(z) + \Big(\Phi_- + (-1)^{j+1}\Phi_+\Big)\widehat{U}_{-1}(z) \\ &\qquad + \sum_{k=0}^{j-1}\kappa_-^{-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I\Bigg].\end{aligned} \tag{3.3.9}$$

Since we expect $(\widehat{U}_j)_{j\in\mathbb{N}}(z) \in \ell^2(\mathbb{N}, \mathbb{C}^2)$, we need a natural boundary condition at $x = +\infty$. Besides, one gets $|\kappa_+| > 1$ and $|\kappa_-| < 1$. Thus, the natural boundary condition takes the form

$$\begin{cases}\Phi_-\Big(-\kappa_+\widehat{U}_0(z) + \widehat{U}_{-1}(z) + w^I\Big) = 0, \\ \Phi_+\Big(\kappa_+\widehat{U}_0(z) + \widehat{U}_{-1}(z) - w^{II}\Big) = 0.\end{cases} \tag{3.3.10}$$

By the definition of $\Phi_\pm$ in (3.2.4), the system (3.3.10) is equivalent to

$$\begin{cases}(g, 1)\Big(-\kappa_+\widehat{U}_0(z) + \widehat{U}_{-1}(z) + w^I\Big) = 0, \\ (g, -1)\Big(\kappa_+\widehat{U}_0(z) + \widehat{U}_{-1}(z) - w^{II}\Big) = 0.\end{cases}$$

Then, we have

$$\widehat{U}_{-1}(z) = \kappa_+(\Phi_- - \Phi_+)\widehat{U}_0(z) - \Phi_- w^I + \Phi_+ w^{II}. \tag{3.3.11}$$

Plugging (3.3.11) into (3.3.9), we get

$$\begin{aligned}\widehat{U}_j(z) = &\frac{1}{\kappa_+ + \kappa_-} \times \sum_{k=j}^{+\infty} (-1)^{j-k}\kappa_+^{j-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I \\ &+ \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \left[\kappa_-\Big(\Phi_- + (-1)^j\Phi_+\Big) + \kappa_+\Big(\Phi_- + (-1)^{j+1}\Phi_+\Big)\Big(\Phi_- - \Phi_+\Big)\right]\widehat{U}_0(z) \\ &+ \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \Big(\Phi_- + (-1)^{j+1}\Phi_+\Big)\Big(-\Phi_- w^I + \Phi_+ w^{II}\Big) \\ &+ \frac{1}{\kappa_+ + \kappa_-} \times \sum_{k=0}^{j-1} \kappa_-^{j-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I.\end{aligned}$$

Under the properties of $\Phi_\pm$ in (3.2.7), the above formula becomes

$$\begin{aligned}\widehat{U}_j(z) = &\kappa_-^j\Big(\Phi_- + (-1)^j\Phi_+\Big)\widehat{U}_0(z) - \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \Big(\Phi_- w^I + (-1)^j\Phi_+ w^{II}\Big) \\ &+ \frac{1}{\kappa_+ + \kappa_-} \times \left[\sum_{k=0}^{j-1} \kappa_-^{j-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I + \sum_{k=j}^{+\infty}(-1)^{j-k}\kappa_+^{j-k}\Big(\Phi_- + (-1)^{j-k-1}\Phi_+\Big)V_k^I\right].\end{aligned} \tag{3.3.12}$$

Secondly, we look at the boundary data $\widehat{U}_0(z)$ and extend the initial data $(f_j)_{j\in\mathbb{N}}$ to the whole line by setting $f_j = 0$ for $j \leq 0$. Since $f_0 = f_{-1} = 0$ and $\Phi_+ + \Phi_- = I$, we can see that

$$\begin{aligned}\widehat{U}_1(z) = &\kappa_-\Big(\Phi_- - \Phi_+\Big)\widehat{U}_0(z) - \frac{\kappa_-}{\kappa_+ + \kappa_-} \times \Big(\Phi_- w^I - \Phi_+ w^{II}\Big) \\ &+ \frac{1}{\kappa_+ + \kappa_-} \times \left[\kappa_- f_1 - \kappa_+ \sum_{k=1}^{+\infty}(-1)^{-k}\kappa_+^{-k}\Big(\Phi_- + (-1)^{-k}\Phi_+\Big)V_k^I\right].\end{aligned} \tag{3.3.13}$$

On the other hand, by the definition of w^I and w^{II} in (3.3.3), we get the following property

$$\sum_{k=1}^{+\infty}(-1)^{-k}\kappa_+^{-k}\Big(\Phi_- + (-1)^{-k}\Phi_+\Big)V_k^I = \Phi_- w^I - \Phi_+ w^{II} - f_1. \tag{3.3.14}$$

Substituting (3.3.14) into (3.3.13), the value of $\widehat{U}_1(z)$ can be reformulated as

$$\widehat{U}_1(z) = \kappa_-(\Phi_- - \Phi_+)\widehat{U}_0(z) - \Phi_- w^I + \Phi_+ w^{II} + f_1.$$

Thus, the equation (3.3.2c) becomes

$$\Gamma A\Big(\kappa_-(\Phi_- - \Phi_+)\widehat{U}_0(z) - \Phi_- w^I + \Phi_+ w^{II} - (\kappa_+ I + 2\lambda_{x\varepsilon}M)\,\widehat{U}_0(z)\Big) = 0. \tag{3.3.15}$$

Indeed, we observe that

$$\kappa_+ I + 2\lambda_{x\varepsilon}M = \kappa_-\Phi_- + (\kappa_+ + 2\lambda_{x\varepsilon}\mu_+)\Phi_+.$$

Then, the equation (3.3.15) can be represented as

$$\Gamma A\Phi_+\widehat{U}_0(z) = -\frac{1}{2\kappa_+}\Gamma A\Big(\Phi_- w^I - \Phi_+ w^{II}\Big). \tag{3.3.16}$$

Besides, one has

$$\Gamma A \Phi_+ = \frac{a}{2g} \times (B_u + gB_v) \times (g, -1).$$

From the Lemma 3.2.1, the equation (3.3.16) is equivalent under the SKC to

$$(g,-1)\widehat{U}_0(z) = -\frac{g}{a\kappa_+(B_u+gB_v)} \Gamma A \Big(\Phi_- w^I - \Phi_+ w^{II} \Big).$$

Together with the boundary condition (3.3.2b), one gets

$$\begin{pmatrix} B_u & B_v \\ g & -1 \end{pmatrix} \widehat{U}_0(z) = -\frac{g}{a\kappa_+(B_u+gB_v)} \begin{pmatrix} 0 \\ \Gamma A \Big(\Phi_- w^I - \Phi_+ w^{II} \Big) \end{pmatrix}.$$

Then, under the SKC, we have

$$\widehat{U}_0(z) = -\frac{g}{a\kappa_+(B_u+gB_v)^2} \times \Gamma A \Big(\Phi_- w^I - \Phi_+ w^{II} \Big) \begin{pmatrix} B_v \\ -B_u \end{pmatrix}. \tag{3.3.17}$$

Substituting (3.3.17) into (3.3.12), the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.3.2a)-(3.3.2c) is given by

$$\begin{aligned}
\widehat{U}_j(z) = &-\frac{\kappa_-^{j+1} g}{a(B_u+gB_v)^2} \times \Gamma A \Big(\Phi_- w^I - \Phi_+ w^{II} \Big) \times \Big(\Phi_- + (-1)^j \Phi_+ \Big) \begin{pmatrix} B_v \\ -B_u \end{pmatrix} \\
&- \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \Big(\Phi_- w^I + (-1)^j \Phi_+ w^{II} \Big) \\
&+ \frac{1}{\kappa_+ + \kappa_-} \times \left[\sum_{k=0}^{j-1} \kappa_-^{j-k} \Big(\Phi_- + (-1)^{j-k-1} \Phi_+ \Big) V_k^I + \sum_{k=j}^{+\infty} (-1)^{j-k} \kappa_+^{j-k} \Big(\Phi_- + (-1)^{j-k-1} \Phi_+ \Big) V_k^I \right].
\end{aligned}$$

Together with the definitions of $\Phi_\pm$ and Γ in (3.2.6) and (3.2.14), respectively, we have

$$\begin{aligned}
&-\frac{g}{a(B_u+gB_v)^2} \times \Gamma A \Big(\Phi_- w^I - \Phi_+ w^{II} \Big) \times \Big(\Phi_- + (-1)^j \Phi_+ \Big) \begin{pmatrix} B_v \\ -B_u \end{pmatrix} \\
&= \frac{1}{4g} \times \left(\frac{B_u - gB_v}{B_u + gB_v} \times (g,1) w^I - (g,-1) w^{II} \right) \times \left(\frac{B_u - gB_v}{B_u + gB_v} \times r_- + (-1)^{j+1} r_+ \right).
\end{aligned}$$

Therefore, the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.3.2a)-(3.3.2c) can be reformulated as

$$\begin{aligned}
\widehat{U}_j(z) = &= \frac{\kappa_-^{j+1}}{4g} \times \left(\frac{B_u - gB_v}{B_u + gB_v} \times (g,1) w^I - (g,-1) w^{II} \right) \times \left(\frac{B_u - gB_v}{B_u + gB_v} \times r_- + (-1)^{j+1} r_+ \right) \\
&- \frac{\kappa_-^j}{\kappa_+ + \kappa_-} \times \Big(\Phi_- w^I + (-1)^j \Phi_+ w^{II} \Big) \\
&+ \frac{1}{\kappa_+ + \kappa_-} \times \left[\sum_{k=0}^{j-1} \kappa_-^{j-k} \Big(\Phi_- + (-1)^{j-k-1} \Phi_+ \Big) V_k^I + \sum_{k=j}^{+\infty} (-1)^{j-k} \kappa_+^{j-k} \Big(\Phi_- + (-1)^{j-k-1} \Phi_+ \Big) V_k^I \right].
\end{aligned}$$

This ends the proof of Proposition 3.3.1. □

Secondly, we can see that the solution $(\widehat{U}_j)_{j\in\mathbb{N}}(z)$ to (3.3.2a)-(3.3.2c) consists of three parts:

$$\widehat{U}_j(z) = \widehat{U}_j^{I}(z) + \widehat{U}_j^{II}(z) + \widehat{U}_j^{III}(z), \tag{3.3.18}$$

where

$$
\begin{aligned}
\widehat{U}_j^I(z) =& \frac{(-1)^j \kappa_-^{j+1}(\varepsilon\xi)}{4g(\varepsilon\xi)} \times \Big((g(\varepsilon\xi), -1)\, w^{II}(\varepsilon\xi) \Big) \times r_+(\varepsilon\xi) \\
& - \frac{\kappa_-^j(\varepsilon\xi)}{\kappa_+(\varepsilon\xi)+\kappa_-(\varepsilon\xi)} \Big(\Phi_-(\varepsilon\xi) w^I(\varepsilon\xi) + (-1)^j \Phi_+(\varepsilon\xi) w^{II}(\varepsilon\xi) \Big) \\
& + \frac{1}{\kappa_+(\varepsilon\xi)+\kappa_-(\varepsilon\xi)} \times \Bigg[\sum_{k=0}^{j-1} \kappa_-^{j-k}(\varepsilon\xi) \Big(\Phi_-(\varepsilon\xi) + (-1)^{j-k-1}\Phi_+(\varepsilon\xi) \Big) V_k^I \\
& \qquad + \sum_{k=j}^{+\infty} (-1)^{j-k} \kappa_+^{j-k}(\varepsilon\xi) \Big(\Phi_-(\varepsilon\xi) + (-1)^{j-k-1}\Phi_+(\varepsilon\xi) \Big) V_k^I \Bigg],
\end{aligned}
\tag{3.3.19}
$$

$$
\widehat{U}_j^{II}(z) = \frac{(-1)^{j+1} \kappa_-^{j+1}(\varepsilon\xi)}{4g(\varepsilon\xi)} \times \frac{B_u - g(\varepsilon\xi)B_v}{B_u + g(\varepsilon\xi)B_v} \times \Big((g(\varepsilon\xi), 1)\, w^I(\varepsilon\xi) \Big) \times r_+(\varepsilon\xi) \tag{3.3.20}
$$

and

$$
\begin{aligned}
\widehat{U}_j^{III}(z) = \frac{\kappa_-^{j+1}(\varepsilon\xi)}{4g(\varepsilon\xi)} \times & \left(\frac{B_u - g(\varepsilon\xi)B_v}{B_u + g(\varepsilon\xi)B_v} \times (g(\varepsilon\xi), 1)\, w^I(\varepsilon\xi) - (g(\varepsilon\xi), -1)\, w^{II}(\varepsilon\xi) \right) \\
\times & \frac{B_u - g(\varepsilon\xi)B_v}{B_u + g(\varepsilon\xi)B_v} \times r_-(\varepsilon\xi).
\end{aligned}
\tag{3.3.21}
$$

Let us extend the initial data $(f_j)_{j\in\mathbb{Z}}$ to the whole line by setting $f_j = 0$ for $j \le 0$. It is easy to verify that $\widehat{U}_j^I(z)$ corresponds to the $\mathcal{Z}-$transform of the solution $(U_j^I)^n$ of the following extended Cauchy problem (3.3.22).

$$
\begin{cases}
\dfrac{\left(U_j^I\right)^{n+1} - \left(U_j^I\right)^n}{\Delta t} + \dfrac{1}{2\Delta x} A \Big(\left(U_{j+1}^I\right)^{n+1} - \left(U_{j-1}^I\right)^{n+1} \Big) = \dfrac{1}{\varepsilon} S \left(U_j^I\right)^{n+1}, & j \in \mathbb{Z},\ n \ge 0, \\
\left(U_j^I\right)^0 = f_j, & j \in \mathbb{Z}.
\end{cases}
\tag{3.3.22}
$$

With $(\widehat{U}_j^{II})_{j\in\mathbb{N}}(z)$ found in (3.3.20), the value of $(U_j^{II})_{j\in\mathbb{N}}^n$ can be obtained by inverting the $\mathcal{Z}$-transform

$$
\left(U_j^{II}\right)^n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{U}_j^{II}(R\mathrm{e}^{i\theta}) R^n \mathrm{e}^{in\theta} d\theta, \quad R > 1. \tag{3.3.23}
$$

Indeed, the value of $\widehat{U}_j^{III}(z)$ can be reformulated as

$$
\widehat{U}_j^{III}(z) = \frac{-B\left(\widehat{U}_0^I(z) + \widehat{U}_0^{II}(z)\right)}{B_u + g(\varepsilon\xi)B_v} \times \kappa_-(\varepsilon\xi) r_-(\varepsilon\xi).
$$

Following Section 3.2.2, $\widehat{U}_j^{III}(z)$ corresponds to the $\mathcal{Z}-$transform of the solution $(U_j^{III})^n$ of the IBVP with the homogeneous initial data

$$\begin{cases} \dfrac{\left(U_j^{III}\right)^{n+1}-\left(U_j^{III}\right)^{n}}{\Delta t}+\dfrac{1}{2\Delta x}A\Big(\left(U_{j+1}^{III}\right)^{n+1}-\left(U_{j-1}^{III}\right)^{n+1}\Big)=\dfrac{1}{\varepsilon}S\left(U_j^{III}\right)^{n+1}, & j\geq 1,\\ \left(U_j^{III}\right)^{0}=0, & j\geq 0,\\ B\left(U_0^{III}\right)^{n}=-B\Big(\left(U_0^{I}\right)^{n}+\left(U_0^{II}\right)^{n}\Big), & n\geq 0,\\ \dfrac{1}{\Delta t}\Gamma\left(\left(U_0^{III}\right)^{n+1}-\left(U_0^{III}\right)^{n}\right)+\dfrac{1}{2\Delta x}\Gamma A\left(\left(U_1^{III}\right)^{n+1}-\displaystyle\sum_{k=0}^{n+1}\mathcal{C}_{n+1-k}\left(U_0^{III}\right)^{k}\right)=\dfrac{1}{\varepsilon}\Gamma S\left(U_0^{III}\right)^{n+1}, & n\geq 0. \end{cases} \tag{3.3.24}$$

3.3.2 The energy method for the Cauchy problem

In this paragraph, we prove the Proposition 3.1.2 by means of the discrete energy method. The energy estimate in the continuous case are obtained using the integration by parts rule. Therefore, we need the corresponding summation by parts rules for the discrete approximations of ∂/∂_x [42]. The idea is to find a symmetric positive definite matrix H, such that HA is symmetric and HS is negative semi-definite. Therefore, we choose

$$H=\begin{pmatrix} a & 0\\ 0 & 1\end{pmatrix}.$$

Now, let us multiply the first equation in (3.3.22) by $((U_j^I)^{n+1})^T H$ and sum over $\mathbb{Z}$, one obtains

$$\sum_{j\in\mathbb{Z}}\left\langle (U_j^I)^{n+1}-(U_j^I)^n, H(U_j^I)^{n+1}\right\rangle+\frac{\Delta t}{2\Delta x}\sum_{j\in\mathbb{Z}}\left\langle A\left((U_{j+1}^I)^{n+1}-(U_{j-1}^I)^{n+1}\right), H(U_j^I)^{n+1}\right\rangle \\ =\frac{\Delta t}{\varepsilon}\sum_{j\in\mathbb{Z}}\left\langle S(U_j^I)^{n+1}, H(U_j^I)^{n+1}\right\rangle, \tag{3.3.25}$$

where $\langle .,.\rangle$ denotes the usual Euclidean inner product. Since H is a symmetric positive definite matrix, we have

$$\sum_{j\in\mathbb{Z}}\left\langle (U_j^I)^{n+1}-(U_j^I)^n, H(U_j^I)^{n+1}\right\rangle\geq\frac{1}{2}\sum_{j\in\mathbb{Z}}\left(\left\langle (U_j^I)^{n+1}, H(U_j^I)^{n+1}\right\rangle-\left\langle (U_j^I)^{n}, H(U_j^I)^{n}\right\rangle\right).$$

Together with the symmetric matrix HA, the second flux term in (3.3.25) becomes

$$\sum_{j\in\mathbb{Z}}\left\langle A\left((U_{j+1}^I)^{n+1}-(U_{j-1}^I)^{n+1}\right), H(U_j^I)^{n+1}\right\rangle=0.$$

Thus, we directly get the inequality

$$\sum_{j\in\mathbb{Z}}\left(\left\langle (U_j^I)^{n+1}, H(U_j^I)^{n+1}\right\rangle-\left\langle (U_j^I)^{n}, H(U_j^I)^{n}\right\rangle\right)\leq\frac{2\Delta t}{\varepsilon}\sum_{j\in\mathbb{Z}}\left\langle S(U_j^I)^{n+1}, H(U_j^I)^{n+1}\right\rangle. \tag{3.3.26}$$

Let us remind that HS is negative semi-definite. Then, from the inequality (3.3.26), for any $n\in\mathbb{N}$, the following inequality holds

$$\sum_{j\in\mathbb{Z}}\left\langle (U_j^I)^{n}, H(U_j^I)^{n}\right\rangle\leq\sum_{j\in\mathbb{Z}}\left\langle f_j, Hf_j\right\rangle. \tag{3.3.27}$$

Furthermore, since H is a symmetric positive definitive matrix, the following inequality holds for some constants $m, k > 0$

$$m \left\langle (U_j^I)^n, H(U_j^I)^n \right\rangle \leq \left\langle (U_j^I)^n, (U_j^I)^n \right\rangle \leq k \left\langle (U_j^I)^n, H(U_j^I)^n \right\rangle . \tag{3.3.28}$$

According to (3.3.27) and (3.3.28), there exists a constant $C > 0$ such that

$$\sum_{j\in\mathbb{Z}} \Delta x \left|\left(U_j^I\right)^n\right|^2 \leq C \sum_{j\in\mathbb{Z}} \Delta x |f_j|^2, \qquad \text{for any } n \in \mathbb{N}, \tag{3.3.29}$$

with the constant C independent of ε and Δx.
This ends the proof of the Proposition 3.1.2.

To complete the proof of the Theorem 3.1.1 for the numerical scheme of the IBVP (3.3.1), observe that from (3.3.29) and setting $f_j = 0$ for $j < 0$, for any $T > 0$, there exists $C_T > 0$ such that

$$\sum_{n=0}^{N} \sum_{j\geq 0} \Delta x \Delta t \left|\left(U_j^I\right)^n\right|^2 \leq C_T \sum_{j\geq 0} \Delta x |f_j|^2, \tag{3.3.30}$$

with $N = T/\Delta t$. Furthermore, from the inequality (3.3.29) and $\Delta x = \Delta t \lambda_{xt}$, one obtains

$$\sum_{n=0}^{N} \Delta t \left|(U_0^I)^n\right|^2 \leq C_T \sum_{j\geq 0} \Delta x |f_j|^2. \tag{3.3.31}$$

3.3.3 The uniform estimate on $(U_j^{II})^n$

The following lemma concerns the estimate on $(U_j^{II})^n$:

Lemma 3.3.2. *Assume that the SKC (3.1.3) is satisfied and let $\lambda_{xt} \leq 3\sqrt{a}/8$ be a positive number. Then, for any $T > 0$, there exists a constant $C_T > 0$ such that for any $\Delta t > 0$ together with $\Delta x = \lambda_{xt}\Delta t$, for any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$, the values of $(U_j^{II})_{j\in\mathbb{N}}$ defined in (3.3.23) satisfy*

$$\sum_{n=0}^{N} \sum_{j\geq 0} \Delta t \Delta x \left|(U_j^{II})^n\right|^2 + \sum_{n=0}^{N} \Delta t \left|(U_0^{II})^n\right|^2 \leq C_T \sum_{j\geq 0} \Delta x |f_j|^2 \tag{3.3.32}$$

where $N := T/\Delta t$ and C_T is independent of $\varepsilon \in (0, +\infty)$.

Proof. By an application of the following Plancherel's theorem for $\mathcal{Z}$-transform, we have

$$\sum_{n\geq 0} R^{-2n} \left|(U_0^{II})^n\right|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\widehat{U}_0^{II}(Re^{i\theta})|^2 d\theta, \quad R > 1$$
$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|\kappa_-(\varepsilon\xi)|^2}{16|g(\varepsilon\xi)|^2} \times \left|\frac{B_u - g(\varepsilon\xi)B_v}{B_u + g(\varepsilon\xi)B_v}\right|^2 \times \left|(g(\varepsilon\xi), 1)w^I(\varepsilon\xi)\right|^2 \times \left(1 + |g(\varepsilon\xi)|^2\right) d\theta.$$

From the Lemma 3.2.1, still under the SKC, $B_u + g(\varepsilon\xi)B_v$ is uniformly bounded away from 0 in $\varepsilon\xi \in \mathbb{C}_+$, and $g(\varepsilon\xi)$ is uniformly bounded in $\varepsilon\xi \in \mathbb{C}_+$. Moreover one has $|\kappa_-(\varepsilon\xi)| < 1$, we therefore obtain

$$\sum_{n\geq 0} R^{-2n} \left|(U_0^{II})^n\right|^2 \lesssim \sum_{k\geq 0} |f_k|^2. \tag{3.3.33}$$

Similarly, by an application of the Plancherel's theorem for $\mathcal{Z}$-transform, under the SKC, we have

$$\sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}\left|(U_j^{II})^n\right|^2 \lesssim \sum_{j\geq 0}|\kappa_-(\varepsilon\xi)|^{2j}\sum_{k\geq 0}|f_k|^2. \tag{3.3.34}$$

Following Lemma 3.2.6, since we assume that the condition (3.2.29) holds, one gets the following property

$$\sum_{j\geq 0}|\kappa_-(\varepsilon\xi)|^{2j} \leq \frac{\Delta t\sqrt{a}}{\Delta x(1-R^{-1})}.$$

Together with $\lambda_{xt} = \Delta x/\Delta t$, the inequality (3.3.34) becomes

$$\frac{R-1}{R}\sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}\Delta x\left|(U_j^{II})^n\right|^2 \lesssim \sum_{k\geq 0}\Delta x|f_k|^2. \tag{3.3.35}$$

Assembling the estimates (3.3.33) and (3.3.35), there exists $C > 0$ such that

$$\frac{R-1}{R}\sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}\Delta x\left|(U_j^{II})^n\right|^2 + \sum_{n\geq 0}\sum_{j\geq 0} R^{-2n}\Delta t\left|(U_0^{II})^n\right|^2 \leq C\sum_{k\geq 0}\Delta x|f_k|^2.$$

By setting in the above formula $R = \mathsf{e}^{\gamma\Delta t}$ for $\gamma > 0$ and $\Delta t > 0$, and using the classical lower bound $\mathsf{e}^{\gamma\Delta t} \geq 1+\gamma\Delta t$, we obtain that there exists a constant $c > 0$ such that

$$\frac{\gamma}{\gamma\Delta t+1}\sum_{n\geq 0}\sum_{j\geq 0}\mathsf{e}^{-2\gamma n\Delta t}\Delta t\Delta x\left|(U_j^{II})^n\right|^2 + \sum_{n\geq 0}\mathsf{e}^{-2\gamma n\Delta t}\Delta t\left|(U_0^{II})^n\right|^2 \leq C\sum_{k\geq 0}\Delta x|f_k|^2.$$

Then, for all $T > 0$, there exists a constant $C_T > 0$ such that

$$\sum_{n=0}^{N}\sum_{j\geq 0}\Delta t\Delta x\left|(U_j^{II})^n\right|^2 + \sum_{n=0}^{N}\Delta t\left|(U_0^{II})^n\right|^2 \leq C_T\sum_{k\geq 0}\Delta x|f_k|^2,$$

with $N = T/\Delta t$. □

3.3.4 Stiff stability analysis

Following Section 3.2.3, for any $T > 0$, there exists $C_T > 0$ such that the solution $(U_j^{III})^n_{j\in\mathbb{N}}$ to (3.3.24) satisfies

$$\sum_{n=0}^{N}\sum_{j\geq 0}\Delta x\Delta t\left|(U_j^{III})^n\right|^2 + \sum_{n=0}^{N}\Delta t\left|(U_0^{III})^n\right|^2 \leq C_T\sum_{n=0}^{N}\Delta t\left|B\Big((U_0^{I})^n + (U_0^{II})^n\Big)\right|^2.$$

Furthermore, from the inequalities (3.3.31) and (3.3.32), one obtains

$$\sum_{n=0}^{N}\Delta t\left|B\Big((U_0^{I})^n + (U_0^{II})^n\Big)\right|^2 \leq C_T\sum_{k\geq 0}\Delta x|f_k|^2.$$

Therefore, we show the uniform estimate on $(U_j^{III})^n_{j\in\mathbb{N}}$

$$\sum_{n=0}^{N}\sum_{j\geq 0}\Delta x\Delta t\left|(U_j^{III})^n\right|^2 + \sum_{n=0}^{N}\Delta t\left|(U_0^{III})^n\right|^2 \leq C_T\sum_{j\geq 0}\Delta x|f_j|^2, \tag{3.3.36}$$

with the positive constant C_T independent of $\varepsilon, \Delta x$ and Δt.

To complete the proof of the Theorem 3.1.1 for the numerical scheme of the IBVP (3.3.1), observe that from the inequalities (3.3.30) and (3.3.36), for any $T > 0$, there exists $C_T > 0$ such that for any $(f_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N}, \mathbb{R}^2)$ and $N := T/\Delta t$, the solution $(U_j^n)_{j\in\mathbb{N}}$ to (3.3.1) satisfies

$$\sum_{n=0}^{N}\sum_{j\geq 0}\Delta x\Delta t|U_j^n|^2 + \sum_{n=0}^{N}\Delta t|U_0^n|^2 \leq C_T\sum_{j\geq 0}\Delta x|f_j|^2, \tag{3.3.37}$$

where the constant C_T independent of $\varepsilon, \Delta x$ and Δt. This is the last step to prove the Theorem 3.1.1.

3.3.5 Numerical experiments

In this paragraph, we present some numerical experiments for the behavior of the numerical solution $(U_j^n)_{j\in\mathbb{N}}$ corresponding to whether or not the SKC (3.1.3) holds. We also look at the numerical solution when the UKC (2.1.5) is wrong. After that, we observe the effective behavior of the energy terms E_1 inside the domaine and E_2 along the boundary, which are defined in (3.2.37).

In our numerical experiments, we choose $a = 1$, $B_v = 1$, $\lambda_{xt} = 1/3$, fix the space step $\Delta x = 5 \times 10^{-3}$, the time step $\Delta t = \lambda_{xt}^{-1}\Delta x$, and let the relaxation rate ε and the boundary data B_u vary. The boundary data is the homogeneous one $b^n \equiv 0$, for any $n \in \mathbb{N}$. The initial data is

$$f_j = \begin{cases} 100 \times \left(\dfrac{13}{30} - x_j\right)\left(x_j - \dfrac{1}{4}\right) \times \begin{pmatrix}1 & -1\end{pmatrix}^T, & \text{if } x_j \in \left[\dfrac{1}{4}, \dfrac{13}{30}\right], \\ \begin{pmatrix}0 & 0\end{pmatrix}^T, & \text{otherwise.}\end{cases}$$

Let us first observe that these data are compatible in the corner $(x,t) = (0,0)$ in the sense that $Bf_0 = 0$. Moreover, the choice of an initial data with support in $[1/4, 13/30]$ is motivated by the property of finite speed of propagation available at the continuous side (2.1.1). More precisely, the exact solution we approximate has characteristic velocities ±1 and therefore vanishes outside some space interval $[0, 0.63]$ for small times in $[0, 0.2]$. Thus, we choose for our experiments the space interval $[0, 1]$ and the time interval $[0, T)$ with $T = 0.2$. Let us mention that the numerical experiments are performed we another discrete right boundary condition at $x = 1$. This is chosen to be the classical homogeneous first order Neumann extrapolation boundary condition $U_{J+1}^n = U_J^n$, for any $n \in \mathbb{N}$, at the rightmost cell J. That boundary condition indeed exhibits convenient stability features for both the inflowing and the outflowing transport equation [36].

3.3.5.1 The behavior of the numerical solution

Firstly, we choose a set of values B_u such that the SKC (3.1.3) is satisfied with $\varepsilon = 10^{-2}$ and also with $\varepsilon = 10^2$. The Figures 3.3.1 and 3.3.2 show numerical solution $(U_j^n)_{j\in\mathbb{N}}$ over the time interval $t \in [0, 0.2)$.

In the first case, $\varepsilon = 10^{-2}$, due to the initial relaxation of solution to the equilibrium system, the numerical solution descends over time (see Figure 3.3.1). In the case $\varepsilon = 10^2$, at time $t < 0.2$, its solution seems to translate to the left and the ghost solutions do not go backward in space for the implicit scheme. After that, the initial condition re-enters the domain from the left boundary (see Figure 3.3.2). It is not so much influenced by relaxation source term but more by the boundary dissipation.

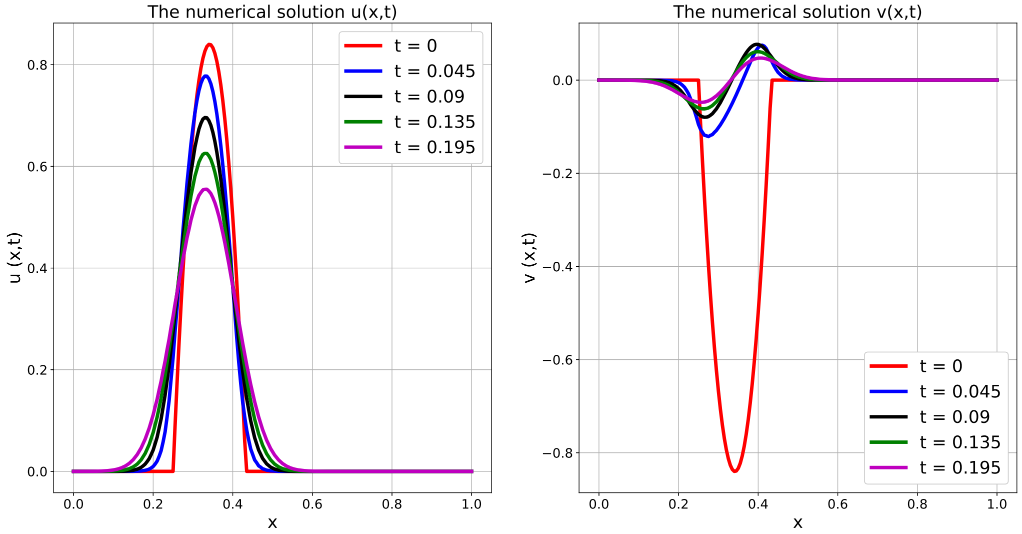

Figure 3.3.1: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^{-2}$. The SKC (3.1.3) is valid with $B_u = -2$.

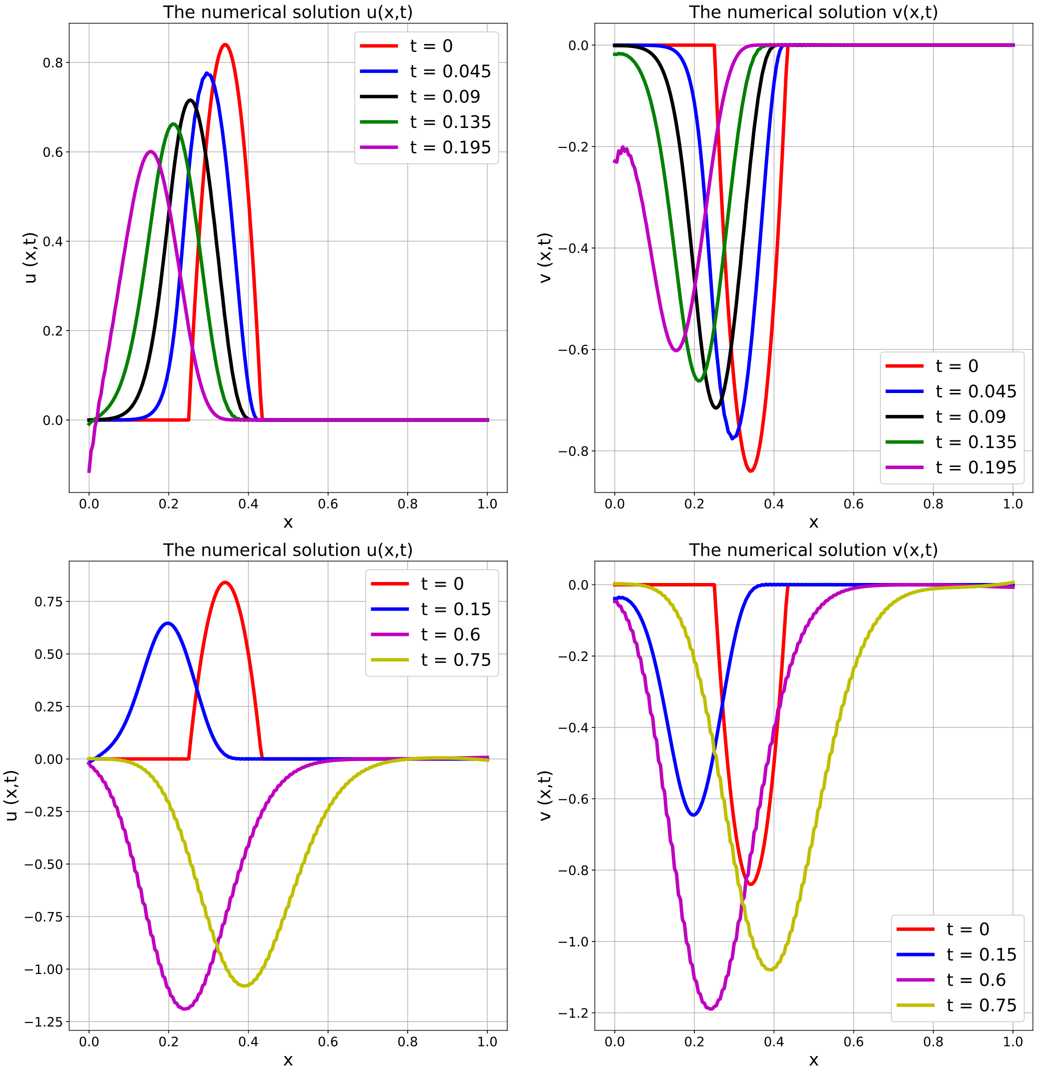

Figure 3.3.2: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^{2}$. The SKC (3.1.3) is valid with $B_u = -2$.

