Étude théorique et numérique de la stabilité GKS pour des schémas d'ordre élevé en présence de bords

Pierre Le Barbenchon

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601

Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique

Spécialité : *Mathématiques et leurs interactions*

Par

Pierre LE BARBENCHON

Étude théorique et numérique de la stabilité GKS pour des schémas d'ordre élevé en présence de bords

Thèse présentée et soutenue à RENNES, le 27 juin 2023

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

Rapporteurs avant soutenance

Pascal NOBLE	Professeur des Universités, INSA de Toulouse, Toulouse
Katharina SCHRATZ	Professeure des Universités, Sorbonne Université, Paris

Composition du Jury

Examinateurs :	Antoine BENOIT	Maître de Conférences, Université du Littoral Côte d'Opale, Calais
	Jean-François COULOMBEL	Directeur de recherche CNRS, Université Toulouse III Paul Sabatier, Toulouse
	Erwan FAOU	Directeur de recherche INRIA, Université de Rennes, Rennes
	Pascal NOBLE	Professeur des Universités, INSA de Toulouse, Toulouse
	Katharina SCHRATZ	Professeure des Universités, Sorbonne Université, Paris
Directeur de thèse :	Benjamin BOUTIN	Maître de Conférences, Université de Rennes, Rennes
Directeur de thèse :	Nicolas SEGUIN	Directeur de recherche INRIA, Université Côte d'Azur, Montpellier

Remerciements

J'aimerais remercier toutes les personnes qui m'ont aidé de près ou de loin pendant ma thèse, mais aussi dans ma vie pour arriver là où j'en suis aujourd'hui.

En premier lieu, je tiens à exprimer ma profonde gratitude envers mes deux directeurs de thèse, Benjamin et Nicolas, sans qui ces trois dernières années n'auraient pas été les mêmes. Comme je le dis souvent, mon choix de thèse ne s'est pas fait sur le sujet, mais bien sur mes encadrants et je serais à jamais ravi d'avoir fait ce choix. Je vous suis extrêmement reconnaissant de m'avoir fait confiance. Merci à toi, Benjamin, pour cet encadrement très privilégié. J'ai adoré ces après-midi passées à faire des maths au tableau avec toi, pouvoir avoir le droit de me tromper sans jugement, de réfléchir sans pression et de discuter de tout et de rien sans limite. Merci aussi pour la qualité de tes cours durant mon parcours scolaire : ils sont pour moi un modèle d'enseignement. Merci à Mathieu et à toi pour l'aide et les discussions sur mon avenir et pour les repas partagés ensemble. Merci à toi, Nicolas, pour ton suivi exemplaire malgré la distance entre Rennes et Montpellier cette dernière année. Je sais qu'en cas de besoin, je pouvais toujours sonner à ta porte. Tes questions toujours pertinentes m'ont apporté le recul qui me manquait parfois. Merci aussi pour tes cours durant mon cursus, que ce soit les TD d'ANUM en L3 ou les cours d'éléments finis de M2 où tu arrivais toujours à nous communiquer ta bonne humeur. Enfin, encore merci à tous les deux pour votre bienveillance et votre sensibilité qui m'inspireront tout le restant de ma vie. Je n'aurais pas pu imaginer meilleur encadrement que le vôtre.

Je voudrais remercier grandement Pascal Noble et Katharina Schratz d'avoir accepté de rapporter ma thèse. Je vous suis très reconnaissant pour votre relecture attentive et soignée. Je remercie aussi Antoine Benoit, Jean-François Coulombel et Erwan Faou d'avoir accepté d'être membres de mon jury de soutenance.

Je voudrais remercier aussi tous les professeurs de l'ENS qui m'ont accompagné durant ma scolarité et mon monitorat à l'ENS : Karine, Arnaud, Jérémy, François, Rémi, David, Nathalie et en particulier, merci à Thibaut, François et Lilian pour leur accompagnement tout au long de l'année de préparation à l'agrégation. Je souhaite également remercier tous les étudiants de l'ENS que j'ai encadrés pendant mon monitorat, cela m'a permis de consolider ma volonté de devenir enseignant. Je voudrais remercier aussi les enseignants-chercheurs de l'IRMAR qui nous ont accompagnés durant notre scolarité, en particulier, Matthieu, Miguel, Vincent, Rémi, Benoit, Christophe et Jean-Christophe.

Toujours présent dès qu'on avait une question, je voudrais remercier tout le personnel de la tour des maths qui nous a accompagnés pendant ces trois années de thèse. Notamment, merci Marie-Aude pour ta disponiblité, merci Florian pour ta gentillesse et ton aide sur les ordres missions et merci Pierre pour ton aide précieuse sur la création du package Python boundaryscheme.

’Ne fais pas tes graphiques au stylo !’, ‘Utilise la magouille de l’angle moitié’, ‘Attention, les nombres négatifs sont vos ennemis !’, ‘La récurrence, c’est comme une échelle’ : par ces quelques phrases qui m’ont marqué, je voudrais remercier tous les professeurs qui m’ont guidé durant ma scolarité à Rouen, notamment M.Béguier, Mme Legendre, Mme Lecouturier, Mme Saint-Sulpice, M.Cerisier, M.Mandon, M.Décultot, Mme Najid, M.Chénel et M.Legros.

Aussi, je souhaiterais remercier vivement François et Sophie avec qui j’ai eu le plaisir d’écrire le livre de logique pendant mes premières années de thèse. Après avoir été votre élève en cours de logique et en préparation agrégation, je vous suis reconnaissant de m’avoir embarqué dans ce périple fou d’écriture. Merci à toi, Sophie, pour ta persévérance sans faille, ton envie de toujours aller plus loin et ta rigueur. Merci pour les moments qu’on a passés chez toi à écrire le livre et à faire de la longe avec ton cheval. Merci à toi, François, pour ton enthousiasme et émerveillement face à tout, ta pédagogie inspirante et ta détermination à aider les autres. Merci pour tous les moments qu’on a passés à faire de la musique ensemble. Toute cette aventure fut une expérience très riche qui m’a appris énormément de choses tant sur le plan technique que sur le plan humain.

Il est évident que je voudrais aussi remercier mes camarades de l’ENS pour tous les bons moments passés ensemble entre le ski, la vidéo des matheux, les jeux, les discussions et les franches rigolades. Merci à Clarence, David, Théo, Tibo, Charlie, Vincent, Silvère, Antoine, Kevin, Sarah, Alain et tous les autres.

Merci à tous mes amis de comédie musicale, merci Alizée, Manon et Lisa, vous avez changé mes années de scolarité à l’ENS en nous embarquant dans ce projet colossal. Les *liaisons dangereuses* furent un tournant dans ma vie. Merci à Romane, Maxence, Marie, Raphaël, Inès, Anaëlle, Audrey, Léa pour tous ces moments passés sur les planches. Cela m’a permis aussi de pouvoir prolonger l’aventure avec *Alice aux pays des merveilles* l’année suivante. Merci à Émilie, Théo, Clémentine, Mathias, Caroline, Etienne, Camille, Clément, Vincent, Antoine et tous les autres pour cette année merveilleuse. Notamment merci à toi, Clémentine, pour ton implication inébranlable, ta grande force et ton amitié sans bornes. Merci pour les moments de colocation qu’on a partagés, en particulier les sorties au théâtre et au cinéma que j’ai beaucoup appréciées.

Ensuite, je voudrais te remercier, Maxence, pour tous les moments passés ensemble, des nuits de colles en prépa, aux matinées de jazz à tout-va en passant par les souvenirs inoubliables des colocations qu’on a vécues ensemble. Ma prépa et ma scolarité ENS n’auraient pas été les mêmes sans toi !

Un immense merci à toi, Emilie, pour ta grande amitié. Je te dois beaucoup, que ce soit, pendant la comédie musicale que j’ai adoré diriger avec toi, et pendant l’année d’agrégation où ton dévouement, ta rigueur et ton aide m’ont permis de surmonter les défis de cette année ardue. Merci !

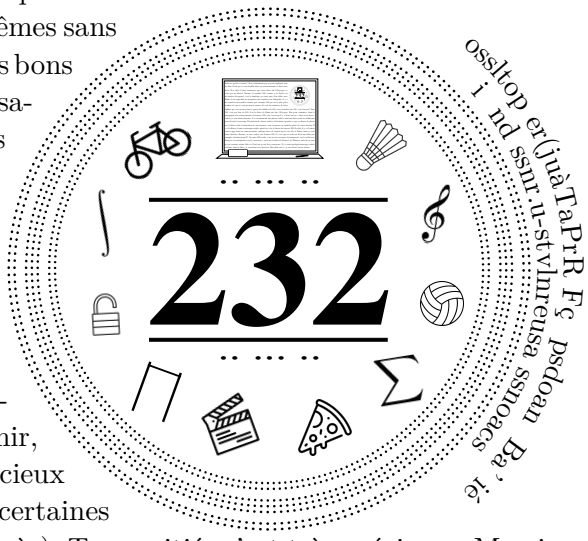
Un immense merci aussi à toi, Mathias, pour ta présence dans ma vie. Tu es et as toujours été là lorsque j’en avais besoin. Merci pour tous les moments de rire, mais aussi pour toutes les grandes discussions sur la vie qu’on a eues. Tu comptes beaucoup pour moi.