Secondly, we choose the value of B_u such that the SKC (3.1.3) is not satisfied. Besides, we also present the numerical solution when the Uniform Kreiss Condition (2.1.5) is wrong. The Figures 3.3.3 and 3.3.4 show numerical solution $(U_j^n)_{j\in\mathbb{N}}$ over the time interval $t \in [0, 0.2)$.

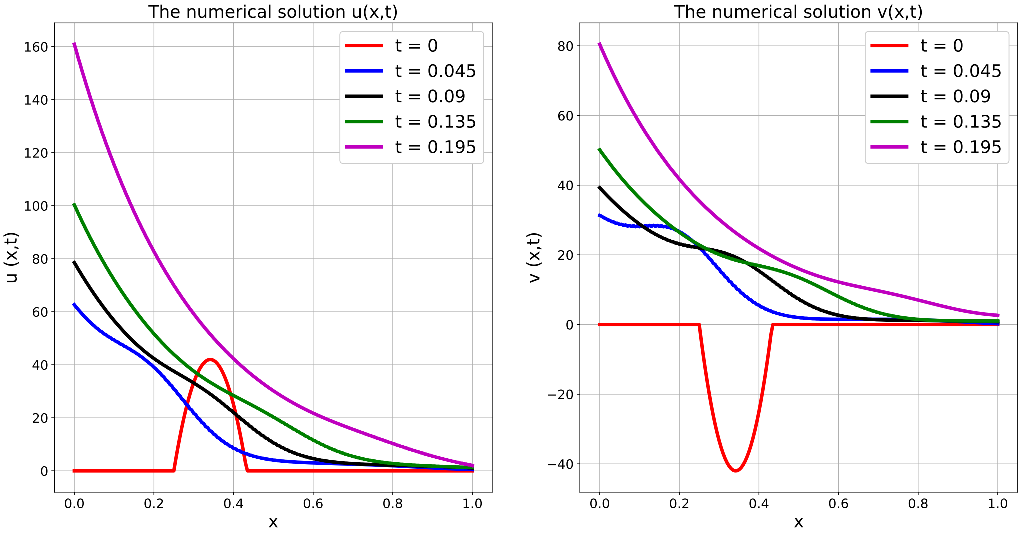

Figure 3.3.3: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^{-2}$. The SKC (3.1.3) is not valid with $B_u = -0.5$.

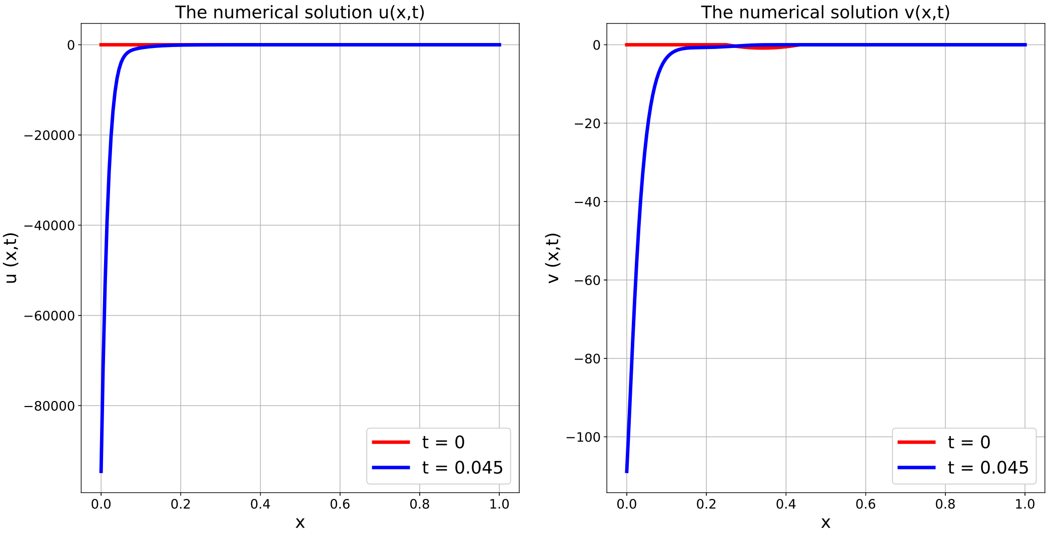

Figure 3.3.4: The numerical solution $u(x,t)$ (left) and $v(x,t)$ (right) for $\varepsilon = 10^2$. The UKC (2.1.5) is wrong with $B_u = -1$.

We can observe that the numerical solution at the boundary rise gradually when the SKC (3.1.3) fails. This is the case for example for $\varepsilon = 10^{-2}$ together with the parameters $(B_u, B_v) = (-1/2, 1)$. When the UKC (2.1.5) does not hold, the behavior is even worse (see Figure 3.3.4).

3.3.5.2 The effective behavior of the energy terms

We present hereafter the effective behavior of the energy terms E_1 and E_2 for $\varepsilon \in (0, +\infty)$, $T = 0.2$ and $N = T/\Delta t$.

According to Table 3.4 and Figures 3.3.5, 3.3.6, we can see that

B_u	$\varepsilon = 10^{-2}$	$\varepsilon = 1$	$\varepsilon = 10^2$
-4	0.038009	0.083375	0.160487
-2	0.038014	0.191338	0.440354
-1	7.54×10^{19}	8.94×10^{30}	3.149×10^{41}
-0.5	696.71	16628.4	101893.6
1	0.030684	0.03537	0.038031
3	0.035051	0.038022	0.046627

B_u	$\varepsilon = 10^{-2}$	$\varepsilon = 1$	$\varepsilon = 10^2$
-4	2.13×10^{-5}	1.05×10^{-3}	1.46×10^{-3}
-2	2.69×10^{-5}	0.43×10^{-2}	0.71×10^{-2}
-1	5.3×10^{21}	8.68×10^{32}	3.58×10^{43}
-0.5	19235.1	125624.5	437872.2
1	2.87×10^{-5}	0.0515	0.0715
3	1.9×10^{-5}	0.0620	0.0894

Table 3.4: The energy terms E_1 (left) and E_2(right).

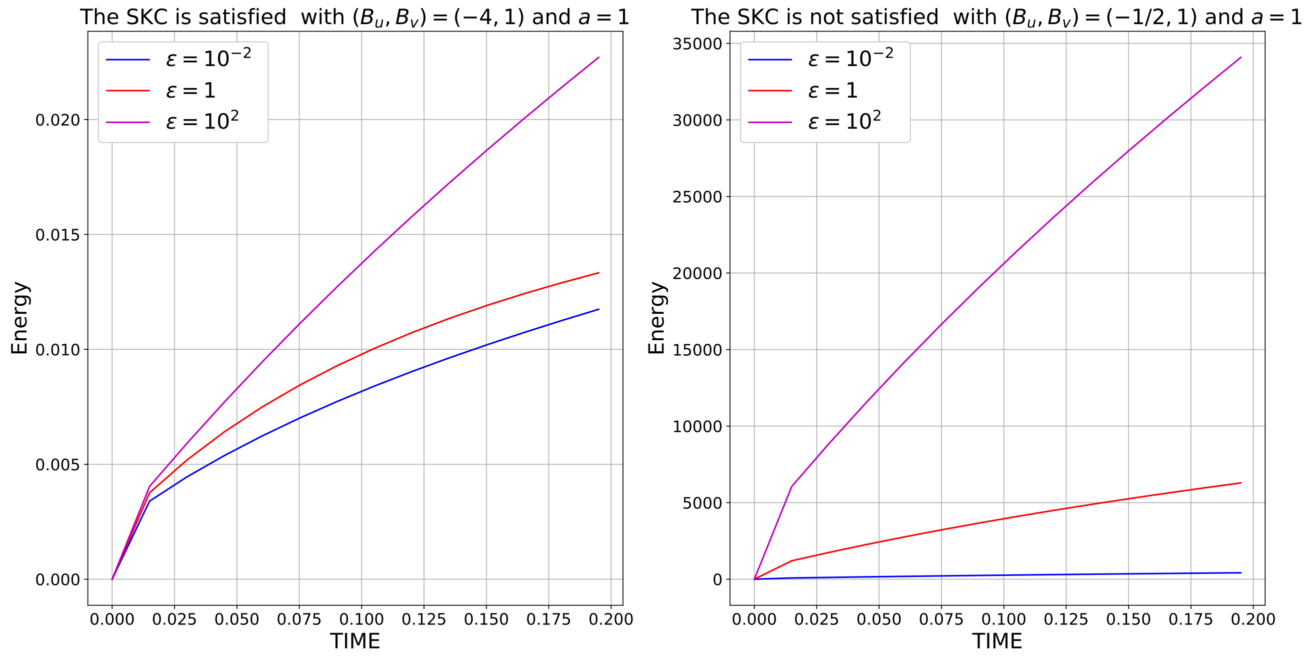

Figure 3.3.5: Energy evolution E_1 for $B_u = -4$ (left) and $B_u = -0.5$ (right).

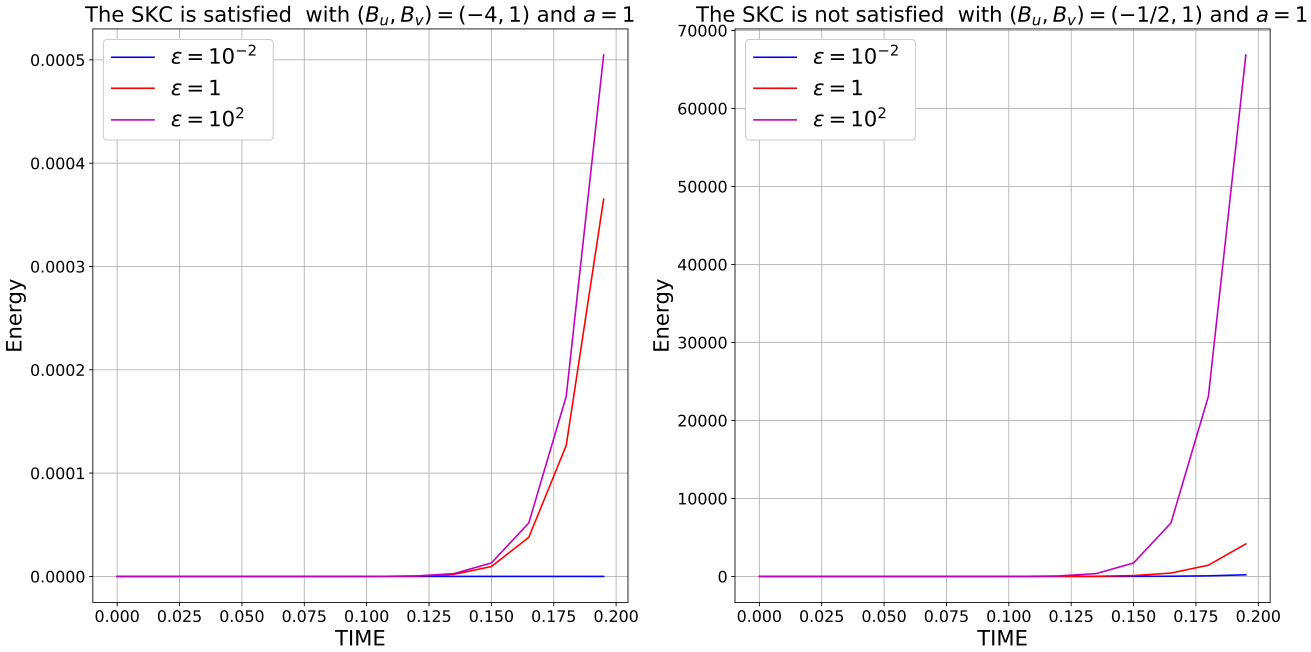

Figure 3.3.6: Energy evolution E_2 for $B_u = -4$ (left) and $B_u = -0.5$ (right).

- For any $\varepsilon \in (0, +\infty)$, the values of E_1 and E_2 rise gradually when the SKC (3.1.3) is not satisfied. This is the case for example for $\varepsilon = 10^2$ together with the parameters $(B_u, B_v) = (-1/2, 1)$. The behavior is even worse when the UKC (2.1.5) is not hold.

- On the boundary $x = 0$, the value of E_2 for $\varepsilon = 10^{-2}$ increase slowly. This is due to the effect of incoming solution at the boundary when the initial relaxation of solution tends to the equilibrium system. In the case $\varepsilon = 10^2$, its value increase fairly rapidly. It is not so much influenced by relaxation source term but more by the boundary dissipation.

Clearly, in our numerical experiment, the numerical solution at the boundary $x = 0$ increase quickly as soon as the SKC (3.1.3) is not valid. The behavior of numerical solution is even worse if the UKC (2.1.5) is not satisfied. Indeed, it seems that the SKC (3.1.3) is also necessary condition to ensure the non-increase rapidly of the energy terms E_1 and E_2 under the effect of the relaxation source term and the boundary dissipation.

Chapter 4

High order numerical schemes for transport equations on bounded domains

The goal is to construct finite difference approximations of the transport equation with nonzero incoming boundary data that achieve the best possible convergence rate in the maximum norm. We construct, implement and analyze the so-called inverse Lax-Wendroff procedure at the incoming boundary. Optimal convergence rates are obtained by combining sharp stability estimates for extrapolation boundary conditions with numerical boundary layer expansions. We illustrate the results with the Lax-Wendroff and $O3$ schemes.

4.1 Introduction

4.1.1 Context and motivation

The implementation of numerical boundary conditions is crucial importance for the simulation of transport and other evolution phenomena. However, a complete analysis of stability and/or accuracy issues depending on the type of boundary and the type of numerical schemes (e.g., finite difference schemes with one or more time levels) is missing. The goal of this chapter is to propose a high order numerical treatment of nonzero incoming boundary data for the transport equation. The methodology is developed here for the one-dimensional problem but it is our hope that the tools used below will be useful for higher dimensional problems. We are thus given a fixed constant velocity $a > 0$, an interval length $L > 0$ and we consider the (continuous) problem

$$\begin{cases} \partial_t u + a\,\partial_x u \;=\; 0\,, & t \geq 0\,, \quad x \in (0,L)\,, \\ u(0,x) \;=\; f(x)\,, & x \in (0,L)\,, \\ u(t,0) \;=\; g(t)\,, & t \geq 0\,. \end{cases} \tag{4.1.1}$$

The requirements on the initial and boundary data, namely f and g, will be made precise below. The solution to (4.1.1) is given by the method of characteristics, which yields the explicit representation formula

$$\forall\,(t,x) \in \mathbb{R}^+ \times (0,L)\,, \quad u(t,x) \;=\; \begin{cases} f\,(x - a\,t)\;, & \text{if } x \geq a\,t\,, \\ g\left(t - \dfrac{x}{a}\right)\;, & \text{if } x \leq a\,t\,. \end{cases} \tag{4.1.2}$$

The question we address is how to construct high order numerical approximations of the solution (4.1.2) to (4.1.1) by means of (explicit) finite difference approximations. This problem has been addressed in [22] in the case of *zero* incoming boundary data (that is, $g = 0$ in (4.1.1)). The focus in [22] is on the outflow boundary ($x = L$ here since a is positive), for which *extrapolation* numerical boundary conditions are analyzed. Fortunately for us, a large part of the analysis in [22] can be used here as a black box and we therefore focus on the incoming boundary. To motivate the analysis of this chapter, let us present a very simple -though illuminating- example for which we just need to introduce the basic notations that will be used throughout this chapter. In all what follows, we consider a positive integer J, that is meant to be large, and define the space step Δx and the grid points $(x_j)_{j\in\mathbb{Z}}$ by

$$\Delta x := \frac{L}{J}, \quad x_j := j\Delta x \quad (j \in \mathbb{Z}).$$

The interval $(0, L)$ corresponds to the cells (x_{j-1}, x_j) with $j = 1, \ldots, J$, but considering the whole real line $\{j \in \mathbb{Z}\}$ will be useful in some parts of the analysis. The time step Δt is then defined as $\Delta t := \lambda\,\Delta x$, where $\lambda > 0$ is a constant that is fixed so that Assumption 4.1.1 below is satisfied. We use from now on the notation $t^n := n\,\Delta t$, $n \in \mathbb{N}$, the quantity u_j^n will play the role of an approximation for the solution u to (4.1.1) at the time t^n on the cell (x_{j-1}, x_j).

We now examine an example where the exact solution to (4.1.1) is approximated by means of the Lax-Wendroff scheme. The approximation reads

$$u_j^{n+1} = u_j^n - \frac{\lambda\,a}{2}\,(u_{j+1}^n - u_{j-1}^n) + \frac{(\lambda\,a)^2}{2}\,(u_{j+1}^n - 2\,u_j^n + u_{j-1}^n)\,, \quad n \in \mathbb{N}\,, \quad 1 \leq j \leq J\,, \tag{4.1.3}$$

where we recall that $\lambda = \Delta t/\Delta x$ is a fixed constant and $a > 0$ is the transport velocity in (4.1.1). The initial condition for (4.1.3) is defined, for instance, by computing the cell averages of the initial condition f in (4.1.1), namely

$$\forall\, 1 \leq j \leq J\,, \quad u_j^0 := \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x\,. \tag{4.1.4}$$

Without any boundary, the Lax-Wendroff scheme is a second order approximation to the transport equation [41]. We would like, of course, to maintain the second order accuracy property when implementing (4.1.3) on an interval. This implementation, however, requires, at each time iteration n, the definition of the boundary (or *ghost cell*) values u_0^n and u_{J+1}^n. At the outflow boundary, we prescribe an extrapolation condition [56, 36], the significance of which will be thoroughly justified in the next sections

$$u_{J+1}^n = 2\,u_J^n - u_{J-1}^n\,, \quad n \in \mathbb{N}\,. \tag{4.1.5}$$

Combining (4.1.3) with (4.1.5), the last interior cell value u_J^n obeys the induction formula

$$u_J^{n+1} = u_J^n - \lambda\,a\,(u_J^n - u_{J-1}^n)\,, \quad n \in \mathbb{N}\,,$$

which is nothing but the upwind scheme. It then only remains to determine the inflow numerical boundary condition u_0^n. Since we wish to approximate the exact solution to (4.1.1) and u_0^n is meant, at least, to approximate the trace $u(t^n, 0)$, it seems reasonable at first sight to prescribe the Dirichlet boundary condition

$$u_0^n = g(t^n)\,, \quad n \in \mathbb{N}\,. \tag{4.1.6}$$

In the case of *zero* incoming boundary data ($g = 0$), and for any sufficiently smooth initial condition f that is “flat” at the incoming boundary, the main result of [22] shows that the above numerical scheme (4.1.3), (4.1.4), (4.1.5), (4.1.6) converges towards the exact solution to (4.1.1) with a rate of convergence $3/2$ in the maximum norm. Numerical simulations even predict that the rate of convergence should be 2, or at least close to 2, for smooth initial data.

However, implementing the above numerical scheme[1] quickly shows that the rate of convergence falls down to 1 when g is nonzero and satisfies the compatibility conditions[2] described hereafter with the initial condition f.

Our goal is to provide with a thorough treatment of nonzero incoming boundary data and to design numerical boundary conditions that recover the optimal rate of convergence in the maximum norm (at least, the same rate of convergence as the one in [22] for zero boundary data). The strategy is not new and is now referred to as the *inverse Lax-Wendroff method*. It consists, as detailed below, in writing Taylor expansions with respect to the space variable x close to the incoming boundary and then using the advection equation (4.1.1) to substitute the normal derivatives $\partial_x^m u(t,0)$ for tangential derivatives $\partial_t^m u(t,0)$, the latter being computed thanks to the boundary conditions in (4.1.1). This strategy is available when the boundary is non-characteristic [3].

The inverse Lax-Wendroff method is a general strategy that has been followed in various directions. We refer for instance to [86, 32, 92, 25] for various implementations related to either hyperbolic or kinetic partial differential equations. In these works, most of the time, the incoming numerical boundary condition prescribes the ghost cell value u_0^n in terms of the boundary datum g but also of *interior cell* values u_j^n with $j \geq 1$. This is the reason why *stability* is a real issue in these works, see for instance the discussion in [92, Section 4], and many rigorous justifications are still open. We develop here a simplified version of some of those previously proposed boundary treatments, but we rigorously justify the convergence with an (almost) optimal rate of convergence. As in [22], the key ingredient in our analysis is an *unconditional* stability result for the Dirichlet boundary conditions which dates back to [37, 38], see an alternative proof in [21].

4.1.2 The inverse Lax-Wendroff method

We first fix from now on some notations. In all this chapter, we are given some fixed integers $p, r \in \mathbb{N}$ and consider an explicit two time step approximation for the solution to (4.1.1)

$$u_j^{n+1} = \sum_{\ell=-r}^{p} a_\ell \, u_{j+\ell}^n \,, \quad n \in \mathbb{N} \,, \quad j = 1, \dots, J \,. \tag{4.1.7}$$

In (4.1.7), the numbers $a_{-r}, \dots, a_p$ are defined in terms of the parameter λ and of the velocity a (see, for instance, (4.1.3) for which $p = r = 1$). These numbers are fixed, which means that (4.1.7) is linear with respect to (u_j^n). For simplicity, we follow [22] and choose as initial data for (4.1.7) the cell averages of the initial condition f in (4.1.1). This means that the vector $(u_1^0, \dots, u_J^0)$ is defined by (4.1.4). For (4.1.7) to define inductively (with respect to n) the vector $(u_1^n, \dots, u_J^n)$, we need to prescribe the ghost cell values $u_{1-r}^n, \dots, u_0^n$ and $u_{J+1}^n, \dots, u_{J+p}^n$. They are depicted in red in Figure 4.1.1 (in that example, $p = r = 2$).

As explained above, we focus here on the inflow boundary and we therefore follow the extrapolation boundary treatment of [22] for the outflow boundary. Namely, if we define the finite difference operator D_- as

$$(D_- v)_j := v_j - v_{j-1} \,,$$

and its iterates D_-^m accordingly, we choose from now on an extrapolation order $k_b \in \mathbb{N}$ for the outflow boundary and prescribe

$$(D_-^{k_b} u^n)_{J+\ell} = 0 \,, \quad n \in \mathbb{N} \,, \quad \ell = 1, \dots, p \,. \tag{4.1.8}$$

[1] One can choose for instance $a = 1$, $\lambda = 5/6$, $L = 6$, $f(x) = \sin(x)$, $g(t) = -\sin(t)$ and increase the integer J geometrically.

[2] The rate of convergence could be even smaller than 1 when the compatibility conditions are not satisfied but that would just reflect the fact that the exact solution (4.1.2) is not smooth (for instance, not even continuous if $f(0) \neq g(0)$).

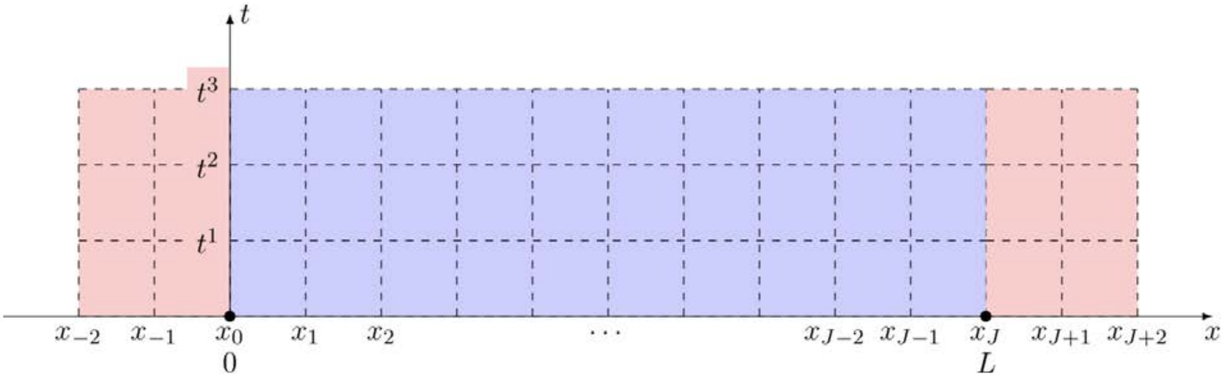

Figure 4.1.1: The mesh on $\mathbb{R}^+ \times (0, L)$ in blue, and the "ghost cell" in red ($r = p = 2$ here).

If $k_b = 0$, this correspond to prescribing homogeneous Dirichlet boundary condition:

$$u^n_{J+\ell} = 0, \quad \forall\, \ell = 1, ..., p,$$

while if $k_b = 1$, this correspond to the standard Neumann numerical boundary condition:

$$u^n_{J+1} = ... = u^n_{J+p} := u^n_J.$$

The example (4.1.5) corresponds to $k_b = 2$ (recall $p = 1$ for the Lax-Wendroff scheme). It now remains to prescribe the inflow values $u^n_{1-r}, \ldots, u^n_0$. Unlike some previous works, we are going to prescribe Dirichlet boundary conditions, meaning for instance that the value u^n_0 will be determined in terms of the boundary datum g only. Let us assume for a while that u^n_j is a second order approximation of $u(t^n, (x_{j-1} + x_j)/2)$ where u is the exact solution (4.1.2) of the continuous problem (4.1.1). Then we *formally* have

$$u^n_0 \approx u\left(t^n, -\frac{\Delta x}{2}\right) \approx u(t^n, 0) - \frac{\Delta x}{2}\, \partial_x u(t^n, 0)\,,$$

where $\approx$ means "equal up to $O(\Delta x^2)$", and we then use (4.1.1) to get

$$u^n_0 \approx u(t^n, 0) + \frac{\Delta x}{2\,a}\, \partial_t u(t^n, 0) \;=\; g(t^n) + \frac{\Delta x}{2\,a}\, g'(t^n)\,.$$

The last term $\Delta x/(2\,a)\, g'(t^n)$ in the previous equality is precisely the correction that is required to recover the second order accuracy when dealing with the Lax-Wendroff scheme (compare with (4.1.6)). More generally speaking, we could have pushed further the above Taylor expansion and obtained as a final (formal) result that u^n_0 should be "close" to some quantity of the form

$$\sum_{\kappa=0}^{K} \frac{\Delta x^\kappa}{\kappa\,!\, a^\kappa}\, \alpha_\kappa\, g^{(\kappa)}(t^n)\,,$$

where K is a truncation order and $\alpha_0, \ldots, \alpha_K$ are numerical constants.

The general form of the Dirichlet boundary conditions that we consider below is

$$u^n_\ell \;=\; \sum_{\kappa=0}^{K} \frac{\Delta x^\kappa}{\kappa\,!\, a^\kappa}\, \alpha_{\kappa,\ell}\, g^{(\kappa)}(t^n)\,, \quad n \in \mathbb{N}\,, \quad \ell = 1 - r, \ldots, 0\,,$$

where the $\alpha_{\kappa,\ell}$'s are numerical constants which will play a role (together with the truncation order K) in the consistency analysis. There are two main choices which we discuss in this chapter. The first one is given in [86, 92]

$$\alpha_{\kappa,\ell} \;:=\; \left(\frac{1}{2} - \ell\right)^\kappa\,, \quad \kappa \in \mathbb{N}\,, \quad \ell = 1 - r, \ldots, 0\,,$$

and is relevant if u_j^n is eventually compared in the convergence analysis with $u(t^n,(x_{j-1}+x_j)/2)$, u being the exact solution (4.1.2). The other possible choice we advocate is

$$\alpha_{\kappa,\ell} := \frac{(-1)^\kappa}{\kappa+1}\left(\ell^{\kappa+1}-(\ell-1)^{\kappa+1}\right), \quad \kappa\in\mathbb{N}, \quad \ell=1-r,\dots,0, \tag{4.1.9}$$

and is relevant if u_j^n is eventually compared (as in Theorem 4.1.2 below) in the convergence analysis with the average of $u(t^n,\cdot)$ on the cell (x_{j-1},x_j). The truncation order K is discussed with our main result in the following paragraph.

4.1.3 Main results

We assume that the approximation (4.1.7) is consistent with the transport operator and that it defines a stable procedure on $\ell^2(\mathbb{Z})$.

Assumption 4.1.1 (Consistency and stability without any boundary)**.** *The coefficients a_{-r}, ..., a_p in* (4.1.7) *satisfy $a_{-r}\,a_p\neq 0$ (normalization), and for some integer $k\geq 1$, there holds:*

$$\forall\, m=0,\dots,k\,,\quad \sum_{\ell=-r}^{p}\ell^m\,a_\ell \;=\; (-\lambda\,a)^m\,,\quad \text{(consistency of order } k\text{)}, \tag{4.1.10}$$

$$\sup_{\theta\in[0,2\pi]}\left|\sum_{\ell=-r}^{p}a_\ell\,e^{i\,\ell\,\theta}\right|\leq 1\,,\quad (\ell^2\text{-stability on }\mathbb{Z}). \tag{4.1.11}$$

Provided that the relations (4.1.10) are satisfied for $m=0$ (conservativity) and $m=1$ (consistency of order 1) with $a>0$, the stability assumption (4.1.11) implies $r\geq 1$, which we assume from now on. Though we view this observation here, as a *necessary* condition for *stability*, the condition $r\geq 1$ is also known to be necessary for convergence by comparing the numerical and continuous dependency domains. Let us observe that (4.1.11) is a necessary and sufficient condition for stability of the iteration process (4.1.7) on $\ell^2(\mathbb{Z})$ in a strong sense, meaning here that the map

$$(v_j)_{j\in\mathbb{Z}}\mapsto\left(\sum_{\ell=-r}^{p}a_\ell v_{j+\ell}\right)_{j\in\mathbb{Z}}$$

is a contraction (its norm is not larger than 1) as an operator on $\ell^2(\mathbb{Z})$. However, (4.1.11) is not sufficient to yield stability in $\ell^\infty(\mathbb{Z})$ for (4.1.7), see [45, 87]. Note that through the dependence of the a_ℓ with respect to $\lambda=\Delta t/\Delta x$, (4.1.11) is usually intended to be true only under a so-called Courant-Friedrichs-Lewy (CFL) condition asking for λ to be less than some constant depending on the scheme and the velocity a. (Indeed, the Bernstein inequality for trigonometric polynomials implies $\lambda|a|\leq\max(p,r)$, see [84]). For the Lax-Wendroff scheme (4.1.3), we have $p=r=1$, the integer k equals 2, and (4.1.11) holds if and only if $\lambda\,a\leq 1$.

Let us observe that Assumption 4.1.1 does not include any *dissipative* behavior for (4.1.7), meaning that we do not assume a bound of the form

$$\forall\theta\in[-\pi,\pi],\qquad\left|\sum_{\ell=-r}^{p}a_\ell e^{i\ell\theta}\right|\leq 1-c\theta^{2q},$$

for some suitable integer q and positive constant c. In that respect, the framework of Assumption 4.1.1 is more general than the works [36, 43, 57, 76] and following works that are based on these pioneering results. We thus expect that our approach may be useful to deal with multidimensional problems in which dissipativity is most of the time excluded (or restrictive).

In Theorem 4.1.2 below and all what follows, the velocity $a > 0$, the length $L > 0$, the parameter $\lambda = \Delta t/\Delta x$ and the extrapolation order $k_b \in \mathbb{N}$ at the outflow boundary are given. Subsequent constants may depend on them. The integer $k \geq 1$ is also fixed such that Assumption 4.1.1 holds. We consider the initial condition (4.1.4) and its evolution by the numerical scheme (4.1.7), (4.1.8), the inflow values being given by

$$u_\ell^n = \sum_{\kappa=0}^{k-1} \frac{\Delta x^\kappa}{(\kappa+1)\,!\,(-a)^\kappa} \left(\ell^{\kappa+1} - (\ell-1)^{\kappa+1}\right) g^{(\kappa)}(t^n)\,, \quad n \in \mathbb{N}\,, \quad 1-r \leq \ell \leq 0\,. \tag{4.1.12}$$

The interation (4.1.7), (4.1.8), (4.1.12) thus as process as follows, see Figure 4.1.2 for an illustration. Given the vector $(u_1^n, ..., u_J^n)$ for some time level n, one first determines the ghost values $\left(u_{1-r}^n, ..., u_0^n, u_{J+1}^n, ..., u_{J+p}^n\right)$ by (4.1.8) and (4.1.12). The new vector $\left(u_1^{n+1}, ..., u_J^{n+1}\right)$ is then determined by applying (4.1.7). It is assume that $J \geq 1$ in order to make the space step $\Delta x = L/J$ meaningful and to have at least one cell in the interval $(0, L)$. Of course, prescribing (4.1.12) is meaningful only if g is sufficiently smooth (say, $g \in \mathcal{C}^{k-1}$). One could push further the Taylor expansion in (4.1.12) and consider higher order correctors but it would require further smoothness on g and it would eventually not improve our convergence result below, so fixing the truncation order $K = k-1$ seems to be the most convenient choice.

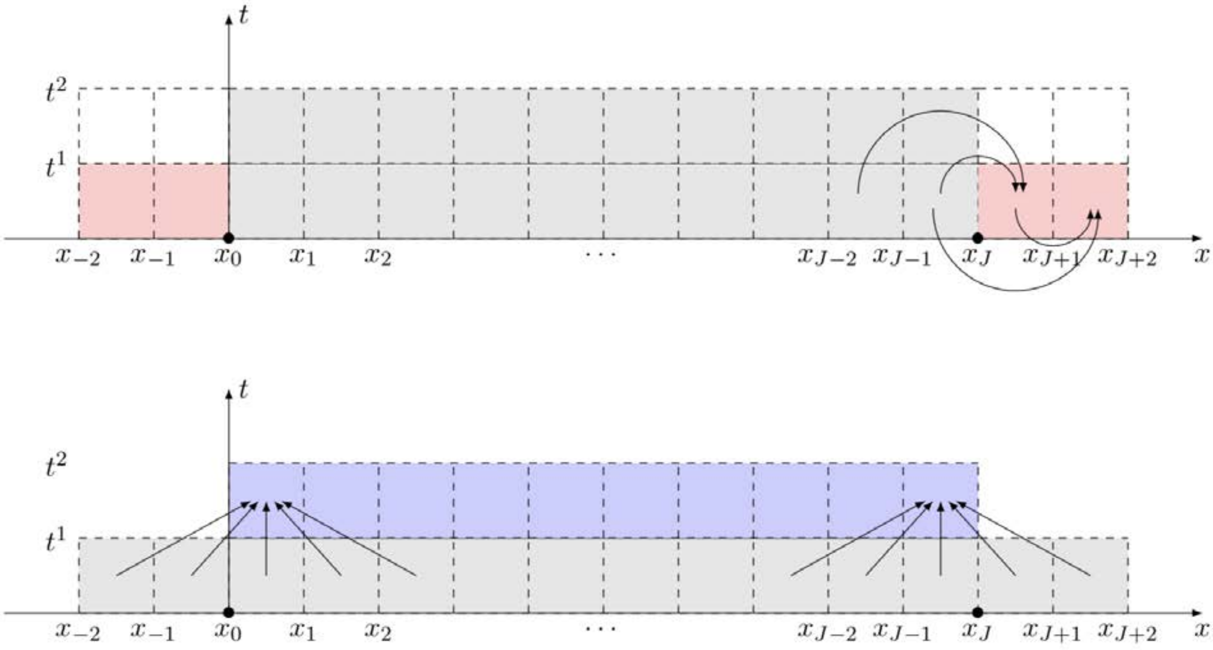

Figure 4.1.2: Top: updating iteratively the ghost values at the outflow boundary ($r = p = k_b = 2$). Bottom: updating the numerical approximation in the interior.