Franchement, Gaspard et Laurine, le confinement n’aurait pas été le même sans vous. Merci pour votre grande écoute et votre bienveillance qui m’inspirent tous les jours. Après les parties enRhumées de Pandemic Legacy et toutes nos discussions qui ont accompagné cette thèse,

j'espère qu'on ne s'arrêtera jamais de se retrouver pour jouer, discuter et passer du temps ensemble.

On ne peut pas faire sa thèse tout seul et je voudrais donc remercier tous les doctorants que j'ai côtoyés au sein du labo. Parmi eux, je souhaite remercier Paul, Alice, Josselin, Mégane, Jérémy, Quentin, Grégoire, Louis, Emeline, Marie, Marc, Yohan, Maxime, Antoine, Héloïse, Arthur et Jean Bussac. Merci Paul et Alice pour les matchs de badminton, les nombreuses discussions le midi dans le bureau 232, les chasses aux trésors et tous les fous rires.

Repose en paix petit bureau 232 (je parle évidemment du bureau lui-même). Pendant ma thèse, le nombre 232 a occupé une place très particulière comme on peut le voir au fil des pages de cette thèse (et pas seulement parce que c'est un nombre décagonal). Je voudrais ainsi remercier François, Thomas et Rémi pour m'avoir accueilli dans votre (notre ?) bureau ces deux dernières années. Je n'ai jamais vécu meilleur accueil que celui d'eux trois de toute ma scolarité. Ces années de thèse n'auraient vraiment pas été les mêmes sans vous. On ne peut pas résumer en quelques lignes tous les bons moments passés ensemble. Merci à notre auteur/réalisateur officiel, François, ta forte amitié, ton écoute sans faille et tes nombreuses qualités m'ont été très chères ces dernières années. Que ce soit durant les multiples projets ensemble, notamment les courts-métrages, la vulgarisation en duo dans les lycées et la colocation, ou durant les moments passés ensemble à discuter et à rire, je suis heureux d'avoir pu passer autant de temps avec toi et je n'ai pas l'intention de m'arrêter là. Merci à notre star en devenir, Thomas, ta gentillesse illimitée, ton aide et tes précieux conseils m'ont beaucoup guidé et permis de voir certaines situations sous un nouvel angle (à deux trois degrés près). Ton amitié m'est très précieuse. Merci à notre bodybuilder préféré, Rémi, ton éternelle bienveillance, ton humour dévastateur et ta joie de vivre communicative m'ont permis d'avancer mieux ces dernières années. Désormais, je te crois, avec regret, quand tu dis venir de la Réunion... Mais tu auras beau essayer de t'éloigner de nous, j'aurai toujours 232 raisons de te considérer comme un vrai ami. Entre les courts-métrages, le badminton, l'activité tractions/pizza, les jeux, j'espère que l'esprit du bureau 232 ne s'estompera jamais. Vous allez terriblement me manquer les gars.



Toujours dans les activités qui me permettent de m'épanouir, jour après jour, je voudrais remercier énormément toutes les personnes que j'ai côtoyées en faisant du jazz ces deux dernières années. Je voudrais remercier tous les enseignants du département jazz du conservatoire, en particulier, merci Sabah pour ton suivi et tes précieux conseils, merci Yannick pour les cours d'arrangement que j'ai suivis avec grande joie, merci Jacques pour les cours de flûte et la découverte plus en profondeur du jazz et de sa pratique. Merci Stéphane pour ton accompagnement durant les cours d'ensemble et merci également à toutes les personnes avec qui j'ai pu jouer, notamment Jules, Fanny, Hugo, Paul, Raphaël, Lucas, Denys, Ewen, Liliane, Noël et Youri avec qui je me suis éclaté !

Stéphane, je voudrais te remercier pour tes excellents cours en MP* tant d'un point de vue pédagogique qu'humain. Si je souhaite être enseignant aujourd'hui, c'est en partie grâce à toi et tu m'as donné le parfait exemple du professeur de prépa que j'aimerais devenir ! Merci

aussi pour la relation qu'on a développée ensemble depuis, les repas partagés avec Carole et ma famille, les moments musicaux privilégiés, les discussions sur la vie et merci beaucoup pour ton aide sur mes choix d'orientation post-thèse.