Our main convergence result is the extension of the main result in [22] to the case of nonzero boundary data.

Theorem 4.1.2 (Main convergence result)**.** *Let $a > 0$, $k \in \mathbb{N}^*$ and $k_b \in \mathbb{N}$. Under Assumption 4.1.1, there exists a constant $C > 0$ such that for any final time $T \geq 1$, any integer $J \in \mathbb{N}^*$, any data $f \in H^{k+1}((0, L))$ and $g \in H^{k+1}((0, T))$ satisfying the compatibility requirements at $t = x = 0$:*

$$\forall\, m = 0, \ldots, k\,, \quad f^{(m)}(0) \;=\; (-a)^{-m}\, g^{(m)}(0)\,,$$

the solution (u_j^n) *to* (4.1.4), (4.1.7), (4.1.8), (4.1.12) *satisfies:*

$$\sup_{0\le n\le T/\Delta t}\ \sup_{1\le j\le J}\left|u_j^n-\frac{1}{\Delta x}\int_{x_{j-1}}^{x_j}u(t^n,x)\,\mathrm{d}x\right|\le C\,T\,e^{C\,T/L}\,\Delta x^{\min(k,k_b)-1/2}\,\big(\|f\|_{H^{k+1}((0,L))}+\|g\|_{H^{k+1}((0,T))}\big),\tag{4.1.13}$$

with u *the exact solution to* (4.1.1), *whose expression is given by* (4.1.2).

Actually, the constant C in (4.1.13) is independent of $L\ge 1$, which is consistent with the convergence result we shall prove below for the half-space problem on $\mathbb{R}^+$ with inflow at $x=0$. As in [22], the loss of $1/2$ in the rate of convergence of Theorem 4.1.2 looks somehow artificial and is mostly a matter of passing from the $\ell_n^\infty\ell_j^2$ topology to $\ell_{n,j}^\infty$. Our next result examines a situation where the optimal convergence rate $\min(k,k_b)$ can be obtained. In order to simplify (and shorten) the proof of Theorem 4.1.3, we only examine here the case of a half-space with extrapolation outflow conditions. The extension of the techniques to the case of an interval is left to the interested reader.

Theorem 4.1.3 (Optimal rate of convergence for the outflow problem)**.** *Let* $a>0$, $k\in\mathbb{N}^*$ *and* $k_b\in\mathbb{N}$. *Under Assumption 4.1.1 and under the additional Assumption 4.3.2 stated hereafter, there exists a constant* $C>0$ *such that for any final time* $T\ge 1$, *any integer* $J\in\mathbb{N}^*$, *any data* $f\in H^{k+1}((-\infty,L))$, *the solution to the scheme*

$$\begin{cases} u_j^0 = \dfrac{1}{\Delta x}\displaystyle\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x\,, & j\le J\,,\\ (D_-^{k_b}u^n)_{J+\ell} = 0\,, & 0\le n\le T/\Delta t\,,\quad 1\le\ell\le p\,,\\ u_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\,u_{j+\ell}^n\,, & 0\le n\le T/\Delta t-1\,,\quad j\le J\,,\end{cases}\tag{4.1.14}$$

satisfies the error estimate

$$\sup_{0\le n\le T/\Delta t}\ \sup_{j\le J}\left|u_j^n-\frac{1}{\Delta x}\int_{x_{j-1}}^{x_j}f(x-a\,t^n)\,\mathrm{d}x\right|\le C\,T\,\Delta x^{k_b}\,\|f\|_{H^{k+1}((0,L))}\,,$$

as long as $k_b<k$.

In other words, the technical Assumption 4.3.2 hereafter, which is verified on many examples such as the Lax-Wendroff and $O3$ schemes, allow to recover the optimal rate $k_b=\min(k_b,k)$ in the case $k_b<k$. Of course, one would also like to improve the rate $\min(k_b,k)-1/2$ in the case $k_b=k$, which is clearly the most natural choice. However, in that case, both the interior and boundary consistency errors scale like Δx^k and, in the framework of Assumption 4.1.1, stability in the interior domain is available only in the ℓ_j^2 topology, so it is quite difficult to derive the convergence rate k in the ℓ_j^∞ topology. Theorem 4.1.3 already indicates that combining the approach of [22] with other techniques (here, boundary layer expansions) may improve some results. We hope to deal with the case $k_b=k$ in the future.

4.2 Convergence analysis for the inverse Lax-Wendroff method

This section is devoted to the proof of Theorem 4.1.2. Following [22], we shall prove Theorem 4.1.2 by using a stability estimate for (4.1.7), (4.1.8), (4.1.12) and a superposition argument, which amounts to considering separately two half-space problems: one in which there is only inflow at $x=0$, and one for which there is only outflow at $x=L$.

4.2.1 Stability estimates for the outflow problem

In this paragraph, we prove Theorem 4.2.1 below that provides us with stability estimates for the outflow problem. Theorem 4.2.1 is a key tool for proving stability estimate for the scheme (4.1.4), (4.1.7), (4.1.8), (4.1.12) on a finite interval, which in turn yields the convergence result of Theorem 4.1.2. Let us recall that Theorem 4.2.1 is already known to hold true thanks to the joint results of [36, 56, 57] and in a more general setting [15, 21, 95]. Before going on, let us fix the space domain that we consider. Since we deal with a constant coefficient linear problem, by translation invariance, there is no loss of generality in considering the half-line $(-\infty, L)$. The grid and the associated ghost cells are depicted in Figure 4.2.1.

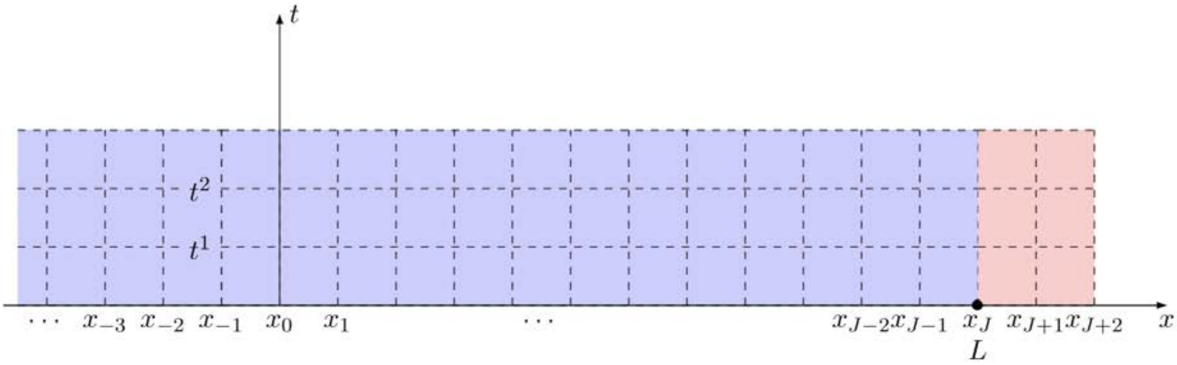

Figure 4.2.1: The mesh on $\mathbb{R}^+ \times (-\infty, L)$ in blue and the "ghost cells" in red ($p = 2$ here).

Theorem 4.2.1 (Stability estimates for the outflow problem [22]). *Let $a > 0$, $k \in \mathbb{N}^*$ and $k_b \in \mathbb{N}$. Under Assumption 4.1.1, there exists a constant $C > 0$ such that for any initial condition $(f_j)_{j \leq J} \in \ell^2$ and for all boundary source terms $(h^n_{J+\ell})_{1 \leq \ell \leq p,\, n \geq 0}$ verifying the growth condition*

$$\forall \, \Gamma > 0, \quad \sum_{n \geq 0} e^{-2\Gamma n} \left(\left(h^n_{J+1}\right)^2 + \ldots + \left(h^n_{J+p}\right)^2 \right) < +\infty,$$

the solution $(u^n_j)_{j \leq J+p,\, n \in \mathbb{N}}$ to the numerical scheme

$$\begin{cases} u^0_j = f_j, & j \leq J \\ (D^{k_b}_- u^n)_{J+\ell} \; = \; h^n_{J+\ell} \, , & n \in \mathbb{N} \, , \quad 1 \leq \ell \leq p \, , \\ u^{n+1}_j \; = \; \displaystyle\sum_{\ell=-r}^{p} a_\ell \, u^n_{j+\ell} \, , & n \in \mathbb{N} \, , \quad j \leq J \, , \end{cases} \tag{4.2.1}$$

satisfies

$$\begin{aligned} \sup_{n \in \mathbb{N}} \left(e^{-2\gamma n \Delta t} \sum_{j \leq J} \Delta x (u^n_j)^2 \right) + \sum_{n \geq 0} \Delta t e^{-2\gamma n \Delta t} \sum_{\ell = 1-r-k_b}^{p} (u^n_{J+\ell})^2 \\ \leq C \left(\sum_{j \leq J} \Delta x (f_j)^2 + \sum_{n \geq 0} \Delta t e^{-2\gamma n \Delta t} \sum_{\ell=1}^{p} (h^n_{J+\ell})^2 \right) , \end{aligned} \tag{4.2.2}$$

for any $\gamma > 0$. In particular, the numerical boundary condition in (4.2.1) satisfy the Uniform Kreiss Lopatinskii Condition.

Before proving Theorem 4.2.1, let us recall that we always assume the ratio $\Delta t/\Delta x$ to be constant. This will be used several times below and is reminiscent of the scale invariance properties of the underlying continuous problem. Observe also that in (4.2.2), the larger the integer k_b the better the trace estimate on the left hand side behaves. In particular, the numerical boundary conditions in (4.2.1) involve the values $(u_{J+\ell}^n)_{1-k_b\leq\ell\leq p}$ and we get "for free" in (4.2.2) not only the control of those terms but also the extra control of $(u_{J+\ell}^n)_{1-r-k_b\leq\ell\leq -k_b}$ (recall $r\geq 1$). The fact that (4.2.2) implies the Uniform Kreiss Lopatinskii Condition is not our main focus here, so instead of recalling many definitions, we rather refer the interested reader to the review [17].

Proof. We shall use Assumption 4.1.1 in the proof below only for $k=1$, that is, we make the "minimal" consistency requirements for the scheme (4.1.7). Unlike [36, 56], the proof of Theorem 4.2.1 is done by induction with respect to the index $k_b\in\mathbb{N}$ and relies on the energy method. Let us start with the case $k_b=0$, which corresponds to Dirichlet numerical boundary conditions.

The case $k_b=0$: we consider the numerical scheme

$$\begin{cases} u_j^0 = f_j, & j\leq J \\ u_{J+\ell}^n = h_{J+\ell}^n\,, & n\in\mathbb{N}\,,\quad 1\leq\ell\leq p\,, \\ u_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, u_{j+\ell}^n\,, & n\in\mathbb{N}\,,\quad j\leq J\,, \end{cases} \tag{4.2.3}$$

A straightforward proof of the stability estimate (4.2.2) for $k_b=0$ was achieved in [21] (even in some cases of multidimensional systems), see also [38] for an earlier general result based on the theory of [43]. We reproduce here the short proof of (4.2.2) for the scheme (4.2.3) for the sake of completeness. Let us now consider the solution $(u_j^n)_{j\leq J+p,\,n\in\mathbb{N}}$ to (4.2.3) at some time index $n\in\mathbb{N}$. We extend the sequence $(u_j^n)_{j\leq J+p}$ by 0 for $j\geq J+p+1$ and still denote $u^n\in\ell^2(\mathbb{Z})$ the resulting sequence. Let us then define

$$\forall\, j\in\mathbb{Z},\quad v_j^{n+1} := \sum_{\ell=-r}^{p} a_\ell u_{j+\ell}^n, \tag{4.2.4}$$

so that $v_j^{n+1}=u_j^{n+1}$ for $j\leq J$ and $v_j^{n+1}=0$ if $j\geq J+p+r+1$. Observe that, due to the boundary conditions in (4.2.3), we do not necessarily have $v_j^{n+1}=u_j^{n+1}$ for $J+1\leq j\leq J+p$, nor for $J+p+1\leq j\leq J+p+r$ (extending also $(u_j^{n+1})_{j\leq J+p}$ by 0 for $j\geq J+p+1$). Now, we can see that

$$\sum_{j\in\mathbb{Z}}\Delta x(v_j^{n+1})^2 = \sum_{j\leq J}\Delta x(u_j^{n+1})^2 + \sum_{j=J+1}^{J+p+r}\Delta x(v_j^{n+1})^2. \tag{4.2.5}$$

On the other hand, by an application of the Plancherel's theorem, we have

$$\begin{aligned}\sum_{j\in\mathbb{Z}}(v_j^{n+1})^2 &= \frac{1}{2\pi}\int_0^{2\pi}\left|\widehat{v}^{n+1}(\xi)\right|^2 d\xi \\ &= \frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{j\in\mathbb{Z}} e^{-ij\xi}v_j^{n+1}\right|^2 d\xi,\end{aligned}$$

where $\widehat{v}^{n+1}$ is the Fourier transform of $(v_j^{n+1})_{j\in\mathbb{Z}}$. Besides, from the definition of $(v_j^{n+1})_{j\in\mathbb{Z}}$ in (4.2.4), the above formula becomes

$$\begin{aligned}\sum_{j\in\mathbb{Z}}(v_j^{n+1})^2 &= \frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{j\in\mathbb{Z}}e^{-ij\xi}\sum_{\ell=-r}^{p}a_\ell u_{j+\ell}^n\right|^2 d\xi = \frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{\ell=-r}^{p}a_\ell e^{i\ell\xi}\sum_{j\in\mathbb{Z}}e^{-i(j+\ell)\xi}u_{j+\ell}^n\right|^2 d\xi\\ &= \frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{\ell=-r}^{p}a_\ell e^{i\ell\xi}\widehat{u}^n(\xi)\right|^2 d\xi = \frac{1}{2\pi}\int_0^{2\pi}\left|\sum_{\ell=-r}^{p}a_\ell e^{i\ell\xi}\right|^2 \left|\widehat{u}^n(\xi)\right|^2 d\xi.\end{aligned} \tag{4.2.6}$$

Under Assumption 4.1.1 and then use the Plancherel's theorem, the formula (4.2.6) can be represented as

$$\sum_{j\in\mathbb{Z}}(v_j^{n+1})^2 \le \frac{1}{2\pi}\int_0^{2\pi}\left|\widehat{u}^n(\xi)\right|^2 d\xi = \sum_{j\in\mathbb{Z}}(u_j^n)^2 = \sum_{j\le J+p}(u_j^n)^2. \tag{4.2.7}$$

According to (4.2.5) and (4.2.7), we can see that

$$\sum_{j\le J}\Delta x(u_j^{n+1})^2 + \sum_{j=J+1}^{J+p+r}\Delta x(v_j^{n+1})^2 \le \sum_{j\le J+p}\Delta x(u_j^n)^2.$$

Equivalently, talking the boundary conditions of (4.2.3) into account, we get

$$\begin{aligned}&\sum_{j\le J}\Delta x(u_j^{n+1})^2 + \sum_{\ell=1}^{p+r}\Delta x(v_{J+\ell}^{n+1})^2 \le \sum_{j\le J}\Delta x(u_j^n)^2 + \sum_{\ell=1}^{p}\Delta x(h_{J+\ell}^n)^2\\ \Leftrightarrow &\underbrace{\sum_{j\le J}\Delta x(u_j^{n+1})^2 - \sum_{j\le J}\Delta x(u_j^n)^2}_{\text{Discrete time derivative}} + \underbrace{\sum_{\ell=1}^{p+r}\Delta x(v_{J+\ell}^{n+1})^2}_{\text{Trace term}} \le \underbrace{\sum_{\ell=1}^{p}\Delta x(h_{J+\ell}^n)^2}_{\text{Source term}}.\end{aligned} \tag{4.2.8}$$

We now derive a bound from below for the trace term arising on the left hand side of (4.2.8). The real numbers $(v_{J+\ell}^{n+1})_{1\le\ell\le p+r}$, depend linearly on $(u_{J+\ell}^n)_{1-r\le\ell\le p}$. The coefficients in each linear combination are taken among the a_ℓ's. Hence the quantity

$$\sum_{\ell=1}^{p+r}(v_{J+\ell}^{n+1})^2$$

can be seen as a non-negative quadratic form in the variables $(u_{J+\ell}^n)_{1-r\le\ell\le p}$. It is also rather easy to see that this quadratic form is positive definite for we have $v_{J+p+r}^{n+1} = a_{-r}u_{J+p}^n$ and therefore, if $(v_{J+\ell}^{n+1})_{1\le\ell\le p+r} \equiv 0$ then we first have $u_{J+p}^n = 0$ and recursively we can also show $(u_{J+\ell}^n)_{p-1\le\ell\le 1-r} \equiv 0$. Hence there exists a fixed constant $c_0 > 0$, that only depends on the (fixed) coefficients a_ℓ in (4.1.7) such that

$$\sum_{\ell=1}^{p+r}(v_{J+\ell}^{n+1})^2 \ge c_0 \sum_{\ell=1-r}^{p}(u_{J+\ell}^n)^2.$$

Reporting in (4.2.8) and using that $\Delta t/\Delta x = \lambda$ is a fixed positive constant, we get for some constant $c > 0$ ($c = c_0/\lambda$ is suitable):

$$\sum_{j\le J}\Delta x(u_j^{n+1})^2 - \sum_{j\le J}\Delta x(u_j^n)^2 + c\Delta t\sum_{\ell=1-r}^{p}(u_{J+\ell}^n)^2 \le \frac{1}{\lambda}\Delta t\sum_{\ell=1}^{p}(h_{J+\ell}^n)^2 \tag{4.2.9}$$

We now apply the following discrete Gronwall type lemma (with the positive parameter $\Gamma := \gamma\Delta t$), see [21] for repeated use of such summation arguments.

Lemma 4.2.2. *Let $(\mathcal{G}_n)_{n\geq 0}$ be a sequence of non-negative real numbers such that*

$$\forall \Gamma > 0, \qquad \sum_{n\geq 0} e^{-2\Gamma n} \mathcal{G}_n < +\infty.$$

Let $(\mathcal{U}_n)_{n\geq n}$, $(\mathcal{B}_n)_{n\geq n}$ be two sequences of non-negative real numbers such that

$$\forall n \in \mathbb{N}, \qquad \mathcal{U}_{n+1} - \mathcal{U}_n + \mathcal{B}_n \leq \mathcal{G}_n.$$

Then there holds for all $\Gamma > 0$

$$\sup_{n\in\mathbb{N}} e^{-2\Gamma n} \mathcal{U}_n + \sum_{n\geq 0} e^{-2\Gamma n} \mathcal{B}_n \leq \mathcal{U}_0 + \sum_{n\geq 0} e^{-2\Gamma n} \mathcal{G}_n.$$

The proof of Lemma 4.2.2 is straightforward and therefore omitted. We apply Lemma 4.2.2 to (4.2.9) and derive the estimate

$$\sup_{n\in\mathbb{N}} \left(e^{-2\gamma n\Delta t} \sum_{j\leq J} \Delta x \left(u_j^n\right)^2 \right) + \sum_{n\geq 0} \Delta t e^{-2\gamma n\Delta t} \sum_{\ell=1-r}^{p} \left(u_{J+\ell}^n\right)^2$$
$$\leq C \left\{ \sum_{j\leq J} \Delta x (f_j)^2 + \sum_{n\geq 0} \Delta t e^{-2\gamma n\Delta t} \sum_{\ell=1}^{p} (h_{J+\ell}^n)^2 \right\}, \tag{4.2.10}$$

which is (4.2.2) for $k_b = 0$.

We emphasize that when dealing with the case $k_b = 0$, we have only used the stability condition (4.1.11) of Assumption 4.1.1, and we have never used (4.1.10) (not even for $m = 0$). This is consistent with the result of [38] which proves that the Dirichlet boundary condition satisfies the Uniform Kreiss Lopatinskii Condition independently of the nature of the boundary (inflow or outflow).

The induction argument: We now assume that the stability estimate (4.2.2) is valid up to some index $k_b \in \mathbb{N}$, and consider the following numerical scheme

$$\begin{cases} u_j^0 = f_j, & j \leq J \\ (D_-^{k_b+1} u^n)_{J+\ell} = h_{J+\ell}^n\,, & n \in \mathbb{N}\,, \quad 1 \leq \ell \leq p\,, \\ u_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell \, u_{j+\ell}^n\,, & n \in \mathbb{N}\,, \quad j \leq J\,. \end{cases} \tag{4.2.11}$$

We use the induction assumption by defining the following sequence

$$\forall n \in \mathbb{N}, \qquad \forall j \leq J + p, \qquad w_j^n := (D_- u^n)_j = u_j^n - u_{j-1}^n,$$

which satisfies

$$\begin{cases} w_j^0 = f_j - f_{j-1}, & j \leq J \\ (D_-^{k_b+1} w^n)_{J+\ell} = h_{J+\ell}^n - h_{J+\ell-1}^n\,, & n \in \mathbb{N}\,, \quad 1 \leq \ell \leq p\,, \\ w_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell \, w_{j+\ell}^n\,, & n \in \mathbb{N}\,, \quad j \leq J\,. \end{cases}$$

Applying the stability estimate (4.2.2) for the index k_b and omitting one of the two non-negative terms on the left hand side of (4.2.2), we have already derived the preliminary estimate

$$\begin{aligned}\sum_{n\geq 0}\Delta t e^{-2\gamma n\Delta t}\sum_{\ell=1-r-k_b}^{p}(u_{J+\ell}^n-u_{J+\ell-1}^n)^2 &\leq C\left\{\sum_{j\leq J}\Delta x(f_j-f_{j-1})^2+\sum_{n\geq 0}\Delta t e^{-2\gamma n\Delta t}\sum_{\ell=1}^{p}(h_{J+\ell}^n)^2\right\}\\ &\leq C\left\{\sum_{j\leq J}\Delta x(f_j)^2+\sum_{n\geq 0}\Delta t e^{-2\gamma n\Delta t}\sum_{\ell=1}^{p}(h_{J+\ell}^n)^2\right\}.\end{aligned} \tag{4.2.12}$$

Let us now turn back to the numerical scheme (4.2.11) to which we are going to apply the so-called energy method. For a given time index $n\in\mathbb{N}$, we compute

$$\sum_{j\leq J}\Delta x(u_j^{n+1})^2-\sum_{j\leq J}\Delta x(u_j^n)^2+\frac{a}{2}\Delta t(u_J^n)^2\leq C\Delta t\sum_{\ell=2-r}^{p}(u_{J+\ell}^n-u_{J+\ell-1}^n)^2. \tag{4.2.13}$$

For example, to simplify the proof, suppose first that we are dealing with the Lax-Wendroff scheme

$$u_j^{n+1}=\frac{1}{2}\nu(\nu+1)u_{j-1}^n+(1-\nu^2)u_j^n+\frac{1}{2}\nu(\nu-1)u_{j+1}^n,$$

where $\nu=\lambda a$. Numerical simulations even predict that the rate of convergence should be 2, or at least close to 2. As a consequence, $k_b\in\{0,1,2\}$. Following Lemma 4.5.1, we claim that there exist a real number $B\in\mathbb{R}$ and a quadratic form Q on $\mathbb{R}^2$ such that

$$|u_j^{n+1}|^2-|u_j^n|^2=B(u_{j+1}^n-2u_j^n+u_{j-1}^n)^2+Q(u_j^n,u_{j+1}^n-u_j^n)-Q(u_{j-1}^n,u_j^n-u_{j-1}^n) \tag{4.2.14}$$

where

$$\begin{aligned}B&=\frac{1}{4}\nu^2(\nu^2-1),\\ Q(x,y)&=\alpha x^2+\beta xy+\gamma y^2,\\ \alpha&=-\nu,\quad \beta=-\nu(1-\nu),\quad \gamma=\frac{1}{2}\nu^2(1-\nu).\end{aligned}$$

Summing (4.2.14) over $j\leq J$, we get

$$\sum_{j\leq J}\Delta x(u_j^{n+1})^2-\sum_{j\leq J}\Delta x(u_j^n)^2\leq B\sum_{j\leq J}\Delta x(u_{j+1}^n-2u_j^n+u_{j-1}^n)^2+\Delta x Q(u_J^n,u_{J+1}^n-u_J^n).$$

Under the CFL condition, one has $B\leq 0$. Thus, the previous inequality becomes

$$\sum_{j\leq J}\Delta x(u_j^{n+1})^2-\sum_{j\leq J}\Delta x(u_j^n)^2\leq\Delta x Q(u_J^n,u_{J+1}^n-u_J^n).$$

By using the Young's inequality, we yield

$$\sum_{j\leq J}\Delta x(u_j^{n+1})^2-\sum_{j\leq J}\Delta x(u_j^n)^2+a(u_J^n)^2\Delta t\leq C(u_{J+1}^n-u_J^n)^2\Delta t. \tag{4.2.15}$$

Similarly, we have the same result (4.2.15) with the $O3$ scheme by using Lemma 4.5.2.

We now apply the summation argument of Lemma 4.2.2 to the inequality (4.2.13) and derive the estimate

$$\sup_{n\in\mathbb{N}}\left(e^{-2\gamma n\Delta t}\sum_{j\le J}\Delta x(u_j^n)^2\right)+\sum_{n\ge 0}\Delta t e^{-2\gamma n\Delta t}(u_J^n)^2 \\ \le C\left\{\sum_{j\le J}\Delta x(f_j)^2+\sum_{n\ge 0}\Delta t e^{-2\gamma n\Delta t}\sum_{\ell=2-r}^{p}(u_{J+\ell}^n-u_{J+\ell-1}^n)^2\right\}.$$

We then combine the latter inequality with the preliminary estimate (4.2.12), which yields

$$\sup_{n\in\mathbb{N}}\left(e^{-2\gamma n\Delta t}\sum_{j\le J}\Delta x(u_j^n)^2\right)+\sum_{n\ge 0}\Delta t e^{-2\gamma n\Delta t}(u_J^n)^2 \\ \le C\left\{\sum_{j\le J}\Delta x(f_j)^2+\sum_{n\ge 0}\Delta t e^{-2\gamma n\Delta t}\sum_{\ell=1}^{p}(h_{J+\ell}^n)^2\right\}. \tag{4.2.16}$$

At this stage, we have almost proved that (4.2.2) holds up to the index k_b+1. Indeed, if we combine the trace estimates provided by (4.2.12) and (4.2.16), we get a full control of the trace of (u^n), that is of $(u_{J+\ell}^n)_{-r-k_b\le\ell\le p,\, n\ge 0}$

$$\sum_{n\ge 0}\Delta t e^{-2\gamma n\Delta t}\sum_{\ell=-r-k_b}^{p}(u_{J+\ell}^n)^2\le C\left\{\sum_{j\le J}\Delta x(f_j)^2+\sum_{n\ge 1}\Delta t e^{-2\gamma n\Delta t}\sum_{\ell=1}^{p}(h_{J+\ell}^n)^2\right\}. \tag{4.2.17}$$

Combining with (4.2.16), we have completed the proof of (4.2.2) for the index k_b+1, which also the proof of Theorem 4.2.1. □

4.2.2 Convergence estimates for the outflow problem

In the previous paragraph, we have proved the stability estimate (4.2.2) in order to highlight the fact that our method automatically yields the verification of the Uniform Kreiss Lopatinskii Condition. However, the exponential weights arising in (4.2.2) and the fact that no “interior” source term is considered in (4.2.1) make this estimate hardly applicable as such in view of the convergence analysis below. We therefore state a slightly weakened but more practical version of Theorem 4.2.1 which will help us proving Theorem 4.2.4 below.

Proposition 4.2.3. *Let $a>0$, $k\in\mathbb{N}^*$ and $k_b\in\mathbb{N}$. Under Assumption 4.1.1, there exists a constant $C>0$ such that for any $N\in\mathbb{N}^*$, any $J\in\mathbb{N}^*$, all initial data $(f_j)_{j\le J}\in\ell^2$, all boundary source terms $(h_{J+\ell}^n)_{1\le\ell\le p,\,0\le n\le N-1}$ and for all interior source terms $(F_j^n)_{j\le J,\,1\le n\le N}\in\ell^2$, the solution $(u_j^n)_{j\le J,\,0\le n\le N}$ to the numerical scheme*

$$\begin{cases} u_j^0 = f_j\,, & j\le J\,,\\ (D_-^{k_b}u^n)_{J+\ell} = h_{J+\ell}^n\,, & 0\le n\le N-1\,,\quad 1\le\ell\le p\,,\\ u_j^{n+1}=\displaystyle\sum_{\ell=-r}^{p}a_\ell\,u_{j+\ell}^n+\Delta t F_j^{n+1}\,, & 0\le n\le N-1\,,\quad j\le J\end{cases} \tag{4.2.18}$$

satisfies

$$\sup_{0\le n\le N}\sum_{j\le J}\Delta x(u_j^n)^2\le C\left(\sum_{j\le J}\Delta x(f_j)^2+(N\Delta t)^2\sup_{1\le n\le N}\sum_{j\le J}\Delta x(F_j^n)^2+\sum_{n=0}^{N-1}\Delta t\sum_{\ell=1}^{p}(h_{J+\ell}^n)^2\right). \tag{4.2.19}$$

Proof. By linearity of (4.2.18), it is sufficient to examine separately the cases $F \equiv 0$ (no interior source term), and $f \equiv 0$, $h \equiv 0$ (interior forcing only). In the case $F \equiv 0$ the estimate (4.2.19) is directly obtained by

- first extending the boundary source terms $(h^n_{J+\ell})_{1\leq \ell \leq p}$ by 0 for $n > N-1$, which does not affect the solution to (4.2.18) for $j \leq J$ at time steps earlier than N,
- then passing to the limit $\gamma \to 0$ in (4.2.2) (and forgetting about the non-negative trace estimate on the left hand side of (4.2.2)).

Thus, we can obtain

$$\sup_{0\leq n\leq N} \sum_{j\leq J} \Delta x (u^n_j)^2 \leq C \left\{ \sum_{j\leq J} \Delta x (f_j)^2 + \sum_{n=0}^{N-1} \Delta t \sum_{\ell=1}^{p} (h^n_{J+\ell})^2 \right\} .$$

In the case $f \equiv 0$, $h \equiv 0$, the solution to (4.2.18) can be written with the Duhamel formula

$$\forall\, 0 \leq n \leq N, \quad u^n = \sum_{m=1}^{n} \Delta t \mathcal{S}^{n-m} F^m$$

where the generator of the discrete semi-group $(\mathcal{S}^n)_{n\in\mathbb{N}}$ is power bounded on $\ell^2((-\infty, J))$. Therefore, we end up with

$$\sup_{0\leq n\leq N} \sum_{j\leq J} \Delta x (u^n_j)^2 \leq C \sum_{n=1}^{N} \Delta t \left(\sum_{j\leq J} \Delta x (F^n_j)^2 \right)^{1/2} \leq C N \Delta t \sup_{1\leq n\leq N} \left(\sum_{j\leq J} \Delta x (F^n_j)^2 \right)^{1/2} .$$

This completes the proof of (4.2.19). □

We are now ready to prove the convergence result for the outflow boundary

Theorem 4.2.4 (Convergence estimate for the outflow problem [22]). *Let $a > 0$, $k \in \mathbb{N}^\star$ and $k_b \in \mathbb{N}$. Under Assumption 4.1.1, there exists a constant $C > 0$ such that for any final time $T \geq 1$, for any $J \in \mathbb{N}^*$ and for any initial condition $f \in H^{k+1}((-\infty, L))$, the solution $(u^n_j)_{j\leq J, 0\leq n\leq T/\Delta t}$ to the numerical scheme*

$$\begin{cases} u^0_j = \dfrac{1}{\Delta x} \displaystyle\int_{x_{j-1}}^{x_j} f(x)\, \mathrm{d}x \,, & j \leq J \,, \\ (D_-^{k_b} u^n)_{J+\ell} = 0 \,, & 0 \leq n \leq T/\Delta t \,, \quad 1 \leq \ell \leq p \,, \\ u^{n+1}_j = \displaystyle\sum_{\ell=-r}^{p} a_\ell \, u^n_{j+\ell} \,, & 0 \leq n \leq T/\Delta t - 1 \,, \quad j \leq J \,, \end{cases} \tag{4.2.20}$$

satisfies

$$\sup_{0\leq n\leq T/\Delta t} \left(\sum_{j\leq J} \Delta x \left(u^n_j - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x - a\, t^n)\, \mathrm{d}x \right)^2 \right)^{1/2} \leq C\, T\, \Delta x^{\min(k, k_b)} \, \|f\|_{H^{k+1}((-\infty, L))} \,. \tag{4.2.21}$$

Proof. In the proof below, we assume $k_b \leq k$, so that the limiting order of convergence arises from the numerical boundary conditions in (4.2.20) and not from the discretization of the transport equation. The proof in the case $k_b > k$ is quite similar and we leave the corresponding modifications to the interested reader. Since the validity of Assumption 4.1.1 for some integer

$k \geq 1$ implies the validity of the relations (4.1.10) for the restricted subset of indices $0 \leq m \leq k_b$, we can even assume without loss of generality $k_b = k$.