Jean, mon ami de toujours, merci à toi d'avoir toujours été là. Je ne pourrais jamais résumer en quelques lignes tout ce qu'on a traversé ensemble et tout ce qu'on a commencé ensemble : que ce soit la carrière des Sugar Honey Ice Tea, les débuts cinématographiques de P&J Movies et tous les autres délires farfelus qu'on a eus ensemble. Notre relation fraternelle m'a fait énormément grandir sur tous les plans. Merci de m'avoir aidé avec l'architecture du package Python boundaryscheme, cela prouve encore une fois à quel point j'ai de la chance de pouvoir compter sur toi dans ma vie. Les moments qu'on a partagés ensemble seront toujours gravés en moi. Merci à Elsa, Julie et Sarah pour tous les moments partagés au lycée, notamment pendant les tournages des différents courts-métrages et tous les moments de rigolades qu'on a eus ensemble.

La suite de mes remerciements va à Catherine et Yves avec qui les échanges ont toujours été très riches dans tous les domaines. Merci à tous les deux pour vos précieux conseils, votre écoute délicate et sans jugement et pour les moments passés ensemble.

Il est important aussi pour moi de remercier tous mes grands-parents qui m'ont aidé de près ou de loin à façonner la personne que je suis aujourd'hui. Merci Mamounette et Papounet pour votre aide et votre soutien durant mes études et pour tous les précieux moments que nous avons partagés malgré la distance qui nous sépare. Merci Papi Toc-Toc et Maï pour votre accueil et pour tous les moments privilégiés partagés avec les cousins chez vous. Et enfin merci Papi Jacques et Mone pour tous les moments passés ensemble depuis que je suis né, votre soutien et votre influence ont été d'une valeur inestimable. Merci !

Sans l'appui inébranlable de ma famille, je n'aurais jamais pu en arriver là. Merci à mes petits frères d'avoir toujours été présents pour moi : merci Henri pour ta générosité sans limite et ton soutien au sein de la famille, merci Antoine pour ta complicité et ton implication dans notre relation, merci Louis pour ta spontanéité et pour toutes les activités (que ce soit musicale ou sportive) que l'on a partagées ensemble. J'ai de la chance de vous avoir, vous m'avez énormément appris tout au long de ma vie. Enfin, merci Papa et Maman pour votre soutien indéfectible dans ma vie, c'est grâce à vous si je suis là aujourd'hui. Merci Papa pour ton accompagnement, ton pragmatisme et ton humour, et merci Maman pour ton écoute, tes attentions et tes mots rassurants. Vous êtes tous les deux des modèles pour moi. Je suis véritablement chanceux d'avoir une famille si aimante et je sais que je ne vous le dis pas souvent, mais je vous aime tous les cinq profondément.

Afin de conclure ces remerciements, je voudrais te remercier Lisa d'illuminer ma vie quotidiennement. Au cours de ces six dernières années, ta présence, ton amour et ton soutien ont été essentiels dans tout ce que j'ai entrepris. Je ne mesurerai jamais assez la chance que j'ai de partager ta vie et de me lever chaque matin à tes côtés. Et merci également à ta famille qui a toujours été très accueillante avec moi.

Table des Matières

LISTE DES SYMBOLES

$\lvert\cdot\rvert$	module d'un nombre complexe ou valeur absolue d'un nombre réel
$\lVert\cdot\rVert$	norme vectorielle ou norme subordonnée à une norme vectorielle
$\lvert\!\lVert\cdot\rVert\!\rvert$	norme d'opérateur
$[\![n : m]\!]$	intervalle d'entiers entre n et m (inclus)
$\mathbb{D}$	disque unité ouvert
$\overline{\mathbb{D}}$	disque unité fermé
$\mathbb{S}$	cercle unité
$\mathcal{U}$	ensemble des complexes de module strictement supérieur à 1
$\overline{\mathcal{U}}$	ensemble des complexes de module supérieur ou égal à 1
$B(a, r)$	disque ouvert de centre a et de rayon r dans $\mathbb{C}$
$\overline{B(a, r)}$	disque fermé de centre a et de rayon r dans $\mathbb{C}$
$\mathcal{C}(a, r)$	cercle de centre a et de rayon r
$\mathcal{P}(A)$	ensemble des partie de l'ensemble A
r	nombre de points à gauche dans l'écriture des schémas
p	nombre de points à droite dans l'écriture des schémas
m	nombre de points nécessaires au bord pour définir les points fantômes des schémas
s	nombre de pas en temps pour les schémas multipas
d	ordre de consistance
J	nombre d'intervalles de la discrétisation spatiale
Δx	pas de la discrétisation spatiale
Δt	pas de la discrétisation temporelle
λ	nombre de Courant défini par $\lambda \overset{\text{def}}{=} \frac{a\Delta t}{\Delta x}$
B	matrice de condition de bord défini en (1.14)
$\mathcal{B}$	matrice de condition de bord défini en (1.16)
ℓ^2	espace $\ell^2([\![-r : -1]\!] \cup \mathbb{N})$