We now denote by $f_\sharp$ the extension of f as a function in $H^{k+1}(\mathbb{R})$ by the linear continuous reflexion operator of [27] and then define

$$\forall\, j \in \mathbb{Z}, \quad w_j^n := \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f_\sharp(x - at^n)dx \tag{4.2.22}$$

to be the cell average of the exact solution $((t,x) \mapsto f_\sharp(x-at))$ to the transport equation over $\mathbb{R}$. With $(u_j^n)_{j \leq J+p,\, 0 \leq n \leq T/\Delta t}$ the solution to the numerical scheme (4.2.20), we define the error $\varepsilon_j^n := u_j^n - w_j^n$, that is a solution to

$$\begin{cases} \varepsilon_j^n = 0\,, & j \leq J\,, \\ (D_-^{k_b}\varepsilon^n)_{J+\ell} = -(D_-^{k_b}w^n)_{J+\ell}\,, & 0 \leq n \leq T/\Delta t - 1, \quad 1 \leq \ell \leq p\,, \\ \varepsilon_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell \varepsilon_{j+\ell}^n + \Delta t e_j^{n+1}, & 0 \leq n \leq T/\Delta t - 1\,, \quad j \leq J\,. \end{cases} \tag{4.2.23}$$

Let us remark that the interior consistency error $(e_j^n)_{j \leq J,\, 0 \leq n \leq T/\Delta t}$ in (4.2.23) is given by

$$e_j^n := -\frac{1}{\Delta t}\left(w_j^n - \sum_{\ell=-r}^{p} a_\ell w_{j+\ell}^{n-1}\right), \tag{4.2.24}$$

which is easily estimated by means of the Cauchy-Schwarz inequality

$$\begin{aligned} \sum_{j \leq J} \Delta x \left(e_j^n\right)^2 &= \frac{\Delta x}{\Delta t^2} \sum_{j \leq J} \left(w_j^n - \sum_{\ell=-r}^{p} a_\ell w_{j+\ell}^{n-1}\right)^2 \\ &= \frac{1}{\Delta x \Delta t^2} \sum_{j \leq J} \left(\int_{x_{j-1}}^{x_j} \left(f_\sharp\left(x - at^n - a\Delta t\right) - \sum_{\ell=-r}^{p} a_\ell f_\sharp\left(x - at^n + \ell\Delta x\right)\right) dx\right)^2 \\ &\leq \frac{1}{\Delta t^2} \int_{\mathbb{R}} \left(f_\sharp\left(x - at^n - a\Delta t\right) - \sum_{\ell=-r}^{p} a_\ell f_\sharp\left(x - at^n + \ell\Delta x\right)\right)^2 dx. \end{aligned} \tag{4.2.25}$$

By an application of the Plancherel theorem and the shifts property of Fourier analysis, we get

$$\begin{aligned} &\int_{\mathbb{R}} \left(f_\sharp\left(x - at^n - a\Delta t\right) - \sum_{\ell=-r}^{p} a_\ell f_\sharp\left(x - at^n + \ell\Delta x\right)\right)^2 dx \\ \leq &\frac{1}{2\pi} \int_{\mathbb{R}} \left| e^{-ia\lambda\Delta x\xi} - \sum_{\ell=-r}^{p} a_\ell e^{i\ell\Delta x\xi} \right|^2 \left|\widehat{f_\sharp}(\xi)\right|^2 d\xi. \end{aligned} \tag{4.2.26}$$

Besides, if $|\Delta x\xi| > 1$ then

$$\left| e^{-ia\lambda\Delta x\xi} - \sum_{\ell=-r}^{p} a_\ell e^{i\ell\Delta x\xi} \right| \leq \left|e^{-ia\lambda\Delta x\xi}\right| + \sum_{\ell=-r}^{p} |a_\ell| \left|e^{i\ell\Delta x\xi}\right| \leq \left(1 + \sum_{\ell=-r}^{p} |a_\ell|\right) |\Delta x\xi|^{k+1}.$$

Otherwise, if $|\Delta x\xi| \leq 1$, we first use Taylor expansion

$$\left| e^{-ia\lambda\Delta x\xi} - \sum_{\ell=-r}^{p} a_\ell e^{i\ell\Delta x\xi} \right| = \left| \sum_{n=0}^{+\infty} \frac{(\Delta x\xi)^n}{n!} \left((-\lambda a)^n - \sum_{\ell=-r}^{p} a_\ell \ell^n\right) \right|.$$

Then,

$$\left|\sum_{n=0}^{+\infty}\frac{(\Delta x\xi)^n}{n!}\left((-\lambda a)^n-\sum_{\ell=-r}^{p}a_\ell\ell^n\right)\right| = |\Delta x\xi|^{k+1}\left|\sum_{n=0}^{+\infty}\frac{(\Delta x\xi)^{n-k-1}}{n!}\left((-\lambda a)^n-\sum_{\ell=-r}^{p}a_\ell\ell^n\right)\right|$$
$$\leq |\Delta x\xi|^{k+1}\left|\sum_{n=0}^{+\infty}\frac{1}{n!}\left((-\lambda a)^n-\sum_{\ell=-r}^{p}a_\ell\ell^n\right)\right| \leq |\Delta x\xi|^{k+1}\sum_{n=0}^{+\infty}\frac{1}{n!}\left|(-\lambda a)^n-\sum_{\ell=-r}^{p}a_\ell\ell^n\right|$$
$$\leq |\Delta x\xi|^{k+1}\sum_{n=0}^{+\infty}\left(\frac{1}{n!}+\frac{\max(p,r)^n}{n!}\sum_{\ell=-r}^{p}|a_\ell|\right) \leq |\Delta x\xi|^{k+1}\left(e+e^{\max(p,r)}\sum_{\ell=-r}^{p}|a_\ell|\right).$$

Therefore, we can conclude that there exists $C>0$ such that

$$\left|e^{-ia\lambda\Delta x\xi}-\sum_{\ell=-r}^{p}a_\ell e^{i\ell\Delta x\xi}\right| \leq C\,(\Delta x|\xi|)^{k_b+1}\,. \tag{4.2.27}$$

Plugging (4.2.27) and (4.2.26) into (4.2.25) and then using the Plancherel's theorem, we have

$$\sum_{j\leq J}\Delta x(e_j^n)^2 \leq C\frac{\Delta x^{2k_b+2}}{\Delta t^2}\times\frac{1}{2\pi}\int_{\mathbb{R}}|\widehat{f_\sharp}(\xi)|^2d\xi = C\frac{\Delta x^{2k_b+2}}{\Delta t^2}\int_{\mathbb{R}}|f_\sharp(\xi)|^2d\xi. \tag{4.2.28}$$

Recalling that the ratio $\Delta t/\Delta x$ is constant and going back the definition of $f_\sharp$, we have obtained the bound

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(e_j^n)^2 \leq C\Delta x^{2k_b}\|f\|^2_{H^{k_b+1}((-\infty,L))}. \tag{4.2.29}$$

The boundary errors in (4.2.23) are dealt with by the following elementary result

Lemma 4.2.5. *Let $m,p\in\mathbb{N}$. There exists a constant $C>0$ such that for any $\Delta x>0$ and for any $v\in H^{m+1}((-\infty,L+p\Delta x))$, defining*

$$v_j := \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j}v(y)dy, \qquad j\leq J+p,$$

one has

$$|(D_-^m v)_{J+\ell}| \leq C\Delta x^m\|v^{(m)}\|_{H^1((-\infty,L+p\Delta x))}, \qquad 1\leq\ell\leq p.$$

Proof. For $1\leq\ell\leq p$, there holds

$$(D_-^m v)_{J+\ell} = \sum_{m'=0}^{m}\binom{m}{m'}(-1)^{m-m'}\frac{1}{\Delta x}\int_{x_{J-m-1+\ell}}^{x_{J-m+\ell}}v(y+m'\Delta x)dy$$
$$= \sum_{m'=0}^{m}\binom{m}{m'}\frac{(-1)^{m-m'}}{(m-1)!}\times\frac{1}{\Delta x}\int_{x_{J-m-1+\ell}}^{x_{J-m+\ell}}\int_y^{y+m'\Delta x}v^{(m)}(z)(y+m'\Delta x-z)^{m-1}dzdy,$$

where we used Taylor's formula and cancellation properties of the binomial coefficients. Applying the Cauchy-Schwarz inequality to each double integral in the latter expression, we get

$$\int_{x_{J-m-1+\ell}}^{x_{J-m+\ell}}\int_y^{y+m'\Delta x}v^{(m)}(z)(y+m'\Delta x-z)^{m-1}dzdy$$
$$\leq \left(\int_{x_{J-m-1+\ell}}^{x_{J-m+\ell}}\int_y^{y+m'\Delta x}v^{(m)}(z)^2(y+m'\Delta x-z)^{2m-2}dzdy\times\int_{x_{J-m-1+\ell}}^{x_{J-m+\ell}}\int_y^{y+m'\Delta x}dzdy\right)^{1/2}$$
$$= (m')^{1/2}\Delta x\left(\int_{x_{J-m-1+\ell}}^{x_{J-m+\ell}}\int_y^{y+m'\Delta x}v^{(m)}(z)^2(y+m'\Delta x-z)^{2m-2}dzdy\right)^{1/2}.$$

Thus, we obtain

$$\begin{aligned}|(D_-^m v)_{J+\ell}| &\le C \sum_{m'=0}^{m} \left(\int_{x_{J-m-1+\ell}}^{x_{J-m+\ell}} \int_{y}^{y+m'\Delta x} v^{(m)}(z)^2 (y+m'\Delta x - z)^{2m-2} dz dy \right)^{1/2} \\ &\le C\Delta x^m \|v^{(m)}\|_{L^\infty((-\infty,L+p\Delta x))}.\end{aligned}$$

By using the imbedding of H^1 in L^∞ in one space dimension, there exits $C > 0$ such that

$$\|v^{(m)}\|_{L^\infty((-\infty,L+p\Delta x))} \le C\|v^{(m)}\|_{H^1((-\infty,L+p\Delta x))}.$$

This ends the proof of Lemma 4.2.5. □

We now apply Lemma 4.2.5 with $m = k_b$ to evaluate the boundary errors in (4.2.23). We get

$$\begin{aligned}\sum_{n=0}^{T/\Delta t-1} \Delta t \sum_{\ell=1}^{p} ((D_-^{k_b}\varepsilon^n)_{J+\ell})^2 &= \sum_{n=0}^{T/\Delta t-1} \Delta t \sum_{\ell=1}^{p} ((D_-^{k_b}w^n)_{J+\ell})^2 \\ &\le CT\Delta x^{2k_b} \|f_\sharp(\cdot - at^n)\|^2_{H^{k_b+1}((-\infty,L+p\Delta x))} \\ &\le CT\Delta x^{2k_b} \|f\|^2_{H^{k_b+1}((-\infty,L))}\end{aligned} \tag{4.2.30}$$

thanks to the continuity of the reflexion operator. We now apply the result in Proposition 4.2.3 for the error problem (4.2.23), we have already get

$$\sup_{0\le n\le T/\Delta t} \sum_{j\le J} \Delta x (e_j^n)^2 \le C \left\{ T^2 \sup_{1\le n\le T/\Delta t} \sum_{j\le J} \Delta x (e_j^n)^2 + \sum_{n=0}^{T/\Delta t-1} \Delta t \sum_{\ell=1}^{p} \left(\left(D_-^{k_b} w^n \right)_{J+\ell} \right)^2 \right\}. \tag{4.2.31}$$

Going back to (4.2.29)-(4.2.30) and using the estimate (4.2.31), we finally estimate

$$\sup_{0\le n\le T/\Delta t} \sum_{j\le J} \Delta x (\varepsilon_j^n)^2 \le CT\Delta x^{2k_b} \|f\|^2_{H^{k_b+1}((-\infty,L))},$$

which completes the proof of Theorem 4.2.4. □

4.2.3 Convergence analysis on a half-line for the inflow problem

In this paragraph, we consider the half-space $(0, +\infty)$ with the Dirichlet boundary condition (4.1.12) at the inflow boundary condition $x = 0$. The ghost cells correspond to the intervals (x_{-r}, x_{1-r}), ..., (x_{-1}, x_0), see Figure 4.2.2.

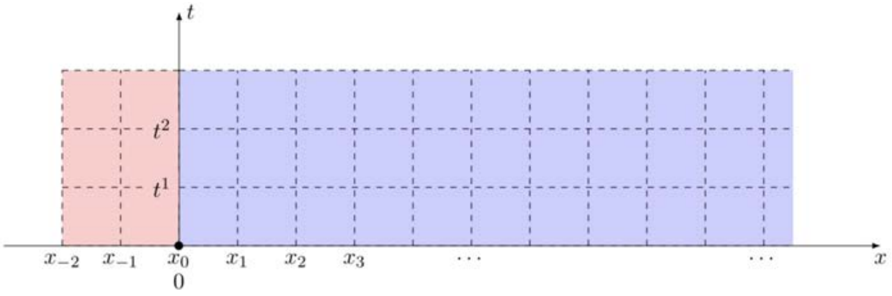

Figure 4.2.2: The mesh on $\mathbb{R}^+$ in blue and the “ghost cells” in red ($r = 2$ here).

We focus here on the inflow source term, and therefore start by proving the main convergence estimate that is the new ingredient for the proof of Theorem 4.1.2.

Theorem 4.2.6 (Convergence estimate for the inflow problem)**.** *Let $a > 0$, $k \in \mathbb{N}^\star$. Under Assumption 4.1.1, there exists a constant $C > 0$ such that for any final time $T \geq 1$, any $J \in \mathbb{N}^*$, any initial condition $f \in H^{k+1}((0,+\infty))$ and boundary source term $g \in H^{k+1}((0,T))$ satisfying the compatibility conditions*

$$\forall\, 0 \leq m \leq k\,, \quad f^{(m)}(0) \;=\; (-a)^{-m}\, g^{(m)}(0)\,, \tag{4.2.32}$$

the solution $(u_j^n)_{j \geq 1-r, n \in \mathbb{N}}$ to the numerical scheme

$$\begin{cases} u_j^0 \;=\; \dfrac{1}{\Delta x} \displaystyle\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x\,, & j \geq 1\,, \\ u_\ell^n \;=\; \displaystyle\sum_{\kappa=0}^{k-1} \dfrac{\Delta x^\kappa}{(\kappa+1)\,!\,(-a)^\kappa} \left(\ell^{\kappa+1} - (\ell-1)^{\kappa+1}\right) g^{(\kappa)}(t^n)\,, & 0 \leq n \leq T/\Delta t\,, \quad 1-r \leq \ell \leq 0\,, \\ u_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, u_{j+\ell}^n\,, & 0 \leq n \leq T/\Delta t - 1\,, \quad j \geq 1\,, \end{cases} \tag{4.2.33}$$

satisfies

$$\sup_{0 \leq n \leq T/\Delta t} \left(\sum_{j \leq J} \Delta x \left(u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} u(t^n,x)\,\mathrm{d}x \right)^2 \right)^{1/2} \leq C\,T\;\Delta x^k \left(\|f\|_{H^{k+1}((0,+\infty))} + \|g\|_{H^{k+1}((0,T))} \right),$$

where u is the exact solution to the half-line transport problem

$$\begin{cases} \partial_t u + a\,\partial_x u \;=\; 0\,, & t \in (0,T)\,,\, x \geq 0\,, \\ u(0,x) \;=\; f(x)\,, & x \geq 0\,, \\ u(t,0) \;=\; g(t)\,, & t \in (0,T)\,. \end{cases} \tag{4.2.34}$$

Proof. For convenience, we first extend g into a function $g_\flat \in H^{k+1}((0,+\infty))$ and then define

$$\forall\, x \in \mathbb{R}\,, \quad f_\flat(x) \;:=\; \begin{cases} f(x)\,, & \text{if } x > 0, \\ g_\flat(-x/a)\,, & \text{if } x < 0. \end{cases}$$

Since f and g satisfy the compatibility conditions (4.2.32), we have $f_\flat \in H^{k+1}(\mathbb{R})$, and the exact solution u to (4.2.34) is given by

$$\forall\,(t,x) \in [0,T] \times (0,+\infty)\,, \quad u(t,x) \;=\; f_\flat(x - a\,t)\,.$$

Let us now define

$$\forall\, j \in \mathbb{Z}\,, \quad \forall\, n \in \mathbb{N}\,, \quad w_j^n \;:=\; \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f_\flat(x - a\,t^n)\,\mathrm{d}x\,,$$

which corresponds to the cell average of the exact solution to (4.2.34). With $(u_j^n)_{j \geq 1-r, 0 \leq n \leq T/\Delta t}$ the solution to the numerical scheme (4.2.33), we define the error $\varepsilon_j^n := u_j^n - w_j^n$, that is a solution to

$$\begin{cases} \varepsilon_j^0 \;=\; 0\,, & j \geq 1\,, \\ \varepsilon_\ell^n \;=\; u_\ell^n - w_\ell^n\,, & 0 \leq n \leq T/\Delta t\,, \quad 1-r \leq \ell \leq 0\,, \\ \varepsilon_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, \varepsilon_{j+\ell}^n + \Delta t\, e_j^{n+1}\,, & 0 \leq n \leq T/\Delta t - 1\,, \quad j \geq 1\,. \end{cases} \tag{4.2.35}$$

Let us remark that the interior consistency error $(e_j^{n+1})_{j\geq 1, 0\leq n\leq T/\Delta t-1}$ in (4.2.35) is given by

$$e_j^n := -\frac{1}{\Delta t}\left(w_j^n - \sum_{\ell=-r}^{p} a_\ell w_{j+\ell}^{n-1}\right),$$

One can proceed as in (4.2.25)-(4.2.28) to have the following inequality

$$\sum_{j\geq 1} \Delta x \left(e_j^{n+1}\right)^2 \leq C\frac{\Delta x^{2k+2}}{\Delta t^2}\int_{\mathbb{R}} |f_\flat(x)|^2 dx.$$

Recalling that the ratio $\Delta t/\Delta x$ is constant and going back to the definition of $f_\flat$, we have obtained the bound

$$\sup_{0\leq n\leq T/\Delta t-1} \left(\sum_{j\geq 1} \Delta x\,(e_j^{n+1})^2\right)^{1/2} \leq C\,\Delta x^k \left(\|f\|_{H^{k+1}((0,+\infty))} + \|g\|_{H^{k+1}((0,T))}\right), \tag{4.2.36}$$

for some constant C that is independent of the final time $T \geq 1$ and the data f and g.

We now turn to the boundary errors in (4.2.35), and wish to estimate the following quantities

$$\sum_{0\leq n\leq T/\Delta t-1} \Delta t\left(\varepsilon_\ell^n\right)^2 = \sum_{0\leq n\leq T/\Delta t-1} \Delta t\left(u_\ell^n - w_\ell^n\right)^2, \quad 1-r\leq \ell\leq 0\,.$$

Let us consider an integer n such that $0\leq n\leq T/\Delta t-1$. From the definition of w_ℓ^n, $\ell\leq 0$, we have

$$u_\ell^n - w_\ell^n = \sum_{\kappa=0}^{k-1} \frac{\Delta x^\kappa}{(\kappa+1)\,!\,(-a)^\kappa}\left(\ell^{\kappa+1} - (\ell-1)^{\kappa+1}\right) g^{(\kappa)}(t^n) - \frac{1}{\Delta x}\int_{x_{\ell-1}}^{x_\ell} g_\flat(t^n - x/a)\,\mathrm{d}x \tag{4.2.37}$$

Applying the Taylor formula[3], we can see that

$$\begin{aligned}\int_{x_{\ell-1}}^{x_\ell} g_\flat(t^n - x/a)dx &= \int_{x_{\ell-1}}^{x_\ell}\sum_{\kappa=0}^{k-1}\frac{x^\kappa}{\kappa!(-a)^\kappa} g_\flat^{(\kappa)}(t^n)dx + \int_{x_{\ell-1}}^{x_\ell}\frac{x^k}{(-a)^k}\int_0^1 \frac{y^{k-1}}{(k-1)!} g_\flat^{(k)}\left(t^n - \frac{xy}{a}\right)dydx\\ &= \sum_{\kappa=0}^{k-1}\frac{\Delta x^{\kappa+1}}{(\kappa+1)!(-a)^\kappa}\left(\ell^{\kappa+1} - (\ell-1)^{\kappa+1}\right) g_\flat^{(\kappa)}(t^n) + \int_{x_{\ell-1}}^{x_\ell}\frac{x^k}{(-a)^k}\int_0^1\frac{y^{k-1}}{(k-1)!} g_\flat^{(k)}\left(t^n - \frac{xy}{a}\right)dydx.\end{aligned} \tag{4.2.38}$$

Plugging (4.2.38) into (4.2.37), one obtains

$$u_\ell^n - w_\ell^n = -\frac{1}{\Delta x}\int_{x_{\ell-1}}^{x_\ell}\frac{x^k}{(-a)^k}\int_0^1\frac{y^{k-1}}{(k-1)!} g_\flat^{(k)}\left(t^n - \frac{xy}{a}\right)dydx.$$

By the Cauchy-Schwarz inequality, we get

$$\left(u_\ell^n - w_\ell^n\right)^2 \leq \frac{C}{\Delta x}\int_{x_{\ell-1}}^{x_\ell}\int_0^1 x^{2\,k} y^{2\,(k-1)}\, g_\flat^{(k)}\left(t^n - \frac{x\,y}{a}\right)^2 \mathrm{d}y\,\mathrm{d}x\,,$$

[3] This is precisely at this point of the analysis that the definition of the coefficients $\alpha_{\kappa,\ell}$ in the inverse Lax-Wendroff method arises. Our choice in (4.1.12) is motivated by the fact that we compare the numerical solution with the cell average of the exact solution.

and we now apply the change of variables $(x,y) \to (u,v) := (x\,y, x)$ to get

$$\left(u_\ell^n - w_\ell^n\right)^2 \,\le\, \int_{x_{\ell-1}}^{x_\ell} \int_v^0 |v|\,|u|^{2\,(k-1)}\, g_\flat^{(k)} \left(t^n - \frac{u}{a}\right)^2 \,\mathrm{d}u\,\mathrm{d}v\,.$$

Restricting to $1-r \le \ell \le 0$, we have

$$\sum_{\ell=1-r}^{0} \left(u_\ell^n - w_\ell^n\right)^2 \,\le\, C\,\Delta x^{2\,k-1} \int_0^{r\,\Delta x/a} g_\flat^{(k)}(t^n+\tau)^2\,\mathrm{d}\tau\,.$$

Summing now with respect to n, we end up with the estimate

$$\sum_{0\le n\le T/\Delta t-1} \Delta t \sum_{\ell=1-r}^{0} (\varepsilon_\ell^n)^2 \,\le\, C\,\Delta x^{2\,k} \int_0^{+\infty} g_\flat^{(k)}(t)^2\,\mathrm{d}t \,\le\, C\,\Delta x^{2\,k}\, \|g\|_{H^k((0,T))}^2\,, \tag{4.2.39}$$

for some constant C that is independent of the final time $T \ge 1$ and the data f and g.

We now apply the main stability estimate for the error problem (4.2.35), for which we refer to the seminal papers [37, 38] and to the more recent works [21, 22]

$$\sup_{0\le n\le T/\Delta t} \left(\sum_{j\ge1} \Delta x\,|\varepsilon_j^n|^2\right)^{1/2} \,\le\, C\,\left\{ T \sup_{1\le n\le T/\Delta t} \left(\sum_{j\ge1} \Delta x\,|e_j^n|^2\right)^{1/2} + \left(\sum_{0\le n\le T/\Delta t-1} \Delta t \sum_{\ell=1-r}^{0} |\varepsilon_\ell^n|^2\right)^{1/2} \right\}\,.$$

The conclusion of Theorem 4.2.6 then comes from the combination of the estimates (4.2.36) and (4.2.39). □

4.2.4 Proof of Theorem 4.1.2

4.2.4.1 Stability estimates on a finite interval

We now turn to the study of the numerical scheme (4.1.4), (4.1.7), (4.1.8), (4.1.12) with homogeneous Dirichlet condition at the inflow, which is an iteration in a finite dimensional space and therefore really corresponds to a numerical scheme that can be implemented in practice. We first prove a stability estimate for (4.1.4), (4.1.7), (4.1.8), (4.1.12) with homogeneous Dirichlet condition at the inflow, which will have various consequence

Proposition 4.2.7. *Let $a>0$, $k \in \mathbb{N}^*$ and $k_\flat \in \mathbb{N}$. Under Assumption 4.1.1, there exists a constant $c>0$ such that for all initial data $(f_j)_{j\le J}$, the solution $(u_j^n)_{1\le j\le J, n\in\mathbb{N}}$ to the numerical scheme*

$$\begin{cases} u_j^0 = f_j\,, & 1\le j\le J\,,\\ u_\ell^n = 0\,, & n\in\mathbb{N},\ 1-r\le \ell\le 0\,,\\ (D_-^{k_\flat} u^n)_{J+\ell} = 0\,, & n\in\mathbb{N},\ 1\le\ell\le p\,,\\ u_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell u_{j+\ell}^n\,, & n\in\mathbb{N},\ 1\le j\le J \end{cases} \tag{4.2.40}$$

satisfies

$$\forall\, n\in\mathbb{N},\quad \left(\sum_{j=1}^{J} \Delta x (u_j^n)^2\right)^{1/2} \le C e^{C n\Delta t/L} \left(\sum_{j=1}^{J} \Delta x (f_j)^2\right)^{1/2}\,. \tag{4.2.41}$$

Proof. The derivation of (4.2.41) follows from the finite speed of propagation of the numerical scheme (4.1.7). (Observe that our argument below does not extend to implicit discretizations of the transport equation.) More precisely, let us assume for simplicity that J is even. Let

$$N_0 := \min\left(E\left(\frac{J}{2p}\right), E\left(\frac{J/2-k_b}{r}\right)\right).$$

The case $n \leq N_0$: we first prove that the solution $(u_j^n)_{1\leq j\leq J}$ can be written as the superposition of solution to two initial boundary value problems of the form (4.2.18) and (4.2.33). More specifically, one introduces $(w_j^n)_{j\leq J, n\in\mathbb{N}}$, which is the solution to the scheme

$$\begin{cases} w_j^0 = \varphi_j, & j \leq J, \\ (D_-^{k_b} w^n)_{J+\ell} = 0, & n\in\mathbb{N},\, 1\leq \ell\leq p, \\ w_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell w_{j+\ell}^n, & n\in\mathbb{N},\, j\leq J \end{cases} \tag{4.2.42}$$

where the initial data $(\varphi_j)_{j\leq J}$ is defined by

$$\varphi_j := \begin{cases} 0, & \text{if } j\leq J/2, \\ f_j, & \text{if } J/2+1\leq j\leq J. \end{cases}$$

We can see that the support of $(w_j^n)_{j\leq J}$ is shifted of p cells to the left at each time iteration and the initial condition $(w_j^0)_{j\leq J}$ vanish for $j\leq J/2$. Therefore,

$$\forall\, n\leq N_0,\, \forall\, j\leq J/2-np, \quad w_j^n = 0.$$

Since $N_0 \leq E(J/2)$, we get $J/2-np \geq J/2-N_0p\geq 0$. Thus, the solution to (4.2.42) satisfies the homogeneous Dirichlet boundary condition

$$\forall\, n\leq N_0,\, \forall\, 1-r\leq \ell\leq 0, \quad w_\ell^n = 0. \tag{4.2.43}$$

We also set $(v_j^n)_{j\geq 1, n\in\mathbb{N}}$, which is the solution to the scheme

$$\begin{cases} v_j^0 = \psi_j, & j\geq 1, \\ v_\ell^n = 0, & n\in\mathbb{N},\, 1-r\leq \ell\leq 0, \\ v_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell v_{j+\ell}^n, & n\in\mathbb{N},\, j\geq 1 \end{cases} \tag{4.2.44}$$

with the initial data $(\psi_j)_{j\geq 1}$ is defined by

$$\psi_j := \begin{cases} f_j, & \text{if } 1\leq j\leq J/2, \\ 0, & \text{if } j\geq J/2+1. \end{cases}$$

Again, we observe that the support of $(v_j^n)_{j\geq 1}$ is shifted of r cells to the right at each time iteration and the initial condition $(v_j^0)_{j\geq 1}$ vanish for $j\geq J/2+1$. Therefore,

$$\forall\, n\leq N_0,\, \forall\, j\geq J/2+1+nr, \quad v_j^n = 0.$$

Since $N_0 \leq E\left(\dfrac{J/2-k_b}{r}\right)$, one obtains $J/2+1+nr\leq J/2+1+N_0r\leq J+1-k_b$. Therefore,

$$\forall n\leq N_0,\, \forall 1\leq \ell\leq p, \quad v_{J-k_b+\ell}^n = 0 = \ldots = 0 = v_{J+1-\ell}^n.$$

Then, the solution to (4.2.44) satisfies

$$\forall\, n \leq N_0,\, \forall 1 \leq \ell \leq p, \quad (D_-^{k_b} v^n)_{J+\ell} = 0. \tag{4.2.45}$$

According to (4.2.42)-(4.2.45), we can write the solution $(u_j^n)_{1\leq j\leq J,\, n\leq N_0}$ as the superposition of solutions to two initial boundary value problems of the form (4.2.42) and (4.2.44). It means that

$$\forall n \leq N_0,\, \forall 1 \leq j \leq J, \quad u_j^n = v_j^n + w_j^n. \tag{4.2.46}$$

Secondly, thanks to the result of Proposition 4.2.3, the solution $(w_j^n)_{j\leq J,\, 0\leq n\leq N_0}$ to the numerical scheme (4.2.42) satisfies

$$\sup_{0\leq n\leq N_0} \sum_{j\leq J} \Delta x (w_j^n)^2 \lesssim \sum_{j\leq J} \Delta x (\varphi_j)^2 = \sum_{J/2+1\leq j\leq J} \Delta x (f_j)^2. \tag{4.2.47}$$

Similarly, $(v_j^n)_{j\geq 1,\, 0\leq n\leq N_0}$ satisfies a similar estimate for the Dirichlet inflow condition

$$\sup_{0\leq n\leq N_0} \sum_{j\geq 1} \Delta x (v_j^n)^2 \lesssim \sum_{j\geq 1} \Delta x (\psi_j)^2 = \sum_{1\leq j\leq J/2} \Delta x (f_j)^2. \tag{4.2.48}$$

From the inequalities (4.2.47) and (4.2.48), one has

$$\sup_{0\leq n\leq N_0} \sum_{j\leq J} \Delta x (w_j^n)^2 + \sup_{0\leq n\leq N_0} \sum_{j\geq 1} \Delta x (v_j^n)^2 \lesssim \sum_{1\leq j\leq J} \Delta x (f_j)^2.$$

Thus,

$$\sup_{0\leq n\leq N_0} \sum_{1\leq j\leq J} \Delta x (w_j^n)^2 + \sup_{0\leq n\leq N_0} \sum_{1\leq j\leq J} \Delta x (v_j^n)^2 \lesssim \sum_{1\leq j\leq J} \Delta x (f_j)^2. \tag{4.2.49}$$

By using the triangle inequality and combining the property (4.2.46) and the estimate (4.2.49), we get

$$\sup_{0\leq n\leq N_0} \sum_{1\leq j\leq J} \Delta x (u_j^n)^2 \leq C \sum_{1\leq j\leq J} \Delta x (f_j)^2, \tag{4.2.50}$$

with the constant $C > 0$ is independent of L and J.
The iteration argument: For any $m \in \mathbb{N}$, let us introduce

$$\theta_m := \sup_{mN_0\leq n\leq (m+1)N_0} \sum_{1\leq j\leq J} \Delta x (u_j^n).$$

The inequality (4.2.50) can be reformulated as

$$\theta_0 \leq C \sum_{1\leq j\leq J} \Delta x (f_j)^2.$$

Now, we can use the same procedure as in (4.2.42)-(4.2.50) with initial data $(u_j^{N_0})_{1\leq j\leq J}$ instead of $(f_j)_{1\leq j\leq J}$. Then, one yields

$$\theta_1 \leq C \sum_{1\leq j\leq J} \Delta x (u_j^{N_0})^2 \leq C\theta_0 \leq C^2 \sum_{1\leq j\leq J} \Delta x (f_j)^2.$$

It remains to iterate the latter estimate for all $m \in \mathbb{N}$,

$$\theta_m \leq C^{m+1} \sum_{1 \leq j \leq J} \Delta x (f_j)^2.$$

Hence, for a given $n \in \mathbb{N}$, we choose $m = E(n/N_0)$. From the above inequality, we have

$$\sum_{1 \leq j \leq J} \Delta x (u_j^n)^2 \leq C^{E(n/N_0)+1} \sum_{1 \leq j \leq J} \Delta x (f_j)^2. \tag{4.2.51}$$

Now, if we assume $k_b \leq J/4$ then

$$\frac{J/2 - k_b}{r} \geq \frac{J}{4r} \geq \frac{J}{4(p+r)}.$$

Besides, we also get

$$\frac{J}{2p} \geq \frac{J}{4(p+r)}.$$

Thus, one obtains

$$N_0 \geq \frac{J}{4(p+r)} = \frac{L\lambda}{4(p+r)} \times \frac{1}{\Delta t}. \tag{4.2.52}$$

From (4.2.51) and (4.2.52), we yield

$$\sum_{1 \leq j \leq J} \Delta x (u_j^n)^2 \leq C^{(n\Delta t)/(LC')+1} \sum_{1 \leq j \leq J} \Delta x (f_j)^2$$

with $C' = \lambda/4(p+r)$. Indeed, we can see that

$$C^{((n\Delta t)/(LC'))+1} = C \times C^{(n\Delta t)/(LC')} = Ce^{((n\Delta t)/(LC'))\ln(C)} \leq C_0 e^{C_0 n\Delta t/L}$$

with $C_0 \geq \max\left\{C, \dfrac{\ln(C)}{C'}\right\}$. Therefore, there exists $C_1 > 0$ such that

$$\forall\, n \in \mathbb{N}, \quad \left(\sum_{j=1}^{J} \Delta x (u_j^n)^2\right)^{1/2} \leq C_1 e^{C_1 n\Delta t/L} \left(\sum_{j=1}^{J} \Delta x (f_j)^2\right)^{1/2}.$$

This ends the proof of Proposition 4.2.7. □

4.2.4.2 Convergence

It remains to combine the convergence estimates of Theorems 4.2.6 and 4.2.4 to prove Theorem 4.1.2. We use a slight modification of the superposition argument in [22] in order to cope with the nonzero incoming data, but we basically follow the same lines. Let us consider a final time $T \geq 1$ and some data $f \in H^{k+1}((0,L))$, $g \in H^{k+1}((0,T))$ that satisfy the compatibility conditions stated in Theorem 4.1.2. We consider some function $\chi \in \mathcal{C}^\infty(\mathbb{R})$ such that

$$\chi(x) = \begin{cases} 0, & \text{if } x \leq 1/3, \\ 1, & \text{if } x \geq 2/3. \end{cases}$$

Then, we decompose the initial condition f as

$$\forall\, x \in (0,L)\,, \quad f(x) \,=\, (1-\chi(x/L))\, f(x) + \chi(x/L)\, f(x)\,.$$

Since $(1-\chi(\cdot/L))\, f$ vanishes on $(2\,L/3,L)$, we can extend it by zero to the interval $(L,+\infty)$ and thus consider $(1-\chi(\cdot/L))\, f$ as an element of $H^{k+1}((0,+\infty))$. Furthermore, the functions $(1-\chi(\cdot/L))\, f$ and g satisfy the same compatibility conditions as f and g at $t=x=0$. We can thus apply Theorem 4.2.6 to the sequence $(v_j^n)_{j\geq 1-r, 0\leq n\leq T/\Delta t}$ that is defined as the solution to the numerical scheme

$$\begin{cases} v_j^0 \,=\, \dfrac{1}{\Delta x}\displaystyle\int_{x_{j-1}}^{x_j} (1-\chi(x/L))\, f(x)\, \mathrm{d}x\,, & j\geq 1\,, \\ v_\ell^n \,=\, \displaystyle\sum_{\kappa=0}^{k-1} \frac{\Delta x^\kappa}{(\kappa+1)\,!\,(-a)^\kappa} \left(\ell^{\kappa+1}-(\ell-1)^{\kappa+1}\right) g^{(\kappa)}(t^n)\,, & 0\leq n\leq T/\Delta t\,, \quad 1-r\leq \ell\leq 0\,, \\ v_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, v_{j+\ell}^n\,, & 0\leq n\leq T/\Delta t-1\,, \quad j\geq 1\,. \end{cases}$$

We obtain the estimate

$$\sup_{0\leq n\leq T/\Delta t} \left(\sum_{j\geq 1} \Delta x \left(v_j^n - \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} v(t^n,x)\, \mathrm{d}x \right)^2 \right)^{1/2} \leq\, C\, T\;\Delta x^k \left(\|f\|_{H^{k+1}((0,L))}+\|g\|_{H^{k+1}((0,T))}\right)\,, \tag{4.2.53}$$

where v is the exact solution to the transport problem

$$\begin{cases} \partial_t v + a\,\partial_x v \,=\, 0\,, & t\in(0,T)\,, \quad x\geq 0\,, \\ v(0,x) \,=\, (1-\chi(x/L))\, f(x)\,, & x\geq 0\,, \\ v(t,0) \,=\, g(t)\,, & t\in(0,T)\,. \end{cases}$$

Similarly, we can view $\chi(\cdot/L)\, f$ as an element of $H^{k+1}((-\infty,L))$ that vanishes on $(-\infty,L/3)$. Theorem 4.2.4 then shows that the solution $(w_j^n)_{j\leq J, 0\leq n\leq T/\Delta t}$ to the numerical scheme

$$\begin{cases} w_j^0 \,=\, \dfrac{1}{\Delta x}\displaystyle\int_{x_{j-1}}^{x_j} \chi(x/L)\, f(x)\, \mathrm{d}x\,, & j\leq J\,, \\ (D_-^{k_b} w^n)_{J+\ell} \,=\, 0\,, & 0\leq n\leq T/\Delta t\,, \quad 1\leq \ell\leq p\,, \\ w_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, w_{j+\ell}^n\,, & 0\leq n\leq T/\Delta t-1\,, \quad j\leq J\,, \end{cases}$$

satisfies

$$\begin{aligned} \sup_{0\leq n\leq T/\Delta t} \left(\sum_{j\leq J} \Delta x \left(w_j^n - \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} \chi((x-at^n)/L) f(x-at^n) dx \right)^2 \right)^{1/2} \\ \leq CT\Delta x^{\min\,(k,k_b)}\|f\|_{H^{k+1}((0,L))}. \end{aligned} \tag{4.2.54}$$

Using the support property of the function χ and the fact that the scheme (4.1.7) is explicit with a finite stencil, we find that for all time iteration n up to the threshold

$$N \,:=\, \min\left(\mathrm{E}\left(\frac{J/3-k_b}{r}\right)\,,\, E\left(\frac{E(J/3)}{p}\right)\right)\,,$$

there holds

$$w_{1-r}^n = \cdots = w_0^n = 0\,, \quad v_{J+1-k_b}^n = \cdots = v_{J+p}^n = 0\,.$$

In particular, the solution $(u_j^n)_{1-r\le j\le J+p, 0\le n\le T/\Delta t}$ to (4.1.4), (4.1.7), (4.1.8), (4.1.12) satisfies

$$\forall\, 0 \le n \le N\,, \quad \forall\, 1-r \le j \le J+p\,, \quad u_j^n = v_j^n + w_j^n\,.$$

Combining then the error estimates (4.2.53) and (4.2.54), we obtain

$$\sup_{0\le n\le N} \left(\sum_{1\le j\le J} \Delta x \left(u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} u(t^n,x)\,\mathrm{d}x \right)^2 \right)^{1/2} \le C_1 T \Delta x^{\min(k,k_b)} \left(\|f\|_{H^{k+1}((0,L))} + \|g\|_{H^{k+1}((0,T))} \right)\,, \tag{4.2.55}$$

where u is the exact solution to (4.1.1).