$\mathcal{E}^s(z)$	espace vectoriel défini page 78
M	nombre de racines distinctes dans $\mathbb{D}$ de l'équation caractéristique
β	multiplicité d'une racine d'une équation
$K_{i,j}(z)$	matrice définie à la Notation 3.6 (page 78)
$H_{i,j}(z)$	matrice de Hermite, définie page 154
Δ_{KL}	déterminant de Kreiss–Lopatinskii
Δ	déterminant intrinsèque de Kreiss–Lopatinskii
$\widetilde{U_j}(z)$	transformée en $\mathcal{Z}$ de $(U_j^n)_n$
$\widehat{U^n}(\xi)$	transformée de Fourier de $(U_j^n)_j$
γ	symbole $\gamma : \xi \mapsto \sum_{k=-r}^{p} a_k e^{ik\xi}$ des schémas
Γ	courbe dans $\mathbb{C}$, souvent utilisé pour la courbe du symbole γ
$\mathrm{Ind}_\Gamma(w)$	indice complexe de $w \in \mathbb{C}$ par rapport à la courbe Γ
ind	indice d'un opérateur de Fredholm
Ran	image d'un opérateur
ker	noyau d'un opérateur
$\Lambda(A)$	spectre de la matrice ou de l'opérateur A
$\Lambda_\varepsilon(A)$	ε-pseudospectre de A
$R_z(A)$	résolvante de A en z, *i.e.* $(A-z)^{-1}$
T_J	matrice Toeplitz
$\widetilde{T_J}$	matrice Quasi-Toeplitz
T_J°	matrice Toeplitz circulante
$T_{\mathbb{Z}}$	opérateur Toeplitz sur $\mathbb{Z}$
$T_{\mathbb{N}}$	opérateur Toeplitz sur $\mathbb{N}$
$\widetilde{T_{\mathbb{N}}}$	opérateur Quasi-Toeplitz sur $\mathbb{N}$
$\lim_{n\to\infty} A_n$	limite d'ensemble, définie en Notation 2.27 (page 58)
$\widetilde{\lim}_{n\to\infty} A_n$	limite d'ensemble à sous-suite près, définie en Définition 2.12 (page 51)
$\begin{bmatrix} n \\ \ell \end{bmatrix}$	coefficient défini en Annexe E
$\#\text{zéros}_f(A)$	nombre de zéros (comptés avec multiplicité) de la fonction f dans l'ensemble A

$\#\text{pôles}_f(A)$	nombre de pôles (comptés avec multiplicité) de la fonction f dans l'ensemble A
$\delta(n)$	symbole de Kronecker en zéro (vaut 1 si $n = 0$ et 0 sinon)
e_k	le k-ième vecteur de la base canonique de $\mathbb{R}^d$ pour un certain d
O, o	notation de Landau
${}^tA$ ou $A^\top$	transposée de la matrice A

PLAN DU MANUSCRIT

Dans ce manuscrit, nous étudions la stabilité forte des schémas numériques explicites à un pas à coefficients constants, posés sur le demi-espace $\mathbb{N}$ et possédant un bord à gauche. On suppose que ces schémas sont consistants avec l'équation de transport scalaire en dimension 1 comportant une donnée de bord à gauche.

L'enjeu est de trouver des stratégies efficaces et robustes pour étudier la stabilité de ces schémas, notamment au travers d'outils numériques et de la condition de Kreiss–Lopatinskii uniforme représentée par le déterminant de Kreiss–Lopatinskii.

Le manuscrit est découpé en trois parties : la première fait office d'introduction du sujet, mais elle fait aussi le lien entre différentes approches de la stabilité et présente la théorie GKS sur laquelle s'appuie le reste du manuscrit, la deuxième donne le détail des articles soumis et présente aussi des détails d'implémentation numérique et la troisième regroupe plusieurs annexes sur des notions qui servent tout au long du manuscrit.

La première partie comporte quatre chapitres :

- **Chapitre 1 : Notions générales sur la stabilité**

  Ce chapitre introduit l'équation de transport dont on veut approcher les solutions et la définition de schémas que l'on utilise. On y donne également les définitions de convergence, consistance et des notions de stabilité mises en jeu. On présente ensuite une première approche de la stabilité utilisant les notions de valeur propre et valeur propre généralisée et on donne une première condition nécessaire et suffisante pour obtenir la stabilité. Enfin, on conclut ce chapitre avec quelques compléments : une deuxième façon de voir le schéma déjà introduit et une analyse de consistance pour les conditions de bord Lax-Wendroff inverse et Lax-Wendroff inverse simplifiée.

- **Chapitre 2 : Analyse spectrale de matrices Toeplitz**

  Ce chapitre est relativement indépendant de la suite du manuscrit, c'est un travail prospectif effectué pendant la première année de la thèse sur une approche complémentaire de la stabilité introduite dans le chapitre précédent. Celle-ci se base sur les notions de matrices et opérateurs Toeplitz qui représentent des schémas numériques. On en étudie les propriétés spectrales grâce aux notions de spectre asymptotique, pseudospectre et grâce au Kreiss Matrix Theorem. On y recense plusieurs questions ouvertes dont les réponses pourraient aider à trouver des outils numériques pour l'analyse de la stabilité, notamment l'utilisation du pseudospectre et de ses propriétés géométriques.

- **Chapitre 3 : Théorie de Gustafsson, Kreiss et Sundström**

  Dans ce chapitre, on présente la stratégie d'étude de stabilité forte que l'on utilise dans toute la suite. En introduisant la transformée en $\mathcal{Z}$, on définit la notion de déterminant de Kreiss–Lopatinskii dont on aura besoin pour l'étude faite dans la deuxième partie du manuscrit. On présente ensuite une deuxième condition nécessaire et suffisante pour obtenir la stabilité : la condition de Kreiss–Lopatinskii uniforme.

- **Chapitre 4 : Revue de la stabilité et présentation des contributions**

  Ce chapitre fait un bilan des discussions des trois premiers chapitres, introduit les contributions de la thèse qui sont détaillées dans la deuxième partie du manuscrit et explique comment ces contributions s'introduisent dans la littérature.

La deuxième partie comporte trois chapitres :

- **Chapitre 5 : Stability of one-step explicit totally upwind schemes**

  Ce chapitre est dédié à l'article [BLBS23a] écrit avec Benjamin Boutin et Nicolas Seguin :

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

  On y ajoute quelques compléments pour préciser ou pour généraliser certaines preuves.

- **Chapitre 6 : Stability of one-step explicit schemes**

  Ce chapitre est dédié à l'article [BLBS23b] écrit avec Benjamin Boutin et Nicolas Seguin :

  Stability of finite difference schemes
  for the hyperbolic initial boundary value problem
  by winding number computations

  On y ajoute quelques compléments pour préciser certaines preuves ou donner des preuves alternatives. Enfin, on conclut ce chapitre en étudiant succinctement la généralisation aux schémas multipas.

- **Chapitre 7 : Aspects d'implémentation numérique**

  Ce chapitre décrit l'aspect numérique de la stratégie décrite dans les deux chapitres précédents, notamment on donne certains détails du code informatique utilisé. Enfin, on explique les folioscopes présents en bas de page de ce manuscrit.

La troisième partie comporte cinq chapitres :

- **Chapitre A : Quelques éléments d'analyse complexe**
- **Chapitre B : Quelques éléments sur la transformée en $\mathcal{Z}$**
- **Chapitre C : Quelques éléments de théorie spectrale**
- **Chapitre D : Quelques éléments sur le schéma leap-frog**
- **Chapitre E : Quelques éléments sur un coefficient binomial modifié**

Première partie

Théories de stabilité : aperçu général et connexions

Chapitre 1

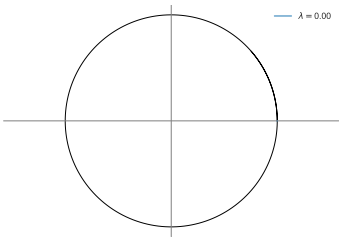
Notions générales sur la stabilité

Dans tout ce manuscrit, on va étudier la stabilité des schémas numériques explicites à un pas à coefficients constants possédant des conditions de bord. Les solutions des schémas numériques que l'on va considérer approchent les solutions de l'équation de transport scalaire 1D à vitesse constante. L'équation de transport est un modèle jouet pour notre étude, mais celle-ci peut se généraliser à d'autres équations hyperboliques linéaires.