It remains, as in Section 4.2.4.1 , to iterate in time the error estimate (4.2.55). We follow again the argument in Section 4.2.4.1. For any time iteration n between N and $2\,N$, we split the solution $(u_j^n)_{1-r\le j\le J+p, 0\le n\le T/\Delta t}$ to (4.1.7), (4.1.4), (4.1.8), (4.1.12) as the sum of the solution to the problem

$$\begin{cases} \tilde{u}_j^N = \dfrac{1}{\Delta x} \displaystyle\int_{x_{j-1}}^{x_j} u(t^N,x)\,\mathrm{d}x\,, & 1 \le j \le J\,, \\ (D_-^{k_b} \tilde{u}^{N+n})_{J+\ell} = 0\,, & 0 \le n \le N\,, \quad 1 \le \ell \le p\,, \\ \tilde{u}_\ell^{N+n} = \displaystyle\sum_{\kappa=0}^{k-1} \frac{\Delta x^\kappa}{(\kappa+1)\,!\,(-a)^\kappa} \left(\ell^{\kappa+1} - (\ell-1)^{\kappa+1} \right) g^{(\kappa)}(t^{N+n})\,, & 0 \le n \le N\,, \quad 1-r \le \ell \le 0\,, \\ \tilde{u}_j^{N+n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, \tilde{u}_{j+\ell}^{N+n}\,, & 0 \le n \le N-1\,, \quad 1 \le j \le J\,, \end{cases}$$

and of the (presumably small) solution to the error problem

$$\begin{cases} \varepsilon_j^N = u_j^N - \dfrac{1}{\Delta x} \displaystyle\int_{x_{j-1}}^{x_j} u(t^N,x)\,\mathrm{d}x\,, & 1 \le j \le J\,, \\ (D_-^{k_b} \varepsilon^{N+n})_{J+\ell} = 0\,, & 0 \le n \le N\,, \quad 1 \le \ell \le p\,, \\ \varepsilon_\ell^{N+n} = 0\,, & 0 \le n \le N\,, \quad 1-r \le \ell \le 0\,, \\ \varepsilon_j^{N+n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell\, \varepsilon_{j+\ell}^{N+n}\,, & 0 \le n \le N-1\,, \quad 1 \le j \le J\,. \end{cases}$$

Since the initial condition $u(\cdot - a\,t^N)$ and the boundary source term $g(t^N + \cdot)$ satisfy the compatibility conditions at the corner $t = x = 0$, we can apply the first step of the proof (leading to the error estimate (4.2.55)) for the $(\tilde{u}_j^{N+n})$ part, and we apply the stability estimate of Proposition 4.2.7 for the (ε_j^{N+n}) part. This leads to the second error estimate

$$\sup_{N\le n\le 2\,N} \left(\sum_{j\le J} \Delta x \left(u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} u(t^n,x)\,\mathrm{d}x \right)^2 \right)^{1/2} \\ \le C_1\,(1+C_2)\,T\,\Delta x^{\min(k,k_b)} \left(\|f\|_{H^{k+1}((0,L))} + \|g\|_{H^{k+1}((0,T))} \right)\,,$$

with $C_2 = C_0 e^{C_0 N/\Delta t}$ and, more generally, to

$$\sup_{\mu\, N \le n \le (\mu+1)\, N} \left(\sum_{j \le J} \Delta x \, \left(u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} u(t^n,x) \, \mathrm{d}x \right)^2 \right)^{1/2} \\ \le \, C_1 \left(\sum_{\nu=0}^{\mu} C_2^\nu \right) \, T \, \Delta x^{\min(k,k_b)} \left(\| f \|_{H^{k+1}((0,L))} + \| g \|_{H^{k+1}((0,T))} \right) .$$

Indeed, we end up with

$$\forall \, \mu \in \mathbb{N}, \quad \sup_{0 \le n \le \mu\, N} \left(\sum_{j=1}^{J} \Delta x \, \left(u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} u(t^n,x) \, \mathrm{d}x \right)^2 \right)^{1/2} \\ \le \, C_1 \left(\sum_{\nu=0}^{\mu-1} C_2^\nu \right) \, T \, \Delta x^{\min(k,k_b)} \left(\| f \|_{H^{k+1}((0,L))} + \| g \|_{H^{k+1}((0,T))} \right) \\ \le \, C_1 \, C_0^\mu \, e^{C_0 \mu N \Delta t / L} \, T \, \Delta x^{\min(k,k_b)} \left(\| f \|_{H^{k+1}((0,L))} + \| g \|_{H^{k+1}((0,T))} \right),$$

where we have assumed $C_0 \ge 2$ without loss of generality. It remains to choose $\mu := E(N\Delta t/T) + 1$, which by definition of N is uniformly bounded with respect to J (N scales like cJ with $c > 0$ and Δt scales like c'/J with $c' > 0$) and we end up with

$$\sup_{0 \le n \le T/\Delta t} \left(\sum_{j=1}^{J} \Delta x \, \left(u_j^n - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} u(t^n,x) \, \mathrm{d}x \right)^2 \right)^{1/2} \\ \le \, C T \, e^{CT/L} \, \Delta x^{\min(k,k_b)} \left(\| f \|_{H^{k+1}((0,L))} + \| g \|_{H^{k+1}((0,T))} \right).$$

The convergence estimate (4.1.13) of Theorem 4.1.2 follows by a direct lower bound for the norm on the left hand side.

4.3 High order outflow boundary layer analysis

In the present section, we explain how the analysis of [6], which dealt with the case of the Dirichlet boundary condition at the outflow boundary, can be extended to the case of high order extrapolation (4.1.8). The goal is to obtain an accurate description of the numerical solution close to the outflow boundary by means of a boundary layer expansion. The leading order term in the expansion corresponds to the exact solution to the transport equation. However, this leading order term does not satisfy the extrapolation condition (4.1.8), leading to a consistency error of magnitude $O(\Delta x^{k_b})$ on the boundary. Under some mild structural assumption on the numerical scheme (4.1.7), we show below that this $O(\Delta x^{k_b})$ error on the boundary gives rise to a boundary layer term which scales as $O(\Delta x^{k_b+1/2})$ in the ℓ^2_j norm. This gain of a factor $\Delta x^{1/2}$ enables us to recover the optimal convergence rate k_b in the maximum norm on the whole spatial domain for $k_b < k$.

4.3.1 An introductive example

Let us go back for a while to the case of the Lax-Wendroff scheme (4.1.3), which we consider here on the left half space

$$u_j^{n+1} = \sum_{\ell=-r}^{p} a_\ell u_{j+\ell}^n, \quad j \le J, n \in \mathbb{N}, \tag{4.3.1}$$

with $p = r = 1$, $a_{-1} = a\lambda(a\lambda + 1)/2$, $a_0 = 1 - a^2\lambda^2$ and $a_1 = a\lambda(a\lambda - 1)/2$. We start with some smooth initial condition f defined on $(-\infty, L)$ which we project as a piecewise constant function

$$u_j^0 := \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x)dx, \quad j \leq J.$$

The exact solution to the transport equation on $(-\infty, L)$ with initial condition f is $u(t, x) = f(x - at)$ (recall $a > 0$). Hence, the consistency analysis of the Lax-Wendroff scheme indicates that u_j^n reads

$$u_j^n = \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x - at^n)dx + \varepsilon_j^n, \tag{4.3.2}$$

where the first term in the expansion on the right hand side yields an $O(\Delta x^2)$ consistency error in the interior domain, but also an $O(\Delta x)$ consistency error on the boundary. If we wish to push forward the above expansion, we need to take into account the boundary consistency error and introduce a corrector which will hopefully not alter the interior consistency error. This can be achieved by observing that the sequence

$$v_j := \kappa^j, \quad j \in \mathbb{Z}, \quad \kappa := -\frac{1 + \lambda a}{1 - \lambda a},$$

is kept unchanged by the Lax-Wendroff scheme on $\mathbb{Z}$, and belongs to $\ell^2(-\infty, J)$ (we assume $0 < \lambda a < 1$ so $|\kappa| > 1$). Hence, to remove the boundary consistency error, we can add a corrector on the right hand side of (4.3.2) in the following way

$$u_j^n = \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x - at^n)dx + \Delta x w^n v_{j-J} + \varepsilon_j^n, \tag{4.3.3}$$

where w^n is defined in such a way that the two first terms on the right hand side satisfy the homogeneous Dirichlet boundary condition for $k_b = 0$ and the first order extrapolation condition for $k_b = 1$ while $k = 2$ for the Lax-Wendroff scheme.

The case $k_b = 0$: At the outflow boundary condition, we first impose the homogeneous Dirichlet boundary condition

$$u_{J+1}^n = 0, \quad n \in \mathbb{N} \tag{4.3.4}$$

Indeed, the value of w^n is given by

$$w^n = -\frac{1}{\Delta x^2 \kappa} \int_{x_J}^{x_{J+1}} f(x - at^n)dx. \tag{4.3.5}$$

Then, we have the following estimate

$$\begin{aligned} \sum_{j \leq J} \Delta x \, |\Delta x w^n v_{j-J}|^2 &\leq \frac{1}{\Delta x} \left(\int_{x_J}^{x_{J+1}} |f(x - at^n)|dx \right)^2 \times \sum_{j \leq J} |\kappa^{j-J-1}|^2 \\ &\lesssim \frac{1}{\Delta x} \left(\int_{x_J}^{x_{J+1}} \|f\|_{H^1((-\infty,L))}dx \right)^2 \times \sum_{j \leq J} |\kappa^{j-J-1}|^2 \\ &\lesssim \Delta x \|f\|^2_{H^1((-\infty,L))}. \end{aligned} \tag{4.3.6}$$

According to (4.3.3)-(4.3.5), we get

$$\forall \, j \leq J, \quad \varepsilon_j^0 = u_j^0 - \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x)dx - \Delta x w^0 v_{j-J} = -\Delta x w^0 v_{j-J}. \tag{4.3.7}$$

and for any $n \in \mathbb{N}$

$$\begin{aligned}\varepsilon_{J+1}^n &= u_{J+1}^n - \frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} f(x-at^n)dx - \Delta x w^n v_1 \\ &= 0 - \frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} f(x-at^n)dx + \frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} f(x-at^n)dx \\ &= 0.\end{aligned} \tag{4.3.8}$$

Besides, from (4.3.1) and (4.3.3), one has for all $j \leq J$ and any $n \in \mathbb{N}$,

$$\begin{aligned}\varepsilon_j^{n+1} &= u_j^{n+1} - \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} f(x-at^{n+1})dx - \Delta x w^{n+1} v_{j-J} \\ &= \sum_{\ell=-r}^{p} a_\ell u_{j+\ell}^n - \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} f(x-at^{n+1})dx + \frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} f(x-at^{n+1})dx \times v_{j-(J+1)} \\ &= \sum_{\ell=-r}^{p} a_\ell \left(\frac{1}{\Delta x}\int_{x_{j+\ell-1}}^{x_{j+\ell}} f(x-at^n)dx + \Delta x w^n v_{j+\ell-J} + \varepsilon_{j+\ell}^n\right) \\ &- \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} f(x-at^{n+1})dx + \frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} f(x-at^{n+1})dx \times v_{j-(J+1)} \\ &= \sum_{\ell=-r}^{p} a_\ell \varepsilon_{j+\ell}^n + \Delta t\left(e_j^{n+1} + \delta_j^{n+1}\right),\end{aligned} \tag{4.3.9}$$

where the consistency error $(e_j^n)_{j\leq J,\, n\in\mathbb{N}}$ and $(\delta^n)_{j\leq J,\, n\in\mathbb{N}}$ read

$$e_j^n = \frac{1}{\Delta t \Delta x}\left(\sum_{\ell=-r}^{p} a_\ell \int_{x_{j+\ell-1}}^{x_{j+\ell}} f(x-at^n)dx - \int_{x_{j-1}}^{x_j} f(x-at^{n+1})dx\right) \tag{4.3.10}$$

and

$$\begin{aligned}\delta_j^n &= \frac{1}{\Delta t}\left(\sum_{\ell=-r}^{p} a_\ell \Delta x w^n v_{j+\ell-J} + \frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} f(x-at^{n+1})dx \times v_{j-(J+1)}\right) \\ &= \frac{1}{\Delta t \Delta x}\left(-\int_{x_J}^{x_{J+1}} f(x-at^n)dx \times \sum_{\ell=-r}^{p} a_\ell v_{j+\ell-J} + \int_{x_J}^{x_{J+1}} f(x-at^{n+1})dx \times v_{j-(J+1)}\right) \\ &= \frac{1}{\Delta t \Delta x}\int_{x_J}^{x_{J+1}} \left(-f(x-at^n) + f(x-at^{n+1})\right) dx \times v_{j-(J+1)}.\end{aligned} \tag{4.3.11}$$

According to (4.3.7)-(4.3.11), we can see that $(\varepsilon_j^n)_{j\leq J,\, 0\leq n\leq T/\Delta t}$ satisfies

$$\begin{cases} \varepsilon_j^0 = -\Delta x w^0 v_{j-J}, & j \leq J, \\ \varepsilon_{J+1}^n = 0, & 0 \leq n \leq T/\Delta t - 1, \\ \varepsilon_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell \varepsilon_{j+\ell}^n + \Delta t F_j^{n+1}, & j \leq J,\ 0 \leq n \leq T/\Delta t - 1, \end{cases} \tag{4.3.12}$$

with the interior source term $(F_j^n)_{j\leq J,\, 1\leq n\leq T/\Delta t}$ is defined by

$$F_j^n = e_j^n + \delta_j^n. \tag{4.3.13}$$

By an application of Proposition 4.2.3 for the scheme (4.3.12), we get the following estimate

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2 \leq C\left(\sum_{j\leq J}\Delta x\left(-\Delta x w^0 v_{j-J}\right)^2 + T^2\sup_{1\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(F_j^n)^2\right). \tag{4.3.14}$$

By using the estimate (4.3.6) and the definition of $(F_j^n)_{j\leq J,\,1\leq n\leq T/\Delta t}$ in (4.3.13), the inequality (4.3.14) becomes

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2 \leq C\left(\Delta x\|f\|^2_{H^1((-\infty,L))} + T^2\sup_{1\leq n\leq T/\Delta t}\left(\sum_{j\leq J}\Delta x(e_j^n)^2 + \sum_{j\leq J}\Delta x(\delta_j^n)^2\right)\right). \tag{4.3.15}$$

One can proceed as in (4.2.24)- (4.2.29) to control the interior consistency error as follows

$$\sup_{1\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(e_j^n)^2 \leq C\Delta x^{2k}\|f\|^2_{H^{k+1}((-\infty,L))}. \tag{4.3.16}$$

Besides, we can see that

$$\begin{aligned}|\delta_j^n| &\leq \frac{1}{\Delta t\Delta x}\int_{x_J}^{x_{J+1}}\left|f(x-at^{n+1}) - f(x-at^n)\right|dx \times |v_{j-(J+1)}|\\ &\leq \frac{1}{\Delta t\Delta x}\int_{x_J}^{x_{J+1}}\Delta t\int_0^1 f'(x-a(t^n+\theta\Delta t))d\theta dx \times |v_{j-(J+1)}|\\ &\lesssim \|f\|_{H^2((-\infty,L))} \times |v_{j-(J+1)}|.\end{aligned}$$

Then, we get the following estimate

$$\sup_{1\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\delta_j^n)^2 \lesssim \Delta x\|f\|^2_{H^2((-\infty,L))}. \tag{4.3.17}$$

Substituting (4.3.16) and (4.3.17) into (4.3.14), we have

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2 \leq C\left(\Delta x\|f\|^2_{H^1((-\infty,L))} + T^2\left(\Delta x^{2k}\|f\|^2_{H^{k+1}((-\infty,L))} + \Delta x\|f\|^2_{H^2((-\infty,L))}\right)\right)$$

and this immediately gives the following estimate

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2 \leq CT^2\Delta x\|f\|^2_{H^{k+1}((-\infty,L))} \tag{4.3.18}$$

On the other hand, from (4.3.3), we get

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x\left(u_j^n - \int_{x_{j-1}}^{x_j} f(x-at^n)dx\right)^2 \lesssim \sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2 + \sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x\left(\Delta x w^n v_{j-J}\right)^2.$$

According to (4.3.6) and (4.3.18), the above inequality becomes

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x\left(u_j^n - \int_{x_{j-1}}^{x_j} f(x-at^n)dx\right)^2 \leq CT^2\Delta x\|f\|^2_{H^{k+1}((-\infty,L))}$$

and this immediately gives the uniform convergence estimate

$$\sup_{0\leq n\leq T/\Delta t}\sup_{j\leq J}\left|u_j^n-\int_{x_{j-1}}^{x_j}f(x-at^n)dx\right|\leq CT\|f\|_{H^{k+1}((-\infty,L))}.$$

The case $k_b=1$: At the outflow boundary condition, we impose the first order extrapolation boundary condition

$$u_{J+1}^n=u_J^n,\quad n\in\mathbb{N}. \tag{4.3.19}$$

We thus introduce the notation

$$\omega_j^n:=\frac{1}{\Delta x}\int_{x_{j-1}}^{x_j}f(x-a\,t^n)\,\mathrm{d}x\,,\quad j\leq J+1\,,\quad n\in\mathbb{N}\,.$$

Indeed, the value of w^n is defined by

$$\begin{aligned}w^n&=-\frac{1}{\Delta x\kappa}\times\frac{(D_-\omega^n)_{J+1}}{(D_-v_{-(J+1)})_{J+1}}\\&=-\frac{1}{\Delta x(\kappa-1)}\times\frac{1}{\Delta x}\left(\int_{x_J}^{x_{J+1}}f(x-at^n)dx-\int_{x_{J-1}}^{x_J}f(x-at^n)dx\right).\end{aligned}\tag{4.3.20}$$

The important observation at this point is that defining w^n requires the real number κ not to equal 1. This fact follows here from a mere verification but it is a general consequence of the analysis in [36] of the Lopatinskii determinant associated with the boundary condition (4.1.8) (see also the proof of Lemma 4.3.5 below). Then, by applying Lemma 4.2.5, we have the following estimate

$$\begin{aligned}\sum_{j\leq J}\Delta x|\Delta x w^n v_{j-J}|^2&=\frac{1}{(\kappa-1)^2}\sum_{j\leq J}\Delta x\left|\frac{1}{\Delta x}\left(\int_{x_J}^{x_{J+1}}f(x-at^n)dx-\int_{x_{J-1}}^{x_J}f(x-at^n)dx\right)\right|^2\times|v_{j-J}|^2\\&\lesssim\Delta x^3\|f\|^2_{H^1((-\infty,L))}.\end{aligned}\tag{4.3.21}$$

Following (4.3.3), (4.3.19) and (4.3.20), we get

$$\forall\, j\leq J,\quad \varepsilon_j^0=u_j^0-\frac{1}{\Delta x}\int_{x_{j-1}}^{x_j}f(x)dx-\Delta x w^0 v_{j-J}=-\Delta x w^0 v_{j-J}. \tag{4.3.22}$$

and for any $n\in\mathbb{N}$

$$(D_-\varepsilon^n)_{J+1}=0. \tag{4.3.23}$$

Besides, from (4.3.1) and (4.3.3), one has for all $j\leq J$ and any $n\in\mathbb{N}$,

$$\varepsilon_j^{n+1}=\sum_{\ell=-r}^{p}a_\ell\varepsilon_{j+\ell}^n+\Delta t\left(e_j^{n+1}+\delta_j^{n+1}\right) \tag{4.3.24}$$

with the consistency error $(e_j^n)_{j\leq J,n\in\mathbb{N}}$ is the same as in (4.3.10) and $(\delta^n)_{j\leq J,n\in\mathbb{N}}$ reads

$$\begin{aligned}\delta_j^n=\frac{1}{\Delta x\Delta t}\times\frac{1}{\kappa-1}\times&\left(\int_{x_J}^{x_{J+1}}\left(f(x-at^{n+1})-f(x-at^n)\right)dx\right.\\&\left.-\int_{x_{J-1}}^{x_J}\left(f(x-at^{n+1})-f(x-at^n)\right)dx\right)\times v_{j-J}.\end{aligned}\tag{4.3.25}$$

According to (4.3.22)-(4.3.25), we can see that $(\varepsilon_j^n)_{j\leq J,\, 0\leq n\leq T/\Delta t}$ satisfies

$$\begin{cases} \varepsilon_j^0 = -\Delta x w^0 v_{j-J}, & j\leq J, \\ (D_-\varepsilon^n)_{J+1} = 0, & 0\leq n\leq T/\Delta t - 1, \\ \varepsilon_j^{n+1} = \displaystyle\sum_{\ell=-r}^{p} a_\ell \varepsilon_{j+\ell}^n + \Delta t F_j^{n+1}, & j\leq J,\, 0\leq n\leq T/\Delta t-1, \end{cases} \tag{4.3.26}$$

with the interior source term $(F_j^n)_{j\leq J,\, 1\leq n\leq T/\Delta t}$ is defined by

$$F_j^n = e_j^n + \delta_j^n. \tag{4.3.27}$$

Again, by an application of Proposition 4.2.3 for the scheme (4.3.26), we get the following estimate

$$\sup_{0\leq n\leq T/\Delta t} \sum_{j\leq J} \Delta x (\varepsilon_j^n)^2 \leq C\left(\sum_{j\leq J} \Delta x \left(-\Delta x w^0 v_{j-J}\right)^2 + T^2 \sup_{1\leq n\leq T/\Delta t} \sum_{j\leq J} \Delta x (F_j^n)^2 \right). \tag{4.3.28}$$

By using the estimates (4.3.16), (4.3.21) and the definition of $(F_j^n)_{j\leq J,\, 1\leq n\leq T/\Delta t}$ in (4.3.27), the inequality (4.3.14) becomes

$$\begin{aligned} \sup_{0\leq n\leq T/\Delta t} \sum_{j\leq J} \Delta x (\varepsilon_j^n)^2 &\leq C\left(\Delta x^3 \|f\|^2_{H^1((-\infty,L))} + T^2 \sup_{1\leq n\leq T/\Delta t} \left(\sum_{j\leq J} \Delta x (e_j^n)^2 + \sum_{j\leq J} \Delta x (\delta_j^n)^2 \right)\right) \\ &\leq C\left(\Delta x^3 \|f\|^2_{H^1((-\infty,L))} + T^2 \left(\Delta x^{2k} \|f\|^2_{H^{k+1}((-\infty,L))} + \sup_{1\leq n\leq T/\Delta t} \sum_{j\leq J} \Delta x (\delta_j^n)^2 \right)\right). \end{aligned} \tag{4.3.29}$$

Now, we can observe that

$$\begin{aligned} &\frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} \left(f(x-at^{n+1}) - f(x-at^n)\right) dx - \frac{1}{\Delta x}\int_{x_{J-1}}^{x_J} \left(f(x-at^{n+1}) - f(x-at^n)\right) dx \\ =&\frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} \Delta t \int_0^1 f'(x-a(t^n+\theta\Delta t)) d\theta dx - \frac{1}{\Delta x}\int_{x_{J-1}}^{x_J} \Delta t \int_0^1 f'(x-a(t^n+\theta\Delta t)) d\theta dx \\ =&\Delta t (D_-\nu^n)_{J+1} \end{aligned}$$

with

$$\nu_j^n = \frac{1}{\Delta x}\int_{x_{j-1}}^{x_j} \Delta t \int_0^1 f'(x-a(t^n+\theta\Delta t)) d\theta dx.$$

By applying Lemma 4.2.5, we have the following estimate

$$\left| \frac{1}{\Delta x}\int_{x_J}^{x_{J+1}} \left(f(x-at^{n+1}) - f(x-at^n)\right) dx - \frac{1}{\Delta x}\int_{x_{J-1}}^{x_J} \left(f(x-at^{n+1}) - f(x-at^n)\right) dx \right| \\ \leq \Delta t \Delta x \|f\|_{H^3((-\infty,L))}.$$

Therefore, we get

$$\sup_{1\leq n\leq T/\Delta t} \sum_{j\leq J} \Delta x (\delta_j^n)^2 \leq \Delta x^3 \|f\|^2_{H^3((-\infty,L))}. \tag{4.3.30}$$

Substituting (4.3.30) into (4.3.29),one has

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2\leq C\left(\Delta x^3\|f\|^2_{H^1((-\infty,L))}+T^2\left(\Delta x^{2k}\|f\|^2_{H^{k+1}((-\infty,L))}+\Delta x^3\|f\|^2_{H^3((-\infty,L))}\right)\right)$$

and this immediately gives the following estimate

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2\leq CT^2\Delta x^3\|f\|^2_{H^{k+1}((-\infty,L))} \tag{4.3.31}$$

On the other hand, from (4.3.3), we get

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x\left(u_j^n-\int_{x_{j-1}}^{x_j}f(x-at^n)dx\right)^2\lesssim\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x(\varepsilon_j^n)^2+\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x\left(\Delta x w^n v_{j-J}\right)^2.$$

According to (4.3.21) and (4.3.31), the above inequality becomes

$$\sup_{0\leq n\leq T/\Delta t}\sum_{j\leq J}\Delta x\left(u_j^n-\int_{x_{j-1}}^{x_j}f(x-at^n)dx\right)^2\leq CT^2\Delta x^3\|f\|^2_{H^{k+1}((-\infty,L))}$$

and this immediately gives the uniform convergence estimate

$$\sup_{0\leq n\leq T/\Delta t}\sup_{j\leq J}\left|u_j^n-\int_{x_{j-1}}^{x_j}f(x-at^n)dx\right|\leq CT\Delta x\|f\|_{H^{k+1}((-\infty,L))}.$$

The above brief sketch is made complete and rigorous below in the general framework of Theorem 4.1.3.

4.3.2 Discrete steady states

Formalizing somehow the previous example in a more general framework, let us now introduce the following definition.

Definition 4.3.1 (Steady state for the numerical scheme)**.** *A sequence $(v_j)_{j\in\mathbb{Z}}$ is called a discrete steady state of the scheme (4.1.7) if it is kept unchanged by the time iteration process on $\mathbb{Z}$, that is, if it satisfies*

$$\forall\, j\in\mathbb{Z},\qquad \sum_{\ell=-r}^{p}a_\ell v_{j+\ell}=v_j. \tag{4.3.32}$$

In order to characterize the discrete steady states, it is natural to introduce the characteristic polynomial

$$A(X):=\sum_{\ell=-r}^{p}a_\ell X^{\ell+r}-X^r. \tag{4.3.33}$$

From the consistency property (4.1.10), any constant sequence is a discrete steady solution for (4.1.7), the same property being available for the continuous model (namely, the transport operator). However, the discrete nature of the differentiation operator involved in the numerical scheme (4.1.7) allows the existence of many other discrete steady solutions. The latter play an important role when considering then the half-space problem with some discrete boundary conditions.

From the non-characteristic assumption $a \neq 0$, it follows that, among the roots of A, $X = 1$ is always a simple root. Let us now introduce the whole set of (pairwise distinct) roots of A together with their multiplicities through the full factorization of A in $\mathbb{C}[X]$

$$A(X) = a_p \prod_{\sigma=1}^{\tau} (X - \kappa_\sigma)^{\mu_\sigma}. \tag{4.3.34}$$

Clearly, looking at the degree of the polynomial A, one has the equality

$$\sum_{\sigma=1}^{\tau} \mu_\sigma = r + p.$$

For convenience, we order the roots of A with decreasing modulus

$$|\kappa_1| \geq |\kappa_2| \geq \ldots \geq |\kappa_\tau|.$$

To make the analysis more intelligible, we will work under the following assumption, which was already present in [6].

Assumption 4.3.2. *The characteristic polynomial A defined in (4.3.33) has a unique root (equal to 1) on the unit circle $\mathbb{S} = \{z \in \mathbb{C},\ |z| = 1\}$. In other words, we assume*

$$\bigcup_{\sigma=1}^{\tau} \{\kappa_\sigma\} \cap \mathbb{S} = \{1\}. \tag{4.3.35}$$

As observed on above example of the Lax-Wendroff scheme, the steady states we are looking at should decrease rapidly as j tends to $-\infty$, so that they only provide with a localized correction (near the boundary) to the usual convergence analysis and belong to $\ell^2_{\Delta x}(-\infty, J)$. We are therefore only concerned with those roots of A that have modulus greater than 1. Lemma 4.3.3 below gives the precise number of such root (counted with their multiplicities). We refer to [6, Lemma 2.1] for the proof.

Lemma 4.3.3 (Unstable roots of A [6])**.** *Under assumptions 4.1.1 and 4.3.2, letting $\kappa_1, \ldots, \kappa_{\tau_+}$ be the roots of A belonging to $\mathbb{U} = \{z \in \mathbb{C},\ |z| > 1\}$ with their corresponding multiplicities $\mu_1, \ldots, \mu_{\tau_+}$, then one has*

$$\sum_{\sigma=1}^{\tau_+} \mu_\sigma = p. \tag{4.3.36}$$

A direct consequence of Lemma 4.3.3 is the following description of steady states for (4.1.7) that belong to $\ell^2(-\infty, J)$. The proof follows from the standard description of the set of solutions to the recurrence relation (4.3.32).

Lemma 4.3.4. *The set of discrete steady solutions of the scheme (4.1.7) that belongs to $\ell^2(-\infty, J)$ is the finite dimensional linear subspace spanned by the p linearly independent sequences $\rho^{(\sigma,\nu)}$*

$$\rho_j^{(\sigma,\nu)} := (j - J)^\nu \kappa_\sigma^{j-J}, \quad j \in \mathbb{Z}, \quad 0 \leq \nu < \mu_\sigma,\ 1 \leq \sigma \leq \tau_+. \tag{4.3.37}$$

Equivalently, these discrete steady solutions in $\ell^2(-\infty, L)$ read

$$v_j = \sum_{\sigma=1}^{\tau_+} p_\sigma(j) \kappa_\sigma^{j-J}, \quad j \in \mathbb{Z}, \tag{4.3.38}$$

where $p_\sigma \in \mathbb{C}_{\mu_\sigma - 1}[X]$ for all index $1 \leq \sigma \leq \tau_+$.

Let us detail the parametrization of the set of (stable) discrete steady states on the two main examples we are concerned with. For the Lax-Wendroff scheme (4.1.3), one has

$$A(X) = -\frac{\lambda a(1-\lambda a)}{2}X^2 + (1-(\lambda a)^2)X + \frac{\lambda a(1+\lambda a)}{2}.$$

The (two simple) roots of A are 1 and

$$\kappa := -\frac{1+\lambda a}{1-\lambda a},$$

with $\kappa \in \mathbb{U}$ assuming, as usual, $0 < \lambda a < 1$. For the half space problem on $(-\infty, J)$, κ is therefore the unique stable root, and 1 counts as an unstable root (see [6]). In particular, assumption 4.3.2 is satisfied. The set of solutions to (4.3.32) that belongs to $\ell^2(-\infty, J)$ is the one-dimensional subspace spanned by the sequence $(\kappa^{j-J})_{j\in\mathbb{Z}}$.

Let us now consider the so-called $O3$ scheme, which is a convex combination of the Lax-Wendroff and Beam-Warming schemes, see [84, 29]. We now have $p = 1$ and $r = 2$, and the scheme reads

$$\begin{aligned} u_j^{n+1} = & -\frac{\lambda a(\lambda a+1)(1-\lambda a)}{6}u_{j-2}^n + \frac{\lambda a(\lambda a+1)(2-\lambda a)}{2}u_{j-1}^n \\ & + \frac{(\lambda a+1)(1-\lambda a)(2-\lambda a)}{2}u_j^n - \frac{\lambda a(1-\lambda a)(2-\lambda a)}{6}u_{j+1}^n, \end{aligned} \tag{4.3.39}$$

with, again, $0 < \lambda a < 1$. Assumption 4.1.1 is then satisfied (with $k = 3$). The roots of the corresponding characteristic polynomial A are

$$\kappa_\pm := \frac{-(1+\lambda a)(5-2\lambda a) \pm \sqrt{(1+\lambda a)(33-15\lambda a)}}{2(1-\lambda a)(2-\lambda a)}, \quad \kappa_0 := 1,$$

each of them being simple. The root κ_- is the only one in $\mathbb{U}$ and κ_+ belongs to the open unit disk $\mathbb{D}$, which is consistent with Lemma 4.3.4 ($p = 1$). In particular, assumption 4.3.2 is satisfied. The plots of corresponding roots (except $\kappa = 1$) according to the value of λa is shown in Figure 4.3.1.

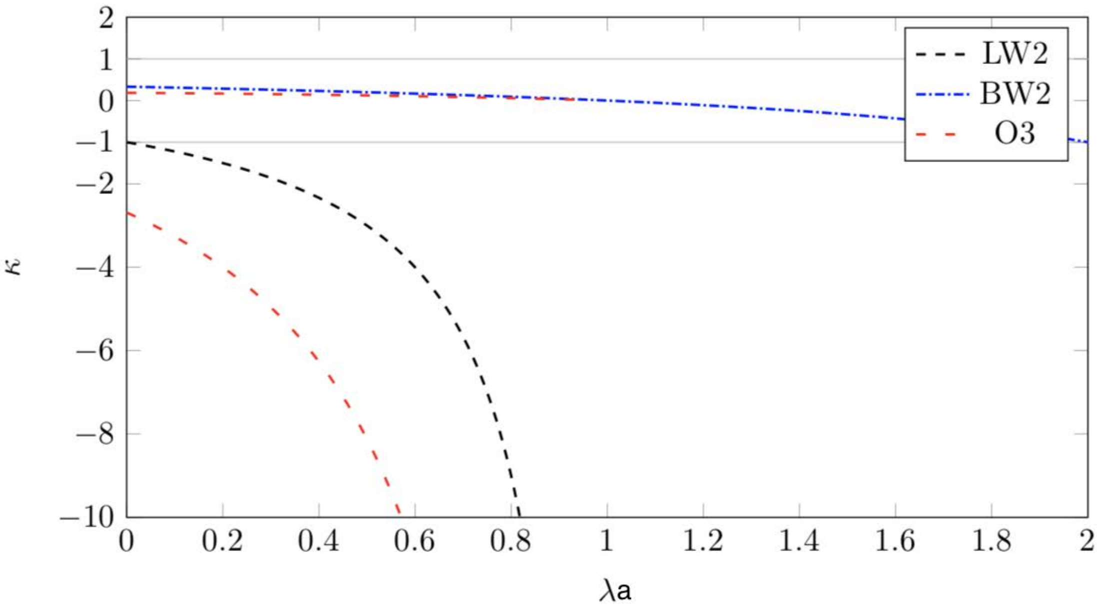

Figure 4.3.1: Generating roots as functions of the CFL number.

4.3.3 The boundary layer expansion. Proof of Theorem 4.1.3

We now start proving Theorem 4.1.3, and for that, we consider some initial condition $f \in H^{k+1}((-\infty, L))$ which, for convenience, we extend to the whole real line $\mathbb{R}$ as an element of $H^{k+1}(\mathbb{R})$. Our aim is to compare the solution to the scheme (4.1.14) (which is set on a half line) with the piecewise constant projection of the exact solution to the transport equation. We thus introduce the notation

$$\omega_j^n := \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} f(x - a\, t^n) \, \mathrm{d}x \, , \quad j \leq J + p \, , \quad n \in \mathbb{N} \, .$$

The consistency analysis in Sections 4.2.1 and 4.2.2 of the scheme (4.1.14) amounts to considering the numerical scheme satisfied by the error $(u_j^n - \omega_j^n)$. It is proved that the resulting boundary consistency errors have size $O(\Delta x^{k_b})$, while the interior consistency errors have size $O(\Delta x^k)$. Here we have $k_b < k$ so the worst term is on the boundary. Following the arguments in [6], we are therefore going to introduce a boundary layer corrector in order to remove the boundary consistency error, up to introducing new initial and interior consistency errors, whose size will be proven to be $O(\Delta x^{k_b + 1/2})$ hence the final result of Theorem 4.1.3. Let us make this argument precise.

The consistent expansion of the numerical solution (u_j^n) takes the form of a corrected version of (ω_j^n), involving now a boundary layer expansion $(v_j^n) \in \ell^2(-\infty, J)$ as for the above introductive example. The aim is to reduce the magnitude of the following convergence error at the boundary

$$\varepsilon_j^n := \omega_j^n - u_j^n + \Delta x^{k_b} v_j^n, \qquad j \leq J + p, \quad n \in \mathbb{N}. \tag{4.3.40}$$

The definition of $(v_j^n)_{j \leq J+p, n \in \mathbb{N}}$ is chosen so as to correct the error at the boundary. The simplest and best way to do so consists in chosing $(v_j^n)_{j \leq J+p, n \in \mathbb{N}}$ so as to get precisely in the ghost cells the relations $(D_-^{k_b} \varepsilon^n)_{J+\ell} = 0, 1 \leq \ell \leq p$. From now on, we formulate the problem in such a way to normalize the generating sequences according to the value of J. In view of the Lemma 4.3.4 and to the above discussion, the problem to be solved writes

$$v_j^n = \sum_{\sigma=1}^{\tau_-} \sum_{\nu=0}^{\mu_\sigma - 1} z_{\sigma,\nu}^n \rho_j^{(\sigma,\nu)}, \quad j \leq J + p, \quad n \in \mathbb{N}, \tag{4.3.41}$$

$$(D_-^{k_b} v^n)_{J+\ell} = -\frac{1}{\Delta x^{k_b}} (D_-^{k_b} \omega^n)_{J+\ell}, \qquad 1 \leq \ell \leq p, \quad n \in \mathbb{N}, \tag{4.3.42}$$

where the sequences $\rho^{(\sigma,\nu)}$ are defined in (4.3.37). Equivalently to (4.3.41), we again can look for the boundary layer corrector $(v_j^n)_{j \leq J+p, n \in \mathbb{N}}$ under the form

$$v_j^n = \sum_{\sigma=1}^{\tau_+} p_{n,\sigma}(j - J) \kappa_\sigma^{j-J}, \tag{4.3.43}$$

where $p_{n,\sigma} \in \mathbb{C}_{\mu_\sigma - 1}[X]$ for all index $1 \leq \sigma \leq \tau_+$. The existence of the corrector (v_j^n) is given by the following result. We recall that in the framework of Theorem 4.1.3, there holds $k_b < k$.

Lemma 4.3.5. *Consider the initial condition $f \in H^{k+1}((-\infty, L))$ extended to the whole real line $\mathbb{R}$. Then, the boundary layer problem (4.3.41)-(4.3.42) admits a unique solution $(v_j^n)_{j \leq J+p}, n \in \mathbb{N}$, and this solution satisfies the estimate*

$$\sup_{n \in \mathbb{N}} \left(\sum_{j \leq J} \Delta x |v_j^n|^2 \right)^{1/2} \leq C \Delta x^{1/2} \|f\|_{H^{k_b+1}((-\infty, L))}, \tag{4.3.44}$$

where the constant $C > 0$ is independent of $\Delta x > 0$, J, L and f.

Proof. Let us fix some integer $n \in \mathbb{N}$. The discrete solution $(u_j^n)_{j \leq J+p}$ to (4.1.14) solves the homogeneous boundary condition (4.1.8), thus equivalently to (4.3.42) one has to find the vector of coordinates $z \in \mathbb{C}^p$ solution to the linear system $A_{k_b} z + b = 0$ where $b = \Delta x^{-k_b}((D_-^{k_b} \omega^n)_{J+\ell})_{1 \leq \ell \leq p}$ and the $p \times p$ matrix A_{k_b} is defined as follows

$$A_{k_b} := \begin{pmatrix} (D_-^{k_b} \rho^{(1)})_1 & \dots & (D_-^{k_b} \rho^{(p)})_1 \\ \vdots & & \vdots \\ (D_-^{k_b} \rho^{(1)})_p & \dots & (D_-^{k_b} \rho^{(p)})_p \end{pmatrix}.$$

where we have relabeled the sequences $\rho^{(\sigma,\nu)}, 1 \leq \sigma \leq \tau_+$, $0 \leq \nu \leq \mu_\sigma - 1$ as $\rho^{(1)}, .., \rho^{(p)}$ in order to make the definition of A_{k_b} easier to read. The latter matrix is somehow the k_bth-order discrete derivative of the so-called confluent Vandermonde matrix. It seems possible to compute the determinant of A_0, see [47], and then to extend this result to higher values of k_b but we prefer to avoid such complicated computations. From the identity of dimensions, we shall just prove that the matrix A_{k_b} is one-to-one, in other words we shall prove that the problem (4.3.41)-(4.3.42), or equivalently (4.3.42)- (4.3.43), admits a trivial kernel.

Dealing with discrete derivatives of geometric sequences $(p_{n,\sigma}(j))_j$, polynomial sequences $(\kappa_\sigma^j)_j$ and of the product of such sequences, the divided difference algebra appears as a suitable tool in our analysis (for more details we refer the interested reader to the references [81, 77, 28]). For consistency in the notations, we recall hereafter the recursive definition of divided differences, but specified for the case of consecutive integer abscissae. Being given a sequence of complex numbers $(w_j)_{j \in \mathbb{Z}}$, one has

$$\begin{aligned} & w[j] = w_j, \quad j \in \mathbb{Z}, \\ & w[j-m, \dots, j] := \frac{1}{m}\Big(w[j-m+1, \dots, j] - w[j-m, \dots, j-1]\Big), \quad j \in \mathbb{Z},\ m \in \mathbb{N}^\star. \end{aligned} \tag{4.3.45}$$

Moreover, the quantity $(D_-^{k_b} w)_j$ is directly related to the divided difference $w[j-k_b, \dots, j]$ by the equality

$$(D_-^k w)_j = k_b! w[j-k_b, \dots, j], \qquad j \in \mathbb{Z}. \tag{4.3.46}$$

Importantly, we may also use the Leibniz formula for divided differences of products of two sequences

$$(w\widetilde{w})[j-k_b, \dots, j] = \sum_{m=0}^{k_b} w[j-k_b, \dots, j-m] \widetilde{w}[j-m, \dots, j], \qquad j \in \mathbb{Z}. \tag{4.3.47}$$

In terms of the D_- operator, using the relation (4.3.46), the Leibniz formula (4.3.47) rewrites under the more recognizable form

$$(D_-^{k_b}(w\widetilde{w}))_j = \sum_{m=0}^{k_b} \binom{k_b}{m} (D_-^{k_b-m} w)_{j-m}\ (D_-^m \widetilde{w})_j, \qquad j \in \mathbb{Z}.$$

Let us continue with the representation formula (4.3.38) of the solution to the boundary layer problem. Looking at the kernel of the linear problem (4.3.42), we have to find polynomials $(p_\sigma)_{1 \leq \sigma \leq \tau_+}$ with respective degrees less than or equal to $(\mu_\sigma - 1)_{1 \leq \sigma \leq \tau_+}$, satisfying the set of equations

$$\sum_{\sigma=1}^{\tau_+} \sum_{m=0}^{k_b} p_\sigma[\ell-k_b, \dots, \ell-m]\ \kappa_\sigma[\ell-m, \dots, \ell] = 0, \qquad 1 \leq \ell \leq p,$$

where we denote, with a slight abuse in the notation, κ_σ for the corresponding geometric sequence $(\kappa_m)_{m \in \mathbb{Z}}$, for any $\sigma = 1, \dots, \tau_+$. Actually, from the identity (4.3.46) and by induction on the integer m (or using (4.3.45)), it is easy to prove that the m-th order divided difference of κ_σ is given by

$$\kappa_\sigma[\ell-m,\dots,\ell]=\frac{1}{m!}(D_-^m\kappa_\sigma)_\ell=\frac{1}{m!}(1-\kappa_\sigma^{-1})^m\kappa_\sigma^\ell,\qquad 1\leq\ell\leq p.$$

Let us introduce, for any integer σ and any polynomial p_σ with degree less than or equal to $\mu_\sigma-1$, the following polynomial Q_σ also with degree less than or equal to $\mu_\sigma-1$

$$Q_\sigma(X)=\sum_{m=0}^{k_b}(1-\kappa_\sigma^{-1})^k p_\sigma[X-k_b,\dots,X-m]. \tag{4.3.48}$$

With these notations, the equations to solve now equivalently read

$$\sum_{\sigma=1}^{\tau_+}Q_\sigma(\ell)\kappa_\sigma^\ell=0,\qquad 1\leq\ell\leq p. \tag{4.3.49}$$

Actually, the above set of equations (4.3.49) exactly corresponds to the generalized Lagrange-Hermite interpolation problem, which is known to be invertible. To that aim it suffices to prove the injectivity of the linear application $\Phi:\mathbb{C}_{p-1}[X]\to\mathbb{C}^p$, mapping any polynomial P to the set of values $(P^{(\nu)}(\kappa_\sigma))_{1\leq\sigma\leq\tau_+,\ 0\leq\nu<\mu_\sigma}$. This follows from the count property (4.3.36). Thus one has necessarily $Q_\sigma=0$ for any $\sigma=1,\dots,\tau_+$. It remains to deduce any of the p_σ polynomial to be also zero.

Observe that for any k, $0\leq k\leq k_b$, and from the divided difference algebra, the polynomial $p_\sigma[X-k_b,\dots,X-k]$ has degree less than $\mu_\sigma-(k_b-k)$ (see (4.3.46)). Thus the highest degree polynomial involved in the sum (4.3.48) is $p_\sigma[X-k_b]$ (for $k=k_b$) that is necessarily zero and thus $p_\sigma=0$. The injectivity of the boundary layer problem (4.3.42)-(4.3.43) is proved, and therefore the matrix A_{k_b} is invertible.

As a consequence, there exist unique coefficients $(\beta_{\sigma,\nu,\ell})$ that depend only on the considered scheme and on k_b (but neither on the initial condition f, nor on the time index n), such that the solution to (4.3.41)-(4.3.42) has the form

$$v_j^n=\Delta x^{-k_b}\sum_{\sigma=1}^{\tau_+}\sum_{\nu=0}^{\mu_\sigma-1}\sum_{\ell=1}^{p}\beta_{\sigma,\nu,\ell}(D_-^{k_b}\omega^n)_{J+\ell}\rho_j^{(\sigma,\nu)}. \tag{4.3.50}$$

Using now triangular inequalities, we obtain, for some constant $C>0$, the upper bound

$$(v_j^n)^2\leq C\Delta x^{-2k_b}\sum_{\ell=1}^{p}\left((D_-^{k_b}\omega^n)_{J+\ell}\right)^2\sum_{\sigma=1}^{\tau_+}\sum_{\nu=0}^{\mu_\sigma-1}\left(\rho_j^{(\sigma,\nu)}\right)^2,\qquad j\leq J.$$

On the one side, we recall the definition (4.3.37) of the sequences $\rho^{(\sigma,\nu)}$ in Lemma 4.3.4, hence the estimate

$$\left(\sum_{j\leq J}\Delta x\sum_{\sigma=1}^{\tau_+}\sum_{\nu=0}^{\mu_\sigma-1}\left(\rho_j^{(\sigma,\nu)}\right)^2\right)^{1/2}\leq C\sqrt{\Delta x}, \tag{4.3.51}$$

where the constant $C>0$ is independent of J and Δx. On the other side, from Lemma 4.2.5 and the continuity of the reflection operator from $H^{k_b+1}((-\infty,L))$ to $H^{k_b+1}(\mathbb{R})$, we have the upper bound

$$\left|\left(D_-^{k_b}\omega^n\right)_{J+\ell}\right|\leq C\Delta x^{k_b}\|f\|_{H^{k_b+1}((-\infty,L))},\quad 1\leq\ell\leq p,\quad n\in\mathbb{N},$$

and thus the required estimate (4.3.44) follows. □

The interested reader will find in [36] a similar argument to the one developed in the proof of Lemma 4.3.5. In [36], the analysis of the determinant of the matrix A_{k_b} arises from the verification of the so-called Uniform Kreiss-Lopatinskii Condition (a condition whose significance is based on the work [43]. Let us now prove Theorem 4.1.3. The error $(\varepsilon_j^n)_{j\le J+p,n\in\mathbb{N}}$ introduced in (4.3.40), and fully defined through Lemma 4.3.5, satisfies the following set of equations[4]

$$\begin{cases} \varepsilon_j^0 \, = \Delta x^{k_b}\, v_j^0\,, & j \le J\,, \\ (D_-^{k_b}\varepsilon^n)_{J+\ell} \, = \, 0\,, & 0 \le n \le T/\Delta t\,, \quad 1 \le \ell \le p\,, \\ \varepsilon_j^{n+1} \, = \, \displaystyle\sum_{\ell=-r}^{p} a_\ell\, \varepsilon_{j+\ell}^n + \Delta t\, F_j^{n+1}\,, & 0 \le n \le T/\Delta t - 1\,, \quad j \le J\,. \end{cases} \tag{4.3.52}$$

Here above, the consistency error F_j^{n+1} consists of two terms: a first one coming from the usual interior consistency error denoted e_j^{n+1}, and a second one coming from the time evolution of the boundary layer corrector denoted δ_j^{n+1}. In other words, we split $F_j^{n+1} = e_j^{n+1} + \delta_j^{n+1}$ with

$$e_j^{n+1} \, := \, \frac{1}{\Delta t}\left(\omega_j^{n+1} - \sum_{\ell=-r}^{p} a_\ell\, \omega_{j+\ell}^n\right)\,, \quad \text{and} \quad \delta_j^{n+1} \, := \, \frac{\Delta x^{k_b}}{\Delta t}\left(v_j^{n+1} - \sum_{\ell=-r}^{p} a_\ell\, v_{j+\ell}^n\right)\,.$$

Considering the scheme (4.3.52), the error $(\varepsilon_j^n)_{j\le J+p,0\le n\le T/\Delta t}$ obeys the stability estimate applicable in the case of the homogeneous Neumann boundary condition, see Proposition 4.2.3

$$\sup_{0\le n\le T/\Delta t} \sum_{j\le J} \Delta x\,(\varepsilon_j^n)^2 \, \le \, C\,\left\{\sum_{j\le J} \Delta x\,(\varepsilon_j^0)^2 + T^2 \sup_{1\le n\le T/\Delta t} \sum_{j\le J} \Delta x\,(F_j^n)^2\right\}\,. \tag{4.3.53}$$

It therefore remains to estimate the initial and interior consistency errors in (4.3.52)

- The initial consistency error. Estimating the initial condition $(\varepsilon_j^0)_{j\le J}$ directly follows from the estimate (4.3.44) in Lemma 4.3.5

$$\sum_{j\le J} \Delta x\,(\varepsilon_j^0)^2 \, \le \, C\,\Delta x^{2\,k_b+1}\,\|f\|^2_{H^{k_b+1}((-\infty,L))}\,.$$

- The interior consistency error. I. Estimating the interior consistency error (e_j^n) related to the projected exact solution (ω_j^n) has already been achieved as in (4.2.24)-(4.2.29) so we just report the result

$$\sup_{1\le n\le T/\Delta t} \sum_{j\le J} \Delta x\,(e_j^n)^2 \, \le \, C\,\Delta x^{2\,k}\,\|f\|^2_{H^{k+1}((-\infty,L))}\,.$$

- The interior consistency error. II. Estimation of the new error term related to (δ_j^n). Observe that, first due to the steady states decomposition from Lemma 4.3.4 and then using successively (4.3.41) and (4.3.50), the interior consistency error arising from the boundary layer corrector rewrites as

$$\delta_j^{n+1} = \frac{\Delta x^{k_b}}{\Delta t}(v_j^{n+1} - v_j^n) = \frac{1}{\Delta t}\sum_{\sigma=1}^{\tau_+}\sum_{\nu=0}^{\mu_\sigma-1}\sum_{\ell=1}^{p} \beta_{\sigma,\nu,\ell}(D_-^{k_b}(\omega^{n+1}-\omega^n))_{J+\ell}\rho_j^{(\sigma,\nu)},$$

Thus, from Cauchy-Schwarz inequalities, there exists a constant C such that

$$\sum_{j\le J} \Delta x|\delta_j^{n+1}|^2 \le C\Delta x \sup_{1\le\ell\le p}\left(D_-^{k_b}\left(\frac{\omega^{n+1}-\omega^n}{\Delta t}\right)_{J+\ell}\right)^2.$$

[4] Here we use $u_j^0 = w_j^0$ for $j \le J$

In the above formula, the discrete in time derivative of ω_j^n rewrites, for any $j \leq J + p$ as

$$\frac{\omega_j^{n+1} - \omega_j^n}{\Delta t} = \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} \frac{f(y - at^n - a\Delta t) - f(y - at^n)}{\Delta t} \, \mathrm{d}y$$
$$= \frac{1}{\Delta x} \int_{x_{j-1}}^{x_j} \underbrace{\frac{1}{\Delta t} \int_{y-at^n}^{y-at^n-a\Delta t} f'(z) \, \mathrm{d}z}_{:=F(y)} \, \mathrm{d}y.$$

Since $f \in H^{k+1}(\mathbb{R})$ with $k > k_b$, we have at least $f \in H^{k_b+2}(\mathbb{R})$ and therefore $F \in H^{k_b+1}(\mathbb{R})$ with

$$\|F^{(k_b+1)}\|_{L^2(\mathbb{R})} \leq a^2 \|f^{(k_b+2)}\|_{L^2(\mathbb{R})},$$

from which we deduce, using in addition Lemma 4.2.5:

$$\left| D_-^{k_b} \left(\frac{\omega^{n+1} - \omega^n}{\Delta t} \right)_{J+\ell} \right| \leq C \Delta x^{k_b} \|f\|_{H^{k_b+2}(\mathbb{R})}, \quad 1 \leq \ell \leq p.$$

Thus, using again the upper bound (4.3.51), the above estimate and the H^{k_b+2}-continuity of the extension operator, one has:

$$\sup_{1 \leq n \leq T/\Delta t} \sum_{j \leq J} \Delta x \left(\delta_j^{n+1} \right)^2 \leq C \Delta x^{2k_b+1} \|f\|^2_{H^{k+1}((-\infty,L))}. \tag{4.3.54}$$

Let us now come back to the stability estimate (4.3.53) and use the three above consistency estimates to get (recall $T \geq 1$ and $k_b < k$)

$$\sup_{0 \leq n \leq T/\Delta t} \left(\sum_{j \leq J} \Delta x \, (\varepsilon_j^n)^2 \right)^{1/2} \leq C \, T \, \Delta x^{k_b+1/2} \, \|f\|_{H^{k+1}((-\infty,L))} \, .$$

From the constructive formula for the boundary layer corrector (v_j^n), we have derived the bound (4.3.44) which, by the triangle inequality, yields the convergence estimate (recall $\varepsilon_j^n = \omega_j^n - u_j^n + \Delta x^{k_b} v_j^n$)

$$\sup_{0 \leq n \leq T/\Delta t} \left(\sum_{j \leq J} \Delta x \, (u_j^n - \omega_j^n)^2 \right)^{1/2} \leq C \, T \, \Delta x^{k_b+1/2} \, \|f\|_{H^{k+1}} \, .$$

Using now the (crude) estimate

$$\sup_{j \leq J} |b_j| \, \leq \, \Delta x^{-1/2} \left(\sum_{j \leq J} \Delta x \, b_j^2 \right)^{1/2} ,$$

we complete the proof of Theorem 4.1.3.

4.4 Numerical experiments

4.4.1 The Lax-Wendroff scheme

We report in this paragraph on various numerical experiments with the Lax-Wendroff scheme (4.1.3) (which corresponds to $p = r = 1$). Assumption 4.1.1 is satisfied provided that $\lambda \, a \leq 1$, and the order k equals 2. In all what follows, we choose $a = 1$ and $\lambda = 5/6$. The interval length is $L = 6$ and the final time T equals 8. The initial condition is $f(x) = \sin x$ and the boundary source term is $g(t) = -\sin t$ so that the exact solution to (4.1.1) is $u(t,x) = \sin(x-t)$. Figure 4.4.1 represents the initial condition $f(x)$ on a grid with 50 cells on $(0,6)$. In Figure 4.4.2, we plot the numerical solutions at $t = 8$ obtained with $k_b = 0$ (homogeneous Dirichlet outflow condition), $k_b = 1$ (first order outflow extrapolation condition) and $k_b = 2$ (second order outflow extrapolation condition). As expected, the Dirichlet condition shows a larger boundary layer, and, especially, the solution with $k_b = 2$ is much nearer to the exact solution than the others.

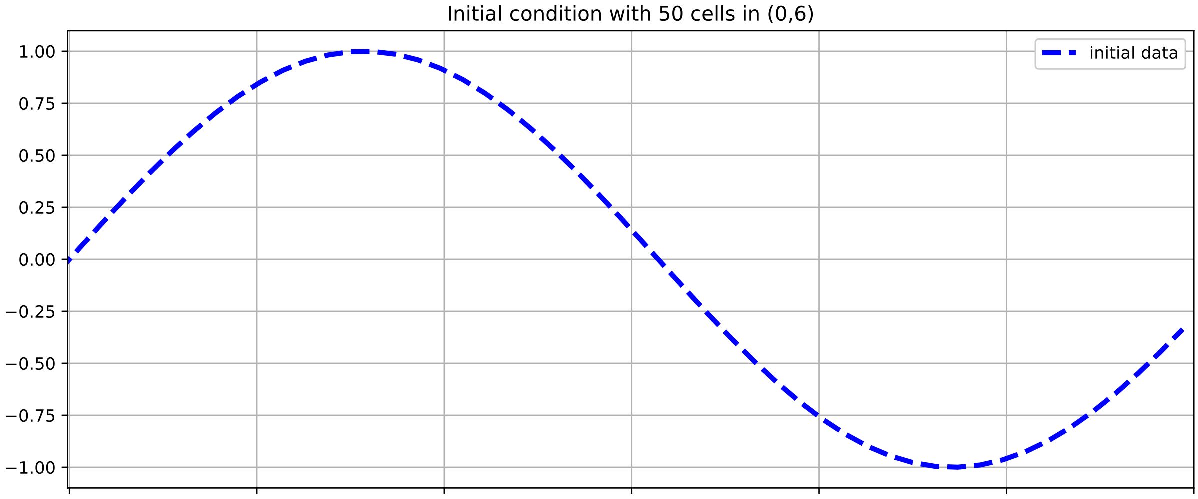

Figure 4.4.1: Initial condition $f(x)$ with 50 cells in $(0,6)$.

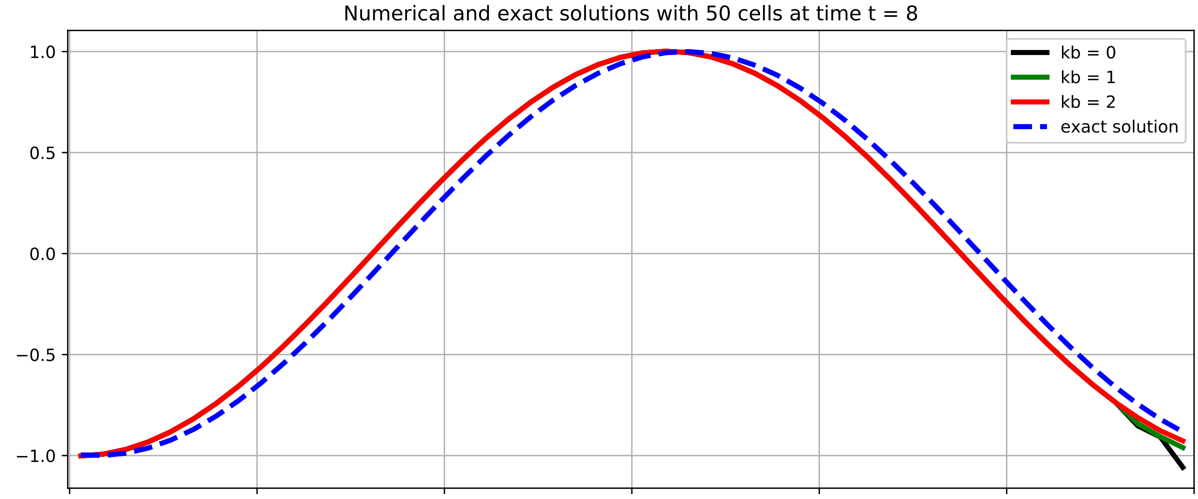

Figure 4.4.2: Numerical and exact solutions at time $t = 8$ with 50 cells in $(0,6)$.

Let us now consider the error of the schemes. With the values of J reported in Table 4.1 below, we first implement the Lax-Wendroff scheme (4.1.3) with the following numerical boundary conditions

$$u^n_{J+1} = u^n_J \,, \quad \text{(first order outflow extrapolation condition)},$$

$$u^n_0 = \begin{cases} -\sin t^n \,, & \text{(Dirichlet inflow condition (4.1.6))}, \\ -\sin t^n - (\Delta x/2)\cos t^n \,, & \text{(inverse Lax-Wendroff inflow condition (4.1.12))}. \end{cases}$$

The errors, as measured in the statement of Theorem 4.1.2, are reported in Table 4.1 below for each of the two cases (either the Dirichlet inflow condition (4.1.6) or the inverse Lax-Wendroff inflow condition (4.1.12)). In either case, the observed convergence rate is 1 since increasing J by a factor 2 decreases the error of the same factor 2. This behavior is fully justified by Theorem 4.1.3 since we have $k_b < k$ here.

Number of cells J	Dirichlet inflow condition	Inverse Lax-Wendroff inflow condition
1000	$4.1 \cdot 10^{-3}$	$5.1 \cdot 10^{-4}$
2000	$2.1 \cdot 10^{-3}$	$2.5 \cdot 10^{-4}$
4000	$1.1 \cdot 10^{-3}$	$1.3 \cdot 10^{-4}$
8000	$5.3 \cdot 10^{-4}$	$6.3 \cdot 10^{-5}$

Table 4.1: The $\ell^\infty_{n,j}$ error for the Lax-Wendroff scheme (4.1.3) with first order outflow extrapolation and either the Dirichlet, or inverse Lax-Wendroff, inflow condition.

Secondly, we now turn to the second order outflow extrapolation condition:

$$\begin{aligned} u_{J+1}^n &= 2\,u_J^n - u_{J-1}^n\,, \quad \text{(second order outflow extrapolation condition (4.1.5))},\\ u_0^n &= \begin{cases} -\sin t^n\,, & \text{(Dirichlet inflow condition (4.1.6))},\\ -\sin t^n - (\Delta x/2)\,\cos t^n\,, & \text{(inverse Lax-Wendroff inflow condition (4.1.12))}. \end{cases} \end{aligned}$$

The errors, as measured in the statement of Theorem 4.1.2, are reported in Table 4.2 below for each of the two cases (either the Dirichlet inflow condition (4.1.6) or the inverse Lax-Wendroff inflow condition (4.1.12)). For the Dirichlet inflow condition, the observed convergence rate is 1 again (despite the more accurate outflow treatment), but one recovers the convergence rate 2 with the inverse Lax-Wendroff inflow condition (4.1.12). However, proving rigorously that this numerical scheme converges with the rate 2 in the maximum norm might be very difficult (it might actually even be wrong !), even for smooth data, since the Lax-Wendroff scheme is known to be unstable in $\ell^\infty(\mathbb{Z})$. Improving the convergence rate 3/2 of Theorem 4.1.2 in the case of the Lax-Wendroff scheme with second order extrapolation outflow condition is left to a future work.

Number of cells J	Dirichlet inflow condition	Inverse Lax-Wendroff inflow condition
1000	$3.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-5}$
2000	$1.8 \cdot 10^{-3}$	$2.9 \cdot 10^{-6}$
4000	$9.3 \cdot 10^{-4}$	$7.3 \cdot 10^{-7}$
8000	$4.7 \cdot 10^{-4}$	$1.8 \cdot 10^{-7}$

Table 4.2: The $\ell^\infty_{n,j}$ error for the Lax-Wendroff scheme (4.1.3) with second order outflow extrapolation (4.1.5) and either the Dirichlet or inverse Lax-Wendroff inflow condition.

4.4.2 The $O3$ scheme

Let us now consider the $O3$ scheme, which is a convex combination of the Lax-Wendroff and Beam-Warming schemes. It reads

$$u_j^{n+1} = a_{-2}\,u_{j-2}^n + a_{-1}\,u_{j-1}^n + a_0\,u_j^n + a_1\,u_{j+1}^n\,, \quad n \in \mathbb{N}\,, \quad j = 1, \dots, J\,,$$

with

$$\begin{aligned} a_{-2} &:= -\frac{\lambda\,a}{6}\,\big(1-(\lambda\,a)^2\big)\,, \quad a_{-1} := \frac{\lambda\,a}{2}\,(1+\lambda\,a)\,(2-\lambda\,a)\,,\\ a_0 &:= \frac{1}{2}\,\big(1-(\lambda\,a)^2\big)\,(2-\lambda\,a)\,, \quad a_1 := -\frac{\lambda\,a}{6}\,(1-\lambda\,a)\,(2-\lambda\,a)\,. \end{aligned}$$

The reader can verify that Assumption 4.1.1 is satisfied provided that $\lambda\, a \leq 1$, and the order k equals 3 ($r = 2$ and $p = 1$ here). To maintain third order accuracy, we implement the latter scheme with the following boundary conditions

$$\begin{aligned} u^n_{J+1} &= 3\, u^n_J - 3\, u^n_{J-1} + u^n_{J-2}\,, \quad \text{(third order outflow extrapolation condition, } k_b = 3\text{)},\\ u^n_0 &= -\sin t^n - (\Delta x/2)\,\cos t^n + (\Delta x^2/6)\,\sin t^n\,, \quad \text{(inverse Lax-Wendroff inflow condition (4.1.12))},\\ u^n_{-1} &= -\sin t^n - (3\,\Delta x/2)\,\cos t^n + (7\,\Delta x^2/6)\,\sin t^n\,, \quad \text{(inverse Lax-Wendroff inflow condition (4.1.12))}. \end{aligned}$$

Number of cells J	Inverse Lax-Wendroff inflow condition
1000	$2.1 \cdot 10^{-8}$
2000	$2.6 \cdot 10^{-9}$
4000	$3.3 \cdot 10^{-10}$

Table 4.3: The $\ell^\infty_{n,j}$ error for the $O3$ scheme (4.1.3) with third order outflow extrapolation and the inverse Lax-Wendroff inflow condition.

The measured errors are reported in Table 4.3. They correspond to a rate of convergence 3. Let us observe that the $O3$ scheme is known to be stable in $\ell^\infty(\mathbb{Z})$, see [87, 29], hence there is a genuine hope of proving rigorously that this rate of convergence does indeed hold (for smooth compatible data). Such a justification is also left to a future work.