1.1 Cadre

On cherche à approcher la solution $u : (t, x) \mapsto u(t, x) \in \mathbb{R}$ de l'équation de transport

$$\begin{cases} \partial_t u + a\partial_x u = 0 & x \in [0,1], t \geqslant 0, \\ u(t,0) = g(t) & t \geqslant 0, \\ u(0,x) = f(x) & x \in [0,1]. \end{cases} \tag{1.1}$$



avec une vitesse du transport a supposée positive, f une donnée initiale et g une donnée de bord à gauche. Puisque la vitesse a est supposée positive, le bord gauche en $x = 0$ est un bord entrant et le bord droit en $x = 1$ est un bord sortant. Pour que la solution u soit régulière, il faut que les données f et g vérifient des conditions de compatibilité que l'on suppose vérifiées pour la suite.

Afin d'approcher les solutions de l'équation (1.1), on discrétise l'espace en $J \in \mathbb{N}^*$ intervalles de la forme $[j\Delta x, (j+1)\Delta x]$ pour $j \in [\![0 : J-1]\!]$ avec $\Delta x = \frac{1}{J}$. On discrétise le temps par le pas $\Delta t > 0$ de sorte que la quantité $\lambda \overset{\text{def}}{=} \frac{a\Delta t}{\Delta x}$ soit constante, celle-ci s'appelle le nombre de Courant et a été introduite par Courant, Friedrichs et Lewy [CFL28] en 1928. Dans l'étude que l'on fait par la suite, on discute de la condition CFL (acronyme des noms Courant, Friedrichs et Lewy) qui est un intervalle dans lequel doit être pris le nombre de Courant λ pour avoir des bonnes propriétés de stabilité.

On va utiliser un schéma numérique dont la solution $(U_j^n)_{j,n}$ se veut proche de la solution $u(n\Delta t, j\Delta x)$ de l'équation de transport (1.1). Il y a de nombreuses façons de définir un schéma numérique, parmi les plus classiques, on trouve les schémas explicites et implicites avec un ou plusieurs pas en temps avec coefficients constants ou variables. Dans ce manuscrit, on se focalise principalement sur les schémas explicites à un pas à coefficients constants. Les schémas multipas sont abordés dans les Section 2.4.4 et Section 6.7, ainsi que dans l'Annexe D.

1.1.1 Équation intérieure

La forme générale des schémas numériques envisagés est explicite à un pas, c'est-à-dire :

$$U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{k+j}^n, \quad j \in [\![0 : J]\!], \ n \geqslant 0, \tag{1.2a}$$

où r et p sont des entiers fixés et $(a_k)_{k=-r}^p$ sont des constantes réelles. Les coefficients a_{-r} et a_p sont supposés non nuls. Les coefficients $(a_k)_{k=-r}^p$ sont indépendants de j, le schéma est donc le même en tout point de la discrétisation. On représente l'interdépendance des valeurs $(U_j^n)_{j,n}$ du schéma (1.2a) dans la Figure 1.1.

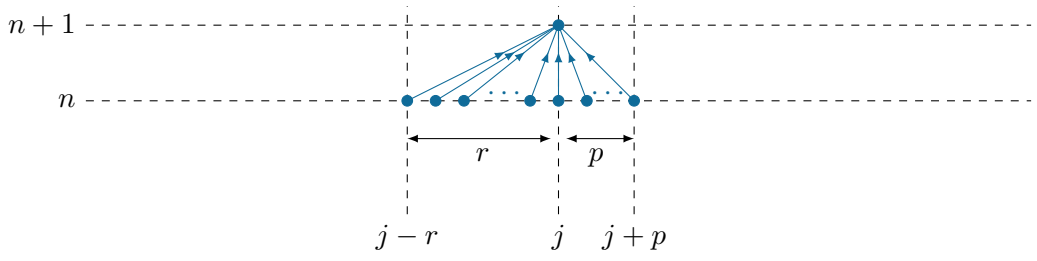
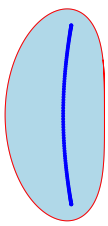


FIGURE 1.1 – Discrétisation du schéma intérieur (1.2a).

Exemple 1.1. Voici quelques exemples de schémas numériques explicites à un pas.