4.5 Appendix: A discrete integration by parts

In this appendix, we study the discrete integration by parts result for the Lax-Wendroff and $O3$ schemes. Let us introduce

$$\nu = \lambda a.$$

4.5.1 The Lax-Wendroff scheme

In order to estimate the discrete integration by parts, the following lemma is useful

Lemma 4.5.1. *Assume that the exact solution to (4.1.1) is approximated by means of the Lax-Wendroff scheme:*

$$u^{n+1}_j = \frac{1}{2}\nu(\nu+1)u^n_{j-1} + (1-\nu^2)u^n_j + \frac{1}{2}\nu(\nu-1)u^n_{j+1}. \tag{4.5.1}$$

Then, there exist real numbers A, B and a quadratic form Q on $\mathbb{R}^2$ such that

$$\begin{aligned} \left|u^{n+1}_j\right|^2 - \left|u^n_j\right|^2 = A\left(u^n_j - u^n_{j-1}\right)^2 + B\left(u^n_{j+1} - 2u^n_j + u^n_{j-1}\right)^2 \\ + Q(u^n_j, u^n_{j+1} - u^n_j) - Q(u^n_{j-1}, u^n_j - u^n_{j-1}). \end{aligned} \tag{4.5.2}$$

Proof. Firstly, we find the real numbers A and B. Let us define the Fourier transform of $(u^n_j)_{j\in\mathbb{Z}}$

$$\widehat{u}^n(\xi) = \sum_{j\in\mathbb{Z}} u^n_j e^{-ij\xi}.$$

Therefore, the numerical scheme (4.5.1) becomes

$$\begin{aligned}\widehat{u}^{n+1}(\xi) &= \left(\frac{1}{2}(\nu^2+\nu)e^{-i\xi} + (1-\nu^2) + \frac{1}{2}(\nu^2-\nu)e^{i\xi}\right)\widehat{u}^n(\xi)\\ &= \left(\nu^2\cos(\xi) + 1 - \nu^2 - i\nu\sin(\xi)\right)\widehat{u}^n(\xi).\end{aligned}$$

Then,

$$\begin{aligned}\left|\widehat{u}^{n+1}(\xi)\right|^2 &= \left(\left(\nu^2\cos(\xi)+1-\nu^2\right)^2 + \nu^2\sin^2(\xi)\right)\left|\widehat{u}^n(\xi)\right|^2\\ &= \left(1+\nu^2(\nu^2-1)(\cos(\xi)-1)^2\right)\left|\widehat{u}^n(\xi)\right|^2\\ &= \left(1+4\nu^2(\nu^2-1)\sin^4\left(\frac{\xi}{2}\right)\right)\left|\widehat{u}^n(\xi)\right|^2.\end{aligned}$$

By an application of the Plancherel's theorem and the above formula, we have

$$\begin{aligned}\sum_{j\in\mathbb{Z}}|u_j^{n+1}|^2 - \sum_{j\in\mathbb{Z}}|u_j^n|^2 &= \frac{1}{2\pi}\int_0^{2\pi}\left|\widehat{u}^{n+1}(\xi)\right|^2 d\xi - \frac{1}{2\pi}\int_0^{2\pi}\left|\widehat{u}^n(\xi)\right|^2 d\xi\\ &= \frac{1}{2\pi}\int_0^{2\pi} 4\nu^2(\nu^2-1)\sin^4\left(\frac{\xi}{2}\right)\left|\widehat{u}^n(\xi)\right|^2 d\xi.\end{aligned} \tag{4.5.3}$$

Now, summing (4.5.2) on $\mathbb{Z}$ and using the Plancherel's theorem, we get

$$\begin{aligned}\sum_{j\in\mathbb{Z}}|u_j^{n+1}|^2 - \sum_{j\in\mathbb{Z}}|u_j^n|^2 &= \frac{1}{2\pi}\int_0^{2\pi}\left(A|1-e^{-i\xi}|^2 + B|e^{i\xi}-2+e^{-i\xi}|^2\right)\left|\widehat{u}^n(\xi)\right|^2 d\xi\\ &= \frac{1}{2\pi}\int_0^{2\pi} 4A\sin^2\left(\frac{\xi}{2}\right)\left|\widehat{u}^n(\xi)\right|^2 d\xi + \frac{1}{2\pi}\int_0^{2\pi} 16B\sin^4\left(\frac{\xi}{2}\right)\left|\widehat{u}^n(\xi)\right|^2 d\xi.\end{aligned} \tag{4.5.4}$$

From (4.5.3) and (4.5.4), we can find

$$A = 0, \quad B = \frac{1}{4}\nu^2(\nu^2-1). \tag{4.5.5}$$

Secondly, we assume that the quadratic form Q on $\mathbb{R}^2$ has a form

$$Q(x,y) = \alpha x^2 + \beta xy + \gamma y^2.$$

Thus, the formula (4.5.2) becomes

$$\begin{aligned}|u_j^{n+1}|^2 - |u_j^n|^2 &= (B+\gamma)(u_{j+1}^n)^2 + (A+4B+\alpha-\beta)(u_j^n)^2 + (A+B-\alpha+\beta-\gamma)(u_{j-1}^n)^2\\ &\quad + (-4B+\beta-2\gamma)u_{j+1}^n u_j^n + 2Bu_{j+1}^n u_{j-1}^n + (-2A-4B-\beta+2\gamma)u_j^n u_{j-1}^n.\end{aligned} \tag{4.5.6}$$

On the other hand, from the numerical scheme (4.5.1), one gets

$$\begin{aligned}|u_j^{n+1}|^2 - |u_j^n|^2 &= \left(\frac{1}{2}\nu(\nu+1)u_{j-1}^n + (1-\nu^2)u_j^n + \frac{1}{2}\nu(\nu-1)u_{j+1}^n\right)^2 - (u_j^n)^2\\ &= \frac{1}{4}\nu^2(\nu-1)^2(u_{j+1}^n)^2 + \left((1-\nu^2)^2-1\right)(u_j^n)^2 + \frac{1}{4}\nu^2(\nu+1)^2(u_{j-1}^n)^2\\ &\quad + \nu(\nu-1)(1-\nu^2)u_{j+1}^n u_j^n + \frac{1}{2}\nu^2(\nu^2-1)u_{j+1}^n u_{j-1}^n + \nu(\nu+1)(1-\nu^2)u_j^n u_{j-1}^n.\end{aligned} \tag{4.5.7}$$

According to (4.5.6) and (4.5.7), we obtain the following equations

$$\begin{cases} B+\gamma = \dfrac{1}{4}\nu^2(\nu-1)^2 \\ A+4B+\alpha-\beta = (1-\nu^2)^2-1 \\ A+B-\alpha+\beta-\gamma = \dfrac{1}{4}\nu^2(\nu+1)^2 \\ -4B+\beta-2\gamma = \nu(\nu-1)(1-\nu^2) \\ 2B = \dfrac{1}{2}\nu^2(\nu^2-1) \\ -2A-4B-\beta+2\gamma = \nu(\nu+1)(1-\nu^2). \end{cases}$$

From the values of A and B in (4.5.5), after some calculations, we can find

$$\alpha = -\nu, \quad \beta = -\nu(1-\nu), \quad \gamma = \frac{1}{2}\nu^2(1-\nu).$$

This ends the proof of Lemma 4.5.1. □

4.5.2 The $O3$ scheme

In order to estimate the discrete integration by parts, we study the following lemma

Lemma 4.5.2. *Assume that the exact solution to (4.1.1) is approximated by means of the O3 scheme:*

$$u_j^{n+1} = -\frac{1}{6}\nu(1-\nu^2)u_{j-2}^n + \frac{1}{2}\nu(\nu+1)(2-\nu)u_{j-1}^n + \frac{1}{2}(1-\nu^2)(2-\nu)u_j^n - \frac{1}{6}\nu(1-\nu)(2-\nu)u_{j+1}^n. \tag{4.5.8}$$

Then, there exist real numbers A, B, C and a quadratic form Q on $\mathbb{R}^3$ such that

$$\begin{aligned} \left|u_j^{n+1}\right|^2 - \left|u_j^n\right|^2 = {} & A\left(u_j^n - u_{j-1}^n\right)^2 + B\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)^2 + C(u_{j+1}^n - 3u_j^n + 3u_{j-1}^n - u_{j-2}^n)^2 \\ & + Q(u_j^n, u_{j+1}^n - u_j^n, u_{j+1}^n - 2u_j^n + u_{j-1}^n) - Q(u_{j-1}^n, u_j^n - u_{j-1}^n, u_j^n - 2u_{j-1}^n + u_{j-2}^n). \end{aligned} \tag{4.5.9}$$

Proof. Firstly, we find the real numbers A, B, C. By taking the Fourier transform of $(u_j^n)_{j\in\mathbb{Z}}$, the numerical scheme (4.5.8) becomes

$$\begin{aligned} \widehat{u}^{n+1}(\xi) = \Bigg(& -\frac{1}{6}\nu(1-\nu^2)e^{-2i\xi} + \frac{1}{2}\nu(\nu+1)(2-\nu)e^{-i\xi} \\ & + \frac{1}{2}(1-\nu^2)(2-\nu) - \frac{1}{6}\nu(1-\nu)(2-\nu)e^{i\xi}\Bigg)\widehat{u}^n(\xi). \end{aligned}$$

Then,

$$\begin{aligned} |\widehat{u}^{n+1}(\xi)|^2 &= \left(1 - \frac{1}{9}\nu(2-\nu)(1-\nu^2)(\cos(\xi)-1)^2\left(4\nu(1-\nu)\sin^2\left(\frac{\xi}{2}\right)+3\right)\right)|\widehat{u}^n(\xi)|^2 \\ &= \left(1 - \frac{16}{9}\nu^2(1-\nu)^2(2-\nu)(\nu+1)\sin^6\left(\frac{\xi}{2}\right) - \frac{4}{3}\nu(1-\nu^2)(2-\nu)\sin^4\left(\frac{\xi}{2}\right)\right)|\widehat{u}^n(\xi)|^2. \end{aligned}$$

By an application of the Plancherel's theorem and the above formula, we have

$$\begin{aligned} \sum_{j\in\mathbb{Z}}|u_j^{n+1}|^2 - \sum_{j\in\mathbb{Z}}|u_j^n|^2 = & \frac{1}{2\pi}\int_0^{2\pi}\left|\widehat{u}^{n+1}(\xi)\right|^2 d\xi - \frac{1}{2\pi}\int_0^{2\pi}\left|\widehat{u}^n(\xi)\right|^2 d\xi \\ = & -\frac{1}{2\pi}\int_0^{2\pi}\frac{16}{9}\nu^2(1-\nu)^2(2-\nu)(\nu+1)\sin^6\left(\frac{\xi}{2}\right)\left|\widehat{u}^n(\xi)\right|^2 d\xi \\ & -\frac{1}{2\pi}\int_0^{2\pi}\frac{4}{3}\nu(1-\nu^2)(2-\nu)\sin^4\left(\frac{\xi}{2}\right)\left|\widehat{u}^n(\xi)\right|^2 d\xi. \end{aligned} \tag{4.5.10}$$

Now, summing (4.5.8) on $\mathbb{Z}$ and using the Plancherel's theorem, we get

$$\begin{aligned}&\sum_{j\in\mathbb{Z}}|u_j^{n+1}|^2-\sum_{j\in\mathbb{Z}}|u_j^n|^2\\ =&\frac{1}{2\pi}\int_0^{2\pi}\left(A|1-e^{-i\xi}|^2+B|e^{i\xi}-2+e^{-i\xi}|^2+C|e^{i\xi}-3+3e^{-i\xi}-e^{-2i\xi}|^2\right)|\widehat{u}^n(\xi)|^2\,d\xi\\ =&\frac{1}{2\pi}\int_0^{2\pi}4A\sin^2\left(\frac{\xi}{2}\right)|\widehat{u}^n(\xi)|^2\,d\xi+\frac{1}{2\pi}\int_0^{2\pi}16B\sin^4\left(\frac{\xi}{2}\right)|\widehat{u}^n(\xi)|^2\,d\xi\\ +&\frac{1}{2\pi}\int_0^{2\pi}64C\sin^6\left(\frac{\xi}{2}\right)|\widehat{u}^n(\xi)|^2\,d\xi.\end{aligned}\tag{4.5.11}$$

From (4.5.10) and (4.5.11), we can find

$$A=0,\quad B=-\frac{1}{12}\nu(1-\nu^2)(2-\nu),\quad C=-\frac{1}{36}\nu^2(1-\nu)^2(2-\nu)(\nu+1).\tag{4.5.12}$$

Secondly, we assume that the quadratic form Q on $\mathbb{R}^3$ has a form

$$Q(x,y,z)=\alpha x^2+\beta y^2+\gamma z^2+\delta xy+\epsilon xz+\rho yz.$$

Thus, the formula (4.5.9) becomes

$$\begin{aligned}|u_j^{n+1}|^2-|u_j^n|^2=&A(u_j^n-u_{j-1}^n)^2+B(u_{j+1}^n-2u_j^n+u_{j-1}^n)^2+C(u_{j+1}^n-3u_j^n+3u_{j-1}^n-u_{j-2}^n)^2\\ &+\alpha(u_j^n)^2+\beta(u_{j+1}^n-u_j^n)^2+\gamma(u_{j+1}^n-2u_j^n+u_{j-1}^n)^2\\ &+\delta u_j^n(u_{j+1}^n-u_j^n)+\epsilon u_j^n(u_{j+1}^n-2u_j^n+u_{j-1}^n)+\rho(u_{j+1}^n-u_j^n)(u_{j+1}^n-2u_j^n+u_{j-1}^n)\\ &-\alpha(u_{j-1}^n)^2-\beta(u_j^n-u_{j-1}^n)^2-\gamma(u_j^n-2u_{j-1}^n+u_{j-2}^n)^2\\ &-\delta u_{j-1}^n(u_j^n-u_{j-1}^n)-\epsilon u_{j-1}^n(u_j^n-2u_{j-1}^n+u_{j-2}^n)-\rho(u_j^n-u_{j-1}^n)(u_j^n-2u_{j-1}^n+u_{j-2}^n).\end{aligned}\tag{4.5.13}$$

On the other hand, from the numerical scheme (4.5.8), one gets

$$\begin{aligned}|u_j^{n+1}|^2-|u_j^n|^2=&\Big(-\frac{1}{6}\nu(1-\nu^2)u_{j-2}^n+\frac{1}{2}\nu(\nu+1)(2-\nu)u_{j-1}^n\\ &+\frac{1}{2}(1-\nu^2)(2-\nu)u_j^n-\frac{1}{6}\nu(1-\nu)(2-\nu)u_{j+1}^n\Big)^2-(u_j^n)^2.\end{aligned}\tag{4.5.14}$$

According to (4.5.12)-(4.5.14), after some calculations, we can find

$$\begin{aligned}&\alpha=-\nu,\quad\beta=-\frac{1}{6}\nu(1-\nu)(1-2\nu),\quad\gamma=-\frac{1}{12}\nu^2(1-\nu)^2(\nu+1),\\ &\delta=-\nu(1-\nu),\quad\epsilon=\frac{1}{3}\nu(1-\nu^2),\quad\rho=\frac{1}{3}\nu(1-\nu)^2(\nu+1).\end{aligned}$$

□

Chapter 5

Stability of stationary solution for nonlinear relaxation balance laws

Stability of stationary solutions of singular systems of balance laws have been analyzed in [80]. The goal of this chapter is to extend this result to the case of nonlinear relaxation balance laws. To investigate the stability of stationary solutions of these systems, we first assume that these systems are endowed with a partially convex entropy. We also define an entropy process solution which generalizes the concept of entropy weak solutions. After that, we construct an associated relative entropy which allows to compare two states which share the same geometric data. The entropy dissipation condition is the key to prove the stability of some stationary states within entropy process solutions. Besides, we also consider another assumption, see in [91], to investigate the asymptotic stability of stationary solution of nonlinear relaxation balance laws.

5.1 Introduction

In this chapter, we consider non-conservation systems of the form

$$\begin{cases} \partial_t u(t,x) + \sum_{i=1}^{d} \partial_i f_i(u,\alpha)(t,x) + \sum_{i=1}^{d} s_i(u,\alpha)(t,x)\partial_i \alpha(t,x) = r(u,\alpha)(t,x), \\ \partial_t \alpha(t,x) = 0. \end{cases} \tag{5.1.1}$$

System (5.1.1) is set on the whole space $x \in \mathbb{R}^d$, and for any time $t \in [0,T)$, $T > 0$. The notation ∂_i denotes the partial derivative with respect to x_i. We assume that there exists a convex bounded subset of $\mathbb{R}^N$, denoted by Ω and called set of the admissible states such that

$$u(t,x) \in \Omega, \quad \forall\,(t,x) \in [0,T) \times \mathbb{R}^d$$

and

$$\alpha(t,x) \in \mathbb{R}, \quad \forall\,(t,x) \in [0,T) \times \mathbb{R}^d.$$

We also define

$$\begin{aligned} &f_i : \Omega \times \mathbb{R} \to \mathbb{R}^N, \quad s_i : \Omega \times \mathbb{R} \to \mathbb{R}^N, \\ &r : \Omega \times \mathbb{R} \to \mathbb{R}^N. \end{aligned}$$

The second equation in (5.1.1) means that α is time-independent, so that this variable is a data, as soon as an initial condition is associated with (5.1.1)

$$\begin{cases} u(0,x) &= u_0(x), \\ \alpha(0,x) &= \alpha(x), \end{cases} \quad \text{for } x \in \mathbb{R}^d. \tag{5.1.2}$$

Let us mention that the third term of the left-hand side of (5.1.1) is studied as a source term if α is smooth. In this chapter, we also consider the analysis applied to non-smooth α, and the term $\sum_{i=1}^{d} s_i(u,\alpha)\partial_i\alpha$ is a non-conservative product. This analysis also applies to the case of systems of conservation laws with discontinuous flux f depending on α, as studied in [1, 49, 55]

System (5.1.1) is endowed with an entropy pair (η, F), which depends on (u,α) and satisfies the following assumptions

(H1) The function $\eta = \eta(u,\alpha) \in \mathcal{C}^2(\Omega \times \mathbb{R}, \mathbb{R})$ is convex with respect to its first variable and there exist two positive constants $\beta_0 < \beta_1$ such that

$$\sigma\left(\partial_u^2\eta\right) \subset [\beta_0, \beta_1], \quad \text{on } \Omega \times \mathbb{R}, \tag{5.1.3}$$

where σ denotes the matrix spectrum and ∂_u denotes the differential with respect to the variables u.

(H2) There exists an entropy flux $F = (F_i(u,\alpha))_{1\le i\le d}$ such that

$$\forall 1 \le i \le d, \quad \partial_u\eta\partial_u f_i = \partial_u F_i \quad \text{and} \quad \partial_u\eta(\partial_\alpha f_i + s_i) = \partial_\alpha F_i.$$

The convexity assumption (5.1.3) is assumed on the whole space $\Omega \times \mathbb{R}$. It may be restrictive and, without loss of generality, we assume that $\eta(u,.) \ge 0$, for all $u \in \Omega$. The existence of the entropy flux F amounts to assume the integrability condition (see [35])

$$\partial_u^2\eta\partial_u f_i = \partial_u f_i^T \partial_u^2\eta. \tag{5.1.4}$$

Let us introduce the quantity L_f by

$$L_f = \sup_{1\le i\le d} \sup_{(u,v)\in\Omega^2} \sup_{w\in\mathbb{R}^N\setminus\{0\}} \left| \frac{w^T\partial_u^2\eta(v,.)\partial_u f_i(u,.)w}{w^T\partial_u^2\eta(v,.)w} \right|. \tag{5.1.5}$$

Remark 5.1.1. *Notice that, in view of (5.1.4), the matrix $\partial_u f_i$ is self-adjoint for the scalar product $\langle w, v\rangle_u = w^T\partial_u^2\eta(u,.)v$. Therefore, the Rayleigh quotient*

$$\sup_{w\in\mathbb{R}^N\setminus\{0\}} \left| \frac{w^T\partial_u^2\eta(u,.)\partial_u f_i(u,.)w}{w^T\partial_u^2\eta(u,.)w} \right| = \sup_{w\in\mathbb{R}^N\setminus\{0\}} \frac{\langle w, \partial_u f_i(u,.)w\rangle_u}{\langle w, w\rangle_u} \tag{5.1.6}$$

provides exactly the largest eigenvalue in absolute value of $\partial_u f_i(u,.)$. The situation in (5.1.5) is more intricate than in (5.1.6) since u might be different of v, but the quantity L_f is bounded in view of the boundedness of Ω and of the regularity of f_i and η.

To investigate the stability of system (5.1.1), it is very important to consider the equilibrium points, namely the values $v \in \Omega$ such that $r(v,.) = 0$.

(H3) A natural assumption is the entropy dissipation condition, see [11, 69, 99], namely, for every $u, v \in \Omega$ with $r(v,.) = 0$,

$$\Big(\partial_u\eta(u,.) - \partial_u\eta(v,.)\Big) \cdot r(u,.) \le 0$$

where the notation $\cdot$ is the scalar product in the same space. Instead of hypothesis (H3), Tzavaras consider another assumption, see in [91], to investigate the asymptotic stability of stationary solution of system (5.1.1). This assumption can be represented as the form

(H3′) For every $u, v \in \Omega$ with $r(v,\cdot) = 0$, there exists $\gamma > 0$ such that

$$-\Big(\partial_u\eta(u,\cdot) - \partial_u\eta(v,\cdot)\Big) \cdot \Big(r(u,\cdot) - r(v,\cdot)\Big) \geq \gamma|u-v|^2.$$

This assumption may be more restrictive than the entropy dissipation condition (H3). However, it is the key to prove Theorem 5.3.1.

In the case non-conservative system (5.1.1), the product $s_i\partial_i\alpha$ are not defined for weak solutions, and generalised theories should be invoked, see [14, 26]. Instead of providing a particular definition of weak solutions, we mainly impose that these solutions satisfy the entropy inequality

$$\partial_t\eta(u,\alpha) + \sum_{i=1}^{d} \partial_i F_i(u,\alpha) \leq \Sigma(u,\alpha), \tag{5.1.7}$$

where $\Sigma(u,\alpha) = \partial_u\eta(u,\alpha)\cdot r(u,\alpha)$. Equation (5.1.7) becomes an equality for smooth solutions because of (H2). The inequality (5.1.7) is well defined for weak solutions due to the fact that the left hand side of this inequality is in a conservative form. In this work, the issue addressed is the role of the entropy inequality (5.1.7) for the stability analysis of non-conservative systems of the form (5.1.1), and more precisely, the nonlinear stability of stationary solutions of (5.1.1). They are very important in applications because they may serve not only as initial conditions, but also as solutions which can be reached in the long time limit. Besides, our analysis is independent of the space dimension, and of the hyperbolicity of system (5.1.1) (except that assumption (H1) implies hyperbolicity when α is constant).

The main tool we use to obtain these results is the relative entropy. Let us briefly recall this notion in the conservative case. Consider a $N \times N$ system of conservation laws with relaxation term

$$\partial_t u(t,x) + \sum_{i=1}^{d} \partial_i f_i(u) = r(u) \tag{5.1.8}$$

endowed with a Lax entropy pair (η, F), η being strictly convex (in a similar sense as in assumption $(H1)$), i.e. admissible weak solutions of (5.1.8) have to satisfy the inequality

$$\partial_t\eta(u) + \sum_{i=1}^{d} \partial_i F_i(u) \leq \Sigma(u)$$

in the weak sense with $\Sigma(u) = \partial_u\eta(u)\cdot r(u)$. The relative entropy associated with system (5.1.8) is defined by

$$h(u,v) = \eta(u) - \eta(v) - \partial_u\eta(v)\cdot(u-v).$$

Note that this function is not symmetric, and we should say that h is the entropy for u relatively to v. It is easy to check that

$$\frac{\beta_0}{2}|u-v|^2 \leq h(u,v) \leq \frac{\beta_1}{2}|u-v|^2 \tag{5.1.9}$$

where $|\cdot|$ is the Euclidian norm of $\mathbb{R}^N$ and $\sigma\left(\partial_u^2\eta\right) \subset [\beta_0, \beta_1]$.

Now, let us consider an admissible weak solution u of (5.1.8) and a constant vector $v \in \mathbb{R}^N$. After some calculations, one obtains

$$\partial_t h(u,v) + \sum_{i=1}^{d} \partial_i\Big(F_i(u) - \partial_u\eta(v)\cdot f_i(u)\Big) \leq \Sigma(u) - \partial_u\eta(v)\cdot r(u). \tag{5.1.10}$$

If we assume that the system (5.1.8) is entropy dissipative, and integrate the above inequality for $x \in \mathbb{R}^d$, the divergence term disappears and we have

$$\frac{d}{dt}\int_{\mathbb{R}^d} h(u,v)dx \leq 0.$$

According to (5.1.9), we then deduce the L^2-stability of constant. Thus, stationary solutions v in the class of admissible weak solutions.

In this work, we extend the previous analysis to systems of the form (5.1.1). For a given α, we are able to compare an entropy process solution to some particular stationary solutions. In Section 5.2, we detail the class of entropy process solution of (5.1.1) we consider in this work, which does not use any explicit definition of the non-conservative term. We then state and prove Theorem 5.2.7, on the nonlinear stability of particular stationary states of (5.1.1). In Section 5.3, instead of using the entropy dissipation condition (H3), we study the condition (H3′) to prove the asymptotic stability of stationary solution of nonlinear relaxation balance laws (5.1.1).

5.2 Stability of stationary solutions

5.2.1 Definition of entropy process solution

We aim at proving that stationary solutions are stable among entropy process solutions (which generalizes the concept of entropy weak solutions and can be obtained by passing to the limit of solution of the numerical scheme, see [31, 33]). The name "entropy process solution" was derived from the notation of bounded measurable process, that is a measurable mapping from a probability space into a space of bounded measurable function [10, 31]. Here, the probability space consists in the interval $(0,1)$, with the Borel σ-algebra and Lebesgue measure, and the set of bounded measurable functions is the bounded subset of $L^\infty([0,T)\times\mathbb{R}^d)$ defined by $\{\nu(.,.,\lambda), \lambda \in (0,1), \|\nu(.,.,\lambda)\|_\infty \leq C\}$, where $C > 0$ is independent of λ. However, since we consider discontinuous α, only in BV for instance, the product $s_i\partial_i\alpha$ in (5.1.1) are not defined. Several theories exist in the literature to define them, but here we only use some basic and natural assumptions. We assume that the products $s_i\partial_i\alpha$ can be described by means of vector-valued Radon measures $\mu_i \in \mathcal{M}\left(\mathbb{R}^+\times\mathbb{R}^d\times(0,1)\right)^N$[1] which satisfy at least the following properties:

(P1) On any open set $B = B_t \times B_x \subset \mathbb{R}^+\times\mathbb{R}^d$ such that $\alpha \in W^{1,\infty}(B_x)$, the measures $\mu_i, 1 \leq i \leq d$, satisfy

$$\forall\varphi \in \mathcal{C}_c^\infty(B), \forall 1 \leq i \leq d, \int_0^1\int_B \varphi d\mu_i(t,x,\lambda) = \int_0^1\int_B \varphi s_i(\nu,\alpha)\partial_i\alpha dtdxd\lambda.$$

(P2) For any component $1 \leq k \leq N$ and any dimension index $1 \leq i \leq d$,

$$s_i^{(k)} \equiv 0 \Rightarrow \mu_i^{(k)} \equiv 0.$$

Let us now define the entropy process solution

Definition 5.2.1. *Let $u_0 \in BV(\mathbb{R}^d,\Omega)^N$, $\alpha \in BV(\mathbb{R}^d)$ and $T > 0$. A function $\nu \in L^\infty([0,T)\times\mathbb{R}^d\times(0,1),\Omega)$ is an entropy process solution of the Cauchy problem* (5.1.1)-(5.1.2) *if there*

[1] More precisely, $\mathcal{M}(X)$ denotes the set of locally bounded Radon measures on a set X, i.e. $\mathcal{M}(X) = (\mathcal{C}_c(X))'$.

exists $(\mu_i)_{1\leq i\leq d} \subset \mathcal{M}(\mathbb{R}^+ \times \mathbb{R}^d \times (0,1))$ satisfying assumptions (P1) and (P2) such that, for all $\varphi \in \mathcal{C}_c^\infty([0,T) \times \mathbb{R}^d)$,

$$\begin{aligned} -\int_0^1 \int_0^T \int_{\mathbb{R}^d} \left(\nu(t,x,\lambda)\partial_t\varphi + \sum_{i=1}^d f_i(\nu,\alpha)\partial_i\varphi \right) dxdtd\lambda + \int_0^1 \int_0^T \int_{\mathbb{R}^d} \varphi d\mu(t,x,\lambda) \\ - \int_{\mathbb{R}^d} u_0(x)\varphi(0,x)dx = \int_0^1 \int_0^T \int_{\mathbb{R}^d} r(\nu,\alpha)\varphi dxdtd\lambda, \end{aligned} \tag{5.2.1}$$

and, for all non-negative $\varphi \in \mathcal{C}_c^\infty([0,T) \times \mathbb{R}^d)$,

$$\begin{aligned} -\int_0^1 \int_0^T \int_{\mathbb{R}^d} \left(\eta(\nu,\alpha)\partial_t\varphi + \sum_{i=1}^d F_i(\nu,\alpha)\partial_i\varphi \right) dxdtd\lambda - \int_{\mathbb{R}^d} \eta(u_0,\alpha)(x)\varphi(0,x)dx \\ \leq \int_0^1 \int_0^T \int_{\mathbb{R}^d} \Sigma(\nu,\alpha)\varphi dxdtd\lambda. \end{aligned} \tag{5.2.2}$$

Remark 5.2.2. *From an entropy weak solution $u(t,x)$ to problem (5.1.1)-(5.1.2), one may easily construct an entropy process solution to problem (5.1.1)-(5.1.2) by setting $\nu(t,x,\lambda) = u(t,x)$ for any $\lambda \in (0,1)$. Reciprocally, if ν is an entropy process solution to problem (5.1.1)-(5.1.2) such that there exists $u \in L^\infty([0,T)\times\mathbb{R}^d)$ such that $\nu(t,x,\lambda) = u(t,x)$ for a.e. $(t,x,\lambda) \in [0,T) \times \mathbb{R}^d \times (0,1)$, then u is an entropy weak solution to problem (5.1.1)-(5.1.2).*

Remark 5.2.3. *The definition 5.2.1 is not sufficient to hope a well-posedness result, without any additional assumption on the measures μ_i, but it is sufficient to obtain the stability results of the next sections. Besides, assumption $(P1)$ is not necessary for the upcoming analysis. We introduce it to ensure that, if $\alpha \in W^{1,\infty}(\mathbb{R}^d)$ the standard definition of entropy process solutions is recovered. It is also important to note that inequalities (5.2.2) exactly correspond to the weak form of (5.1.7) by setting $\nu(t,x,\lambda) = u(t,x)$ for any $\lambda \in (0,1)$, so that the measures μ_i do not appear there.*

5.2.2 Relative entropy and nonlinear stability

5.2.2.1 Relative entropy

In [60], Kruzhkov is able to compare two entropy weak solutions using the doubling variable technique. In [60], such method has been extended in order to compare an entropy weak solution with an approximate solution. In the case of systems of conservation laws, these techniques no longer work. Basically, the family of entropy pairs (η, F) is not sufficiently rich to control the difference between two solutions. As mention in the introduction, it seems impossible to construct a relative entropy for system (5.1.1) to compare two solutions (ν,α) and (v,β). Nonetheless, one can define a relative entropy between two solutions ν and v, α being given and common.

Definition 5.2.4. *Let $\nu, v \in \Omega$. The relative entropy of ν with respect to v is defined by*

$$\begin{aligned} h :\ & \Omega \times \Omega \times \mathbb{R} \to \mathbb{R}^+ \\ & (\nu, v, \alpha) \mapsto \eta(\nu,\alpha) - \eta(v,\alpha) - \partial_u\eta(v,\alpha)\cdot(\nu - v) \end{aligned} \tag{5.2.3}$$

and the corresponding relative entropy fluxes $q : \Omega \times \Omega \times \mathbb{R} \to \mathbb{R}^d$ are

$$q_i(\nu,v,\alpha) = F_i(\nu,\alpha) - F_i(v,\alpha) - \partial_u\eta(v,\alpha)\cdot\Big(f_i(\nu,\alpha) - f_i(v,\alpha)\Big), \quad \forall i \in \{1,...,d\}. \tag{5.2.4}$$

On the other hand, it follows from the definition of h that

$$h(\nu, v, .) = \int_0^1 \int_0^\theta (\nu - v)^T \partial_u^2 \eta(v + \gamma(\nu - v), .)(\nu - v) d\gamma d\theta, \tag{5.2.5}$$

which leads to the following results

Lemma 5.2.5. *Assume that the entropy η satisfies (H1). Then, the relative entropy is convex with respect to its first variable and for all $\nu, v \in \Omega$, we have*

$$\frac{\beta_0}{2}|\nu - v|^2 \leq h(\nu, v, .) \leq \frac{\beta_1}{2}|\nu - v|^2. \tag{5.2.6}$$

Lemma 5.2.6. *Let L_f be defined by (5.1.5). Assume that the entropy pair (η, F) satisfies (H1) and (H2). Then, for all $(\nu, v) \in \Omega^2$, we have*

$$|q_i(\nu, v, .)| \leq L_f h(\nu, v, .). \tag{5.2.7}$$

Proof. Denote by $w = \nu - v$, the definition of the relative entropy h in (5.2.5) becomes

$$h(\nu, v, .) = \int_0^1 \int_0^\theta w^T \partial_u^2 \eta(v + \gamma w, .) w d\gamma d\theta. \tag{5.2.8}$$

Denoting by $\mathbb{A}_\gamma$ the symmetric definite positive matrix $\partial_u^2 \eta(v + \gamma w, .)$, and by $\langle ., . \rangle_{\mathbb{A}_\gamma}$, the scalar product on $\mathbb{R}^N$ defined by $\langle v_1, v_2 \rangle_{\mathbb{A}_\gamma} = v_1^T \mathbb{A}_\gamma v_2$, the relation (5.2.8) can be rewritten

$$h(\nu, v, .) = \int_0^1 \int_0^\theta \langle w, w \rangle_{\mathbb{A}_\gamma} d\gamma d\theta.$$

On the other hand, it follows from the assumption (H2) of the entropy flux F that

$$\begin{aligned} q_i(\nu, v, .) &= \int_0^1 \left(\partial_u \eta(v + \theta w, .) - \partial_u \eta(v, .)\right) \left(\partial_u f_i(v + \theta w, .)\right)^T d\theta \\ &= \int_0^1 \int_0^\theta \langle w, \left(\partial_u f_i(v + \theta w, .)\right)^T w \rangle_{\mathbb{A}_\gamma} d\gamma d\theta. \end{aligned}$$

for all $1 \leq i \leq d$. The quantity L_f introduced in (5.1.5) has been designed so that

$$\left| \langle w, \left(\partial_u f_i(v + \theta w, .)\right)^T w \rangle_{\mathbb{A}_\gamma} \right| \leq \langle w, w \rangle_{\mathbb{A}_\gamma}.$$

Therefore, we obtain

$$|q_i(\nu, v, .)| \leq L_f \int_0^1 \int_0^\theta \langle w, w \rangle_{\mathbb{A}_\gamma} d\gamma d\theta = L_f h(\nu, v, .).$$

This ends the proof of Lemma 5.2.6. □

5.2.2.2 Nonlinear stability

For a given $\alpha \in \mathcal{C}^1(\mathbb{R}^d)$, consider a smooth, and thus entropy conservative, entropy process solution ν of (5.1.1), and a time-independent function v. Let us compute the equation satisfied by the relative entropy h

$$\begin{aligned}\partial_t h(\nu, v, \alpha) &= \partial_t \eta(\nu, \alpha) - \partial_u \eta(v, \alpha) \cdot \partial_t \nu \\ &= \Sigma(\nu, \alpha) - \sum_{i=1}^d \partial_i F_i(\nu, \alpha) - \partial_u \eta(v, \alpha) \cdot \left[r(\nu, \alpha) - \sum_{i=1}^d \partial_i f_i(\nu, \alpha) - \sum_{i=1}^d s_i(\nu, \alpha) \partial_i \alpha \right] \\ &= - \sum_{i=1}^d \partial_i \left[F_i(\nu, \alpha) - \partial_u \eta(v, \alpha) \cdot f_i(\nu, \alpha) \right] + \Sigma(\nu, \alpha) - \partial_u \eta(v, \alpha) \cdot r(\nu, \alpha) \\ &\quad - \sum_{i=1}^d \partial_i \left[\partial_u \eta(v, \alpha) \right] \cdot f_i(\nu, \alpha) + \partial_u \eta(v, \alpha) \cdot \sum_{i=1}^d s_i(\nu, \alpha) \partial_i \alpha.\end{aligned} \tag{5.2.9}$$

The two last terms are not in conservation form, but one could make them vanishing adding some assumptions on v. Following [80], for any given constant vector $H_0 \in \mathbb{R}^N$, we introduce $\mathcal{S}(H_0)$, the set of $(v, \alpha) \in \Omega \times \mathbb{R}$ such that

(S1) $\partial_u \eta(v, \alpha) = H_0$.

(S2) For all $1 \leq i \leq d$ and $1 \leq k \leq N$, $H_0^{(k)} s_i^{(k)} \equiv 0$.

Let us emphasize that the above discussion on the smooth solution is extended to entropy process solutions by the following theorem

Theorem 5.2.7. *Let $H_0 \in \mathbb{R}^N$ and consider the set $\mathcal{S}(H_0)$ defined by (S1)-(S2), assumed to be nonempty. Consider $\alpha \in BV(\mathbb{R}^d)$ and a function $v \in BV(\mathbb{R}^d, \Omega)$ such that $(v, \alpha) \in \mathcal{S}(H_0)$ almost everywhere and satisfy the entropy dissipation condition (H3). Then, v is a stationary entropy process solution of system (5.1.1).*

Moreover, let $T > 0$, $u_0 \in BV(\mathbb{R}^d, \Omega)^N$, and $\nu \in L^\infty\big([0, T) \times \mathbb{R}^d \times (0, 1), \Omega)\big)$ an associated entropy process solution. Then, there exists a positive constant L_f, independent of ν, v and α such that the following nonlinear stability property holds for all $R > 0$ and for almost every $t \in [0, T]$

$$\int_0^1 \int_{B(0,R)} h(\nu(t, x, \lambda), v(x), \alpha(x)) dx d\lambda \leq \int_{B(0,R+L_f t)} h(u_0(x), v(x), \alpha(x)) dx. \tag{5.2.10}$$

Proof. First, let us remark that the stability inequality (5.2.10) implies that v is a stationary solution of system (5.1.1). Indeed, if we choose $\nu(0, x, \lambda) = u_0(x) = v(x)$, for any $\lambda \in (0, 1)$, the right-hand side of (5.2.10) is null, by the properties of h, see Lemma 5.2.5. Therefore, ν being an entropy process solution and v being time-independent, one may deduce that v is stationary entropy process solution using once again the properties of h.

Let us now rewrite the calculations described above, but in the weak sense. By assumptions $(S1)$ and $(S2)$, $H_0 \cdot s_i(., \alpha) = 0$ for all i. In other words, this means that if the i-th component of H_0 is non-zero, then $s_i \equiv 0$. We now use the definition of ν and assumption (P2) on the non-conservative product to obtain, for all $\varphi \in \mathcal{C}_c^\infty([0, T) \times \mathbb{R}^d)$,

$$\begin{aligned}- \int_0^1 \int_0^T \int_{\mathbb{R}^d} H_0 \cdot \left[\nu \partial_t \varphi + \sum_{i=1}^d f_i(\nu, \alpha) \partial_i \varphi \right] dx dt d\lambda - \int_{\mathbb{R}^d} H_0 \cdot u_0(x) \varphi(0, x) dx \\ = \int_0^1 \int_0^T \int_{\mathbb{R}^d} H_0 \cdot r(\nu, \alpha) \varphi dx dt d\lambda.\end{aligned}$$

Now, using the entropy inequality (5.2.2) for ν and the fact that v is independent of time,

$$
\begin{aligned}
&-\int_0^1\int_0^T\int_{\mathbb{R}^d}\left[h(\nu,v,\alpha)\partial_t\varphi+\sum_{i=1}^d q_i(\nu,v,\alpha)\partial_i\varphi\right]dxdtd\lambda-\int_{\mathbb{R}^d}h(u_0,v,\alpha)\varphi(0,x)dx\\
&\leq\int_0^1\int_0^T\int_{\mathbb{R}^d}\left[\partial_u\eta(\nu,\alpha)-H_0\right]\cdot r(\nu,\alpha)\varphi dxdtd\lambda-\int_0^1\int_0^T\int_{\mathbb{R}^d}\sum_{i=1}^d\partial_i\left[F_i(v,\alpha)-H_0\cdot f_i(v,\alpha)\right]\varphi dxdtd\lambda,
\end{aligned}
\tag{5.2.11}
$$

for all non-negative $\varphi\in\mathcal{C}_c^\infty([0,T)\times\mathbb{R}^d)$. Under the entropy dissipation condition (H3), the first term of the right-hand side in (5.2.11) is non-positive. Indeed, since the equilibrium solution of $r(v,.)=0$, one could make the last term of the right-hand side in (5.2.11) vanishing. Therefore, the inequality (5.2.11) becomes

$$
\int_0^1\int_0^T\int_{\mathbb{R}^d}\left[h(\nu,v,\alpha)\partial_t\varphi+\sum_{i=1}^d q_i(\nu,v,\alpha)\partial_i\varphi\right]dxdtd\lambda+\int_{\mathbb{R}^d}h(u_0,v,\alpha)\varphi(0,x)dx\geq 0. \tag{5.2.12}
$$

To obtain inequality (5.2.10), we introduce L_f in (5.1.5) such that

$$
|q_i|\leq L_f h \tag{5.2.13}
$$

which is comparable to the maximum of the spectral radii of $\partial_u f_i$ (see more details in Lemma 5.2.6). It suffices now to introduce, t and R being fixed

$$
w_\varepsilon(\tau)=\begin{cases}1, & \text{if } 0\leq\tau\leq t,\\ 1+(t-\tau)/\varepsilon, & \text{if } t<\tau\leq t+\varepsilon,\\ 0, & \text{if } t+\varepsilon<\tau,\end{cases} \tag{5.2.14}
$$

and

$$
\chi_\varepsilon(\tau,x)=\begin{cases}1, & \text{if } |x|\leq R+L_f(t-\tau),\\ 1+(R+L_f(t-\tau)-|x|)/\varepsilon, & \text{if } 0<|x|-R-L_f(t-\tau)\leq\varepsilon,\\ 0, & \text{if } R+L_f(t-\tau)+\varepsilon<|x|.\end{cases} \tag{5.2.15}
$$

and take $\varphi(\tau,x)=\chi_\varepsilon(\tau,x)w_\varepsilon(\tau)$ (we omit the passage from Lipschitz continuous functions to $\mathcal{C}_c^\infty$ functions. Plugging this test function into (5.2.12) yields

$$
\begin{aligned}
\frac{1}{\varepsilon}\int_0^1\int_t^{t+\varepsilon}\int_{B(0,R+\varepsilon)}h(\nu,v,\alpha)(\tau,x)\chi_\varepsilon(\tau,x)dxd\tau d\lambda\leq\int_{B(0,R+L_ft+\varepsilon)}h(u_0,v,\alpha)\chi_\varepsilon(0,x)dx\\
-\frac{1}{\varepsilon}\int_0^1\int_0^{t+\varepsilon}\int_{B(0,R+L_f(t-\tau)+\varepsilon)}w_\varepsilon(\tau)\left[L_fh(\nu,v,\alpha)+\frac{x}{|x|}q(\nu,v,\alpha)\right]dxd\tau d\lambda.
\end{aligned}
$$

By definition of L_f, the last integral is non-negative, so that, letting ε tend to 0 provides inequality (5.2.10). □

We provide here one example of equations which enter in this framework. In one dimension, the standard shallow water model with Darcy-Weisbach friction reads

$$
\begin{cases}\partial_t h+\partial_x(hU)=0,\\ \partial_t(hU)+\partial_x\left(hU^2+\dfrac{1}{2}gh^2\right)=-gh\partial_x\alpha-\kappa(h,hU)U|U|,\\ \partial_t\alpha=0,\end{cases} \tag{5.2.16}
$$

where h is the height of water, assumed to remain positive, u is the depth-averaged velocity, α plays the role of the bathymetry, g is the gravity constant and $\kappa(U, hU) \geq 0$ is a Darcy-Weisbach friction.

This system of equations may be endowed with an entropy inequality of the form (5.1.7), setting

$$\eta(u, \alpha) = hU^2/2 + gh(h/2 + \alpha) \quad \text{and} \quad F(u, \alpha) = U\left(\eta(u, \alpha) + gh^2/2\right)$$

where $u = (h, hU)^T$. The convexity of η with respect to u is classical and one can see that η is only linear in α.

The description of all possible stationary solutions is very difficult in practice. The simplest ones correspond to a "lake at rest" and are defined by

$$h + \alpha = Z_0 \quad \text{and} \quad U = 0 \quad \text{a.e} \tag{5.2.17}$$

where Z_0 is a given real constant greater than the maximum of α. On the other hand, we can compute

$$\partial_u \eta(u, \alpha) = \begin{pmatrix} -U^2/2 + g(h+\alpha) \\ U \end{pmatrix}$$

As a consequence, assumption (S2) yields $U = 0$, since $s_1 = s_2 = (0, gh)^T$. Next, assumption (S1) corresponds to equality $h + \alpha = Z_0$. Indeed, assumption (H3) becomes $-\kappa(h, hU)U^2|U| < 0$. To sum up, we have

Corollary 5.2.8. *Stationary solution of the shallow water equation* (5.2.16) *given by* (5.2.17) *(lake at rest) are nonlinear stable, in the sense of theorem 5.2.7.*

5.3 Asymptotic stability of stationary solution

We now use the hypothesis (H3′) to prove the following theorem

Theorem 5.3.1. *Let $H_0 \in \mathbb{R}^N$ and consider the set $\mathcal{S}(H_0)$ defined by (S1)-(S2), assumed to be nonempty. Consider $\alpha \in BV(\mathbb{R}^d)$ and a function $v \in BV(\mathbb{R}^d, \Omega)$ such that $(v, \alpha) \in \mathcal{S}(H_0)$ almost everywhere and satisfy the entropy dissipation condition (H3′). Then, v is a stationary entropy process solution of system* (5.1.1).

Moreover, let $T > 0$, $u_0 \in BV(\mathbb{R}^d, \Omega)^N$, and $\nu \in L^\infty\left([0, T) \times \mathbb{R}^d \times (0,1), \Omega)\right)$ an associated entropy process solution. Then, there exist positive constants L_f and γ, independent of ν, v and α such that the following nonlinear stability property holds for all $R > 0$ and for almost every $t \in [0, T]$

$$\begin{aligned} \int_0^1 \int_{B(0,R)} h(\nu(t,x,\lambda), v(x), \alpha(x)) dx d\lambda + \gamma \int_0^1 \int_0^t \int_{B(0,R+L_f(t-\tau))} |\nu(\tau,x,\lambda) - v(x)|^2 dx d\tau d\lambda \\ \leq \int_{B(0,R+L_f t)} h(u_0(x), v(x), \alpha(x)) dx. \end{aligned} \tag{5.3.1}$$

Proof. We first can see that by the properties of h, see Lemma 5.2.5, if we choose $\nu(0, x, \lambda) = u_0(x) = v(x)$, for any $\lambda \in (0, 1)$, the right hand side of (5.3.1) is null. Besides, v is independent of time and ν is an entropy process solution. Therefore, v is stationary entropy process solution by using the properties of h in Lemma 5.2.5.

Now, we give more details of the calculations described above in the weak sense. By assumptions (S1)-(S2), (P2) and the definition of ν, for all $\varphi \in \mathcal{C}_c^\infty([0,T)\times\mathbb{R}^d)$, we get the following estimate

$$-\int_0^1\int_0^T\int_{\mathbb{R}^d} H_0\cdot\left[\nu\partial_t\varphi+\sum_{i=1}^d f_i(\nu,\alpha)\partial_i\varphi\right]dxdtd\lambda-\int_{\mathbb{R}^d} H_0\cdot u_0(x)\varphi(0,x)dx \\ =\int_0^1\int_0^T\int_{\mathbb{R}^d} H_0\cdot r(\nu,\alpha)\varphi dxdtd\lambda.$$

Now, using the entropy inequality (5.2.2) for ν and the fact that v is independent of time, we have

$$-\int_0^1\int_0^T\int_{\mathbb{R}^d}\left[h(\nu,v,\alpha)\partial_t\varphi+\sum_{i=1}^d q_i(\nu,v,\alpha)\partial_i\varphi\right]dxdtd\lambda-\int_{\mathbb{R}^d} h(u_0,v,\alpha)\varphi(0,x)dx \\ \leq\int_0^1\int_0^T\int_{\mathbb{R}^d}\left[\partial_u\eta(\nu,\alpha)-H_0\right]\cdot r(\nu,\alpha)\varphi dxdtd\lambda-\int_0^1\int_0^T\int_{\mathbb{R}^d}\sum_{i=1}^d\partial_i\left[F_i(v,\alpha)-H_0\cdot f_i(v,\alpha)\right]\varphi dxdtd\lambda, \tag{5.3.2}$$

for all non-negative $\varphi\in\mathcal{C}_c^\infty([0,T)\times\mathbb{R}^d)$. Since the equilibrium solution of $r(v,\cdot)=0$, one could make the last term of the right hand side in (5.3.2) vanishing. Indeed, under the assumption (H3′), the inequality (5.3.2) becomes

$$\int_0^1\int_0^T\int_{\mathbb{R}^d}\left[h(\nu,v,\alpha)\partial_t\varphi+\sum_{i=1}^d q_i(\nu,v,\alpha)\partial_i\varphi\right]dxdtd\lambda+\int_{\mathbb{R}^d} h(u_0,v,\alpha)\varphi(0,x)dx \\ \geq\gamma\int_0^1\int_0^T\int_{\mathbb{R}^d}|\nu-v|^2dxdtd\lambda. \tag{5.3.3}$$

Let $w_\varepsilon(\tau)$ and $\chi_\varepsilon(\tau,x)$ be the same as in (5.2.14) and (5.2.15), respectively. We now select the test function $\varphi(\tau,x)=\chi_\varepsilon(\tau,x)w_\varepsilon(\tau)$ and introduce it to (5.3.3). This gives

$$\frac{1}{\varepsilon}\int_0^1\int_t^{t+\varepsilon}\int_{B(0,R+\varepsilon)} h(\nu,v,\alpha)(\tau,x)\chi_\varepsilon(\tau,x)dxd\tau d\lambda+\gamma\int_0^1\int_0^t\int_{B(0,R+L_f(t-\tau)+\varepsilon)}|\nu-v|^2dxdtd\lambda \\ \leq\int_{B(0,R+L_ft+\varepsilon)} h(u_0,v,\alpha)\chi_\varepsilon(0,x)dx \\ -\frac{1}{\varepsilon}\int_0^1\int_0^{t+\varepsilon}\int_{B(0,R+L_f(t-\tau)+\varepsilon)} w_\varepsilon(\tau)\left[L_fh(\nu,v,\alpha)+\frac{x}{|x|}q(\nu,v,\alpha)\right]dxd\tau d\lambda.$$

By definition of L_f, the last integral is non-negative. Besides, letting ε tend to 0, the above inequality can be rewritten as

$$\int_0^1\int_{B(0,R)} h(\nu(t,x,\lambda),v(x),\alpha(x))dxd\lambda+\gamma\int_0^1\int_0^t\int_{B(0,R+L_f(t-\tau))}|\nu(\tau,x,\lambda)-v(x)|^2dxd\tau d\lambda \\ \leq\int_{B(0,R+L_ft)} h(u_0(x),v(x),\alpha(x))dx.$$

This ends the proof of Theorem 5.3.1. □

Example of relaxation systems that verify (H3′) are given in [53, 61, 90]. The discrete velocity BGK-models in [100, Theorem 5.2] provide yet another example verifying (H3′).

Remark 5.3.2. *In [79], Ruggeri and Serre studied the asymptotic stability of constant states which can be extended to non constant stationary solutions of systems of balance laws, as the lake at rest states for the shallow-water equations with bathymetry and friction. A key assumption to prove the asymptotic stability is the Kawashima's condition. The technique is based on the construction of an appropriate Lyapunov functional involving the entropy and a so-called compensation term. Following the above idea, we tried to sum up the obtention of the Lyapunov functional and of the associated non-conservation systems* (5.1.1)*, but we failed to extend the result of Ruggeri and Serre to our case. We postpone this possibility to a further work.*

Bibliography

[1] B. Andreianov, K. H. Karlsen, and N. H. Risebro. A theory of L^1-dissipative solvers for scalar conservation laws with discontinuous flux. *Arch. Ration. Mech. Anal.*, 201(1):27–86, 2011.

[2] A. Arnold, M. Ehrhardt, and I. Sofronov. Discrete transparent boundary conditions for the Schrödinger equation: fast calculation, approximation, and stability. *Commun. Math. Sci.*, 1(3):501–556, 2003.

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

[4] C. Besse, M. Ehrhardt, and I. Lacroix-Violet. Discrete artificial boundary conditions for the linearized Korteweg–de Vries equation. *Numer. Methods Partial Differential Equations*, 32(5):1455–1484, 2016.

[5] C. Besse, P. Noble, and D. Sanchez. Discrete transparent boundary conditions for the mixed KDV-BBM equation. *J. Comput. Phys.*, 345:484–509, 2017.

[6] B. Boutin and J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems: numerical boundary layers. *Numer. Math. Theory Methods Appl.*, 10(3):489–519, 2017.

[7] B. Boutin, T. H. T. Nguyen, and N. Seguin. A stiffly stable fully discrete scheme for the damped wave equation using discrete transparent boundary condition. Preprint, May 2020.

[8] B. Boutin, T. H. T. Nguyen, and N. Seguin. A stiffly stable semi-discrete scheme for the characteristic linear hyperbolic relaxation with boundary. *ESAIM: Mathematical Modelling and Numerical Analysis*, 2020.

[9] B. Boutin, T. H. T. Nguyen, A. Sylla, S. Tran-Tien, and J.-F. Coulombel. High order numerical schemes for transport equations on bounded domains. Preprint, Dec. 2019.

[10] Y. Brenier, C. De Lellis, and L. Székelyhidi, Jr. Weak-strong uniqueness for measure-valued solutions. *Comm. Math. Phys.*, 305(2):351–361, 2011.

[11] G.-Q. Chen, C. D. Levermore, and T.-P. Liu. Hyperbolic conservation laws with stiff relaxation terms and entropy. *Commun. Pure Appl. Math.*, 47(6):787–830, 1994.

[12] J. F. Clarke. Gas dynamics with relaxation effects. *Reports on Progress in Physics*, 41(6):807–864, jun 1978.

[13] P. Colella, A. Majda, and V. Roytburd. Theoretical and numerical structure for reacting shock waves. *SIAM J. Sci. Stat. Comput.*, 7:1059–1080, 1986.

[14] J.-F. Colombeau. *Multiplication of distributions*, volume 1532 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1992. A tool in mathematics, numerical engineering and theoretical physics.

[15] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. Lecture, Aug. 2011.

[16] J.-F. Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems II. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze*, X(1):37–98, Dec. 2011.

[17] 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, American Institute of Mathematical Sciences*, pages 97–225, 2013.

[18] J.-F. Coulombel. Résolvante, stabilité et applications. *Matapli*, 103:91–122, 2014.

[19] J.-F. Coulombel. Fully discrete hyperbolic initial boundary value problems with nonzero initial data. *Confluentes Mathematici*, 7(2):17–47, May 2015.

[20] J.-F. Coulombel. Transparent numerical boundary conditions for evolution equations: Derivation and stability analysis. *Annales de la Faculté des Sciences de Toulouse. Mathématiques.*, 28(2):259–327, 2019.

[21] 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.

[22] J.-F. Coulombel and F. Lagoutière. The neumann numerical boundary condition for transport equations. *Kinet. Relat. Models*, to appear, 2020.

[23] C. Courtès, F. Lagoutière, and F. Rousset. Error estimates of finite difference schemes for the Korteweg-de Vries equation. *IMA Journal of Numerical Analysis*, 2019.

[24] C. M. Dafermos. The second law of thermodynamics and stability. *Arch. Rational Mech. Anal.*, 70(2):167–179, 1979.

[25] G. Dakin, B. Després, and S. Jaouen. Inverse Lax-Wendroff boundary treatment for compressible Lagrange-remap hydrodynamics on Cartesian grids. *J. Comput. Phys.*, 353:228–257, 2018.

[26] G. Dal Maso, P. G. Lefloch, and F. Murat. Definition and weak stability of nonconservative products. *J. Math. Pures Appl. (9)*, 74(6):483–548, 1995.

[27] R. Dautray and J.-L. Lions. *Analyse mathématique et calcul numérique pour les sciences et les techniques. Tome 3*. Collection du Commissariat à l'Énergie Atomique: Série Scientifique. Masson, Paris, 1985.

[28] C. de Boor. Divided differences. *Surv. Approx. Theory*, 1:46–69, 2005.

[29] B. Després. Finite volume transport schemes. *Numer. Math.*, 108(4):529–556, 2008.

[30] R. J. DiPerna. Uniqueness of solutions to hyperbolic conservation laws. *Indiana Univ. Math. J.*, 28(1):137–188, 1979.

[31] R. Eymard, T. Gallouët, and R. Herbin. Existence and uniqueness of the entropy solution to a nonlinear hyperbolic equation. *Chinese Ann. Math. Ser. B*, 16(1):1–14, 1995.

[32] F. Filbet and C. Yang. An inverse Lax-Wendroff method for boundary conditions applied to Boltzmann type models. *J. Comput. Phys.*, 245:43–61, 2013.

[33] U. S. Fjordholm, R. Käppeli, S. Mishra, and E. Tadmor. Construction of approximate entropy measure-valued solutions for hyperbolic systems of conservation laws. *Found. Comput. Math.*, 17(3):763–827, 2017.

[34] K. O. Friedrichs. Symmetric hyperbolic linear differential equations. *Comm. Pure Appl. Math.*, 7:345–392, 1954.

[35] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1996.

[36] M. Goldberg. On a boundary extrapolation theorem by Kreiss. *Math. Comp.*, 31(138):469–477, 1977.

[37] M. Goldberg and E. Tadmor. Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. I. *Math. Comp.*, 32(144):1097–1107, 1978.

[38] M. Goldberg and E. Tadmor. Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comp.*, 36(154):603–626, 1981.

[39] M. J. Grote and J. B. Keller. Exact nonreflecting boundary conditions for the time dependent wave equation. volume 55, pages 280–297. 1995. Perturbation methods in physical mathematics (Troy, NY, 1993).

[40] R. B. Guenther and J. W. Lee. *Partial differential equations of mathematical physics and integral equations.* Dover Publications, Inc., Mineola, NY, 1996. Corrected reprint of the 1988 original.

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

[42] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods. 2nd ed.* Hoboken, NJ: John Wiley & Sons, 2nd ed. edition, 2013.

[43] B. Gustafsson, H.-O. Kreiss, and A. Sundström. Stability theory of difference approximations for mixed initial boundary value problems. ii. *Mathematics of Computation*, 26, 09 1972.

[44] B. Hanouzet and R. Natalini. Global existence of smooth solutions for partially dissipative hyperbolic systems with a convex entropy. *Arch. Ration. Mech. Anal.*, 169(2):89–117, 2003.

[45] G. W. Hedstrom. Norms of powers of absolutely convergent fourier series. *Michigan Math. J.*, 13(4):393–416, 12 1966.

[46] R. L. Higdon. Initial-boundary value problems for linear hyperbolic systems. *SIAM Rev.*, 28(2):177–217, 1986.

[47] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[48] M. Inglard, F. Lagoutière, and H. H. Rugh. Ghost solutions with centered schemes for one-dimensional transport equations with Neumann boundary conditions. working paper or preprint, Oct. 2018.

[49] E. Isaacson and B. Temple. Nonlinear resonance in systems of conservation laws. *SIAM J. Appl. Math.*, 52(5):1260–1278, 1992.

[50] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.*, 21(2):441–454, 1999.

[51] S. Jin and Z. Xin. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Commun. Pure Appl. Math.*, 48(3):235–276, 1995.

[52] E. Jury. Theory and application of the Z-transform method. *Wiley and Sons, New York*, 1964.

[53] M. A. Katsoulakis and A. E. Tzavaras. Contractive relaxation systems and the scalar multidimensional conservation law. *Comm. Partial Differential Equations*, 22(1-2):195–233, 1997.

[54] M. Kazakova and P. Noble. Discrete transparent boundary conditions for the linearized green-naghdi system of equations. *SIAM Journal on Numerical Analysis*, 58, 10 2017.

[55] C. Klingenberg and N. H. Risebro. Convex conservation laws with discontinuous coefficients. Existence, uniqueness and asymptotic behavior. *Comm. Partial Differential Equations*, 20(11-12):1959–1990, 1995.

[56] H.-O. Kreiss. Difference approximations for hyperbolic differential equations. In *Numerical Solution of Partial Differential Equations (Proc. Sympos. Univ. Maryland, 1965)*, pages 51–58. Academic Press, 1966.

[57] H.-O. Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. i. *Mathematics of Computation - Math. Comput.*, 22, 10 1968.

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

[59] H.-O. Kreiss and E. Lundqvist. On difference approximations with wrong boundary values. *Math. Comp.*, 22:1–12, 1968.

[60] S. N. Kružkov. First order quasilinear equations with several independent variables. *Mat. Sb. (N.S.)*, 81 (123):228–255, 1970.

[61] C. Lattanzio and A. E. Tzavaras. Structural properties of stress relaxation and convergence from viscoelasticity to polyconvex elastodynamics. *Arch. Ration. Mech. Anal.*, 180(3):449–492, 2006.

[62] P. D. Lax. Hyperbolic systems of conservation laws. II. *Comm. Pure Appl. Math.*, 10:537–566, 1957.

[63] P. D. Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves.* Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1973. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 11.

[64] T.-P. Liu. Hyperbolic conservation laws with relaxation. *Commun. Math. Phys.*, 108:153–175, 1987.

[65] A. Majda and S. Osher. Initial-boundary value problems for hyperbolic equations with uniformly characteristic boundary. *Comm. Pure Appl. Math.*, 28(5):607–675, 1975.

[66] K. Mattsson. Boundary procedures for summation-by-parts operators. *J. Sci. Comput.*, 18(1):133–153, 2003.

[67] K. Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *J. Sci. Comput.*, 51(3):650–682, 2012.

[68] K. Mattsson and J. Nordström. Summation by parts operators for finite difference approximations of second derivatives. *J. Comput. Phys.*, 199(2):503–540, 2004.

[69] I. Müller and T. Ruggeri. *Rational extended thermodynamics*, volume 37 of *Springer Tracts in Natural Philosophy.* Springer-Verlag, New York, second edition, 1998. With supplementary chapters by H. Struchtrup and Wolf Weiss.

[70] T. Myint-U and L. Debnath. *Linear partial differential equations for scientists and engineers. 4th ed.* Basel: Birkhäuser, 4th ed. edition, 2007.

[71] H. O. Kreiss and G. Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. *Mathematical Aspects of Finite Elements in Partial Differential Equations*, 12 1974.

[72] H. O. Kreiss and G. Scherer. On the existence of energy estimates for difference approximations for hyperbolic systems. *Technical report, Uppsala University, Dept of Scientific Computing, Uppsala, Sweden.*, 01 1977.

[73] P. Olsson. Summation by parts, projections, and stability. I. *Math. Comp.*, 64(211):1035–1065, S23–S26, 1995.

[74] P. Olsson. Summation by parts, projections, and stability. II. *Math. Comp.*, 64(212):1473–1493, 1995.

[75] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signals and Systems (2Nd Ed.).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[76] S. Osher. Systems of difference equations with general homogeneous boundary conditions. *Transactions of The American Mathematical Society - TRANS AMER MATH SOC*, 137:177–177, 03 1969.

[77] T. Popoviciu. Introduction à la théorie des différences divisées. *Bull. Math. Soc. Roumaine Sci.*, 42(1):65–78, 1940.

[78] J. V. Ralston. Note on a paper of Kreiss. *Comm. Pure Appl. Math.*, 24(6):759–762, 1971.

[79] T. Ruggeri and D. Serre. Stability of constant equilibrium state for dissipative balance laws system with a convex entropy. *Quart. Appl. Math.*, 62(1):163–179, 2004.

[80] N. Seguin. Stability of stationary solutions of singular systems of balance laws. *Confluentes Math.*, 10(2):93–112, 2018.

[81] J. F. Steffensen. Note on divided differences. *Danske Vid. Selsk. Mat.-Fys. Medd.*, 17(3):12, 1939.

[82] J. J. Stoker. *Water waves. The mathematical theory with applications. Reprint of the 1957 original.* New York, NY: Wiley, reprint of the 1957 original edition, 1992.

[83] B. Strand. Summation by parts for finite difference approximations for d/dx. *J. Comput. Phys.*, 110(1):47–67, 1994.

[84] G. Strang. Trigonometric polynomials and difference methods of maximum accuracy. *Journal of Mathematics and Physics*, 41, 04 1962.

[85] J. C. Strikwerda. *Finite difference schemes and partial differential equations. 2nd ed.* Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2nd ed. edition, 2004.

[86] S. Tan and C.-W. Shu. Inverse Lax-Wendroff procedure for numerical boundary conditions of conservation laws. *J. Comput. Phys.*, 229(21):8144–8166, 2010.

[87] V. Thomée. Stability of difference schemes in the maximum-norm. *J. Differential Equations*, 1:273–292, 1965.

[88] L. N. Trefethen. *Spectral methods in Matlab.*, volume 10. Philadelphia, PA: SIAM, 2000.

[89] L. N. Trefethen and J. A. C. Weideman. The exponentially convergent trapezoidal rule. *SIAM Rev.*, 56(3):385–458, 2014.

[90] A. E. Tzavaras. Materials with internal variables and relaxation to conservation laws. *Arch. Ration. Mech. Anal.*, 146(2):129–155, 1999.

[91] A. E. Tzavaras. Relative entropy in hyperbolic relaxation. *Commun. Math. Sci.*, 3(2):119–132, 2005.

[92] F. Vilar and C.-W. Shu. Development and stability analysis of the inverse Lax-Wendroff boundary treatment for central compact schemes. *ESAIM Math. Model. Numer. Anal.*, 49(1):39–67, 2015.

[93] W.-C. Wang and Z. Xin. Asymptotic limit of initial-boundary value problems for conservation laws with relaxational extensions. *Comm. Pure Appl. Math.*, 51(5):505–535, 1998.

[94] G. B. Whitham. *Linear and nonlinear waves.* John Wiley & Sons, Hoboken, NJ, 1974.

[95] L. Wu. The semigroup stability of the difference approximations for initial-boundary value problems. *Mathematics of Computation*, 64:71–88, 01 1995.

[96] Z. Xin and W.-Q. Xu. Stiff well-posedness and asymptotic convergence for a class of linear relaxation systems in a quarter plane. *J. Differ. Equations*, 167(2):388–437, 2000.

[97] W.-A. Yong. Boundary conditions for hyperbolic systems with stiff source terms. *Indiana Univ. Math. J.*, 48(1):115–137, 1999.

[98] W.-A. Yong. Singular perturbations of first-order hyperbolic systems with stiff source terms. *J. Differ. Equations*, 155(1):89–132, 1999.

[99] W.-A. Yong. Basic aspects of hyperbolic relaxation systems. In *Advances in the theory of shock waves*, volume 47 of *Progr. Nonlinear Differential Equations Appl.*, pages 259–305. Birkhäuser Boston, Boston, MA, 2001.

[100] W.-A. Yong. Entropy and global existence for hyperbolic balance laws. *Arch. Ration. Mech. Anal.*, 172(2):247–266, 2004.

UNIVERSITÉ DE
RENNES 1

Titre : Approximation numérique des conditions aux bords et des termes source raides dans les équations hyperboliques

Mot clés : systèmes hyperboliques d'équations aux dérivées partielles, conditions au bord, méthode des différences finies, stabilité, convergence.

Résumé : Ce travail est consacré à l'étude théorique et numérique de systèmes hyperboliques d'équations aux dérivées partielles et aux équations de transport, avec des termes de relaxation et des conditions aux bords.

Dans la première partie, on étudie la stabilité raide d'approximations numériques par différences finies du problème mixte donnée initiale-donnée au bord pour l'équation des ondes amorties dans le quart de plan. Dans le cadre du schéma discret en espace, nous proposons deux méthodes de discrétisation de la condition de Dirichlet. La première est la technique de sommation par partie et la seconde est basée sur le concept de condition au bord transparente. Nous proposons également une comparaison numérique des deux méthodes, en particulier de leur domaine de stabilité.

La deuxième partie traite de schémas numériques d'ordre élevé pour l'équation de transport avec une donnée entrante sur domaine borné. Nous construisons, implémentons et analysons la procédure de Lax-Wendroff inverse au bord entrant. Nous obtenons des taux de convergence optimaux en combinant des estimations de stabilité précises pour l'extrapolation des conditions au bord avec des développements de couche limite numérique.

Dans la dernière partie, nous étudions la stabilité de solutions stationnaires pour des systèmes non conservatifs avec des terms géométrique et de relaxation. Nous démontrons que les solutions stationnaires sont stables parmi les solutions entropique processus, qui généralisent le concept de solutions entropiques faibles. Nous supposons essentiellement que le système est complété par une entropie partiellement convexe et que, selon la dissipation du terme de relaxation, la stabilité ou la stabilité asymptotique des solutions stationnaires est obtenue.

Title: Numerical approximation of boundary conditions and stiff source terms in hyperbolic equations

Keywords: hyperbolic systems of partial differential equations, boundary conditions, finite difference methods, stability, convergence

Abstract: The dissertation focuses on the study of the theoretical and numerical analysis of hyperbolic systems of partial differential equations and transport equations, with relaxation terms and boundary conditions.

In the first part, we consider the stiff stability for numerical approximations by finite differences of the initial boundary value problem for the linear damped wave equation in a quarter plane. Within the framework of the difference scheme in space, we propose two methods of discretization of Dirichlet boundary condition. The first is the technique of summation by part and the second is based on the concept of transparent boundary conditions. We also provide a numerical comparison of the two numerical methods, in particular in terms of stability domain.

The second part is about high order numerical schemes for transport equations with nonzero incoming boundary data on bounded domains. We construct, implement and analyze the so-called inverse Lax-Wendroff procedure at incoming boundary. We obtain optimal convergence rates by combining sharp stability estimate for extrapolation boundary conditions with numerical boundary layer expansions.

In the last part, we study the stability of stationary solutions for non-conservative systems with geometric and relaxation source term. We prove that stationary solutions are stable among entropy process solution, which is a generalisation of the concept of entropy weak solutions. We mainly assume that the system is endowed with a partially convex entropy and, according to the entropy dissipation provided by the relaxation term, stability or asymptotic stability of stationary solutions is obtained.