- le schéma décentré amont, dit *upwind*, est défini de la manière suivante :

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

La constante r vaut 1 et la constante p vaut 0. Ce schéma vient du fait qu'on discrétise la dérivée partielle temporelle $\partial_t u$ par $\frac{U_j^{n+1}-U_j^n}{\Delta t}$ et la dérivée partielle spatiale $\partial_x u$ par $\frac{U_j^n - U_{j-1}^n}{\Delta x}$. Ainsi, la discrétisation de l'équation de transport (1.1) donne

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a \frac{U_j^n - U_{j-1}^n}{\Delta x} = 0$$

qui est une autre écriture de (Upw) par définition de λ. Le terme *upwind* vient du fait que la vitesse a est positive, si la vitesse a était négative, le schéma upwind serait défini sur les points j et $j+1$.

- le schéma *Beam-Warming*, introduit par [WB76], est défini de la manière suivante :

$$U_j^{n+1} = \frac{\lambda(\lambda-1)}{2} U_{j-2}^n + \lambda(2-\lambda) U_{j-1}^n + \frac{(\lambda-1)(\lambda-2)}{2} U_j^n. \tag{BW}$$

La constante r vaut 2 et la constante p vaut 0. Ce schéma peut être vu comme un schéma upwind du deuxième ordre. Les deux schémas upwind et Beam-Warming sont des schémas totalement décentrés ($p = 0$) que l'on étudiera plus en détail dans le Chapitre 5. En présence de bord, si la vitesse a de (1.1) était négative, le schéma de Beam-Warming serait défini sur les points j, $j+1$ et $j+2$, on aurait alors $r = 0$ et $p = 2$.

- le schéma *O3* est défini de la manière suivante :

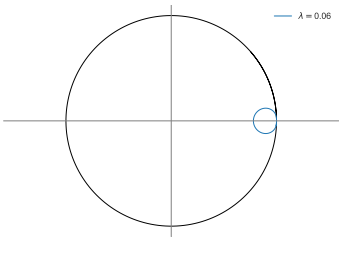
$$U_j^{n+1} = \left(\frac{\lambda^3}{6} - \frac{\lambda}{6}\right) U_{j-2}^n + \left(\lambda + \frac{\lambda^2}{2} - \frac{\lambda^3}{2}\right) U_{j-1}^n + \left(1 - \frac{\lambda}{2} - \lambda^2 + \frac{\lambda^3}{2}\right) U_j^n + \left(\frac{\lambda^2}{2} - \frac{\lambda^3}{6} - \frac{\lambda}{3}\right) U_{j+1}^n. \tag{O3}$$

La constante r vaut 2 et la constante p vaut 1. Il s'appelle « O3 » car on verra dans la suite que c'est un schéma consistant d'ordre 3. Par exemple, Dakin, Desprès et Jaouen [DDJ18] utilisent ce schéma pour leurs illustrations numériques.

- le schéma *Lax-Wendroff 5* est défini de la manière suivante :

$$U_j^{n+1} = \frac{\lambda(\lambda-2)(\lambda-1)(\lambda+1)(\lambda+2)}{120} U_{j-3}^n - \frac{\lambda(\lambda-1)(\lambda-3)(\lambda+1)(\lambda+2)}{24} U_{j-2}^n + \frac{\lambda(\lambda-2)(\lambda-3)(\lambda+1)(\lambda+2)}{12} U_{j-1}^n + \left(1 - \frac{\lambda(\lambda^4-3\lambda^3-5\lambda^2+15\lambda+4)}{12}\right) U_j^n + \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)(\lambda+2)}{24} U_{j+1}^n - \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)(\lambda+1)}{120} U_{j+2}^n. \tag{LW5}$$

La constante r vaut 3 et la constante p vaut 2. Le schéma Lax-Wendroff usuel, introduit



dans [LW60], s'écrit

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = \frac{\Delta t}{2\Delta x^2}a^2(U_{j+1}^n - 2U_j^n + U_{j-1}^n),$$

ou de manière équivalente,

$$U_j^{n+1} = \frac{\lambda^2 + \lambda}{2}U_{j-1}^n + (1-\lambda^2)U_j^n + \frac{\lambda^2 - \lambda}{2}U_{j+1}^n. \tag{LW}$$

Lorcher et Munz [LM06] déclinent ce schéma à différents ordres et donnent notamment le schéma Lax-Wendroff 5 que l'on abrège par « LW5 ». Les schémas LW5 et O3 illustreront le propos du Chapitre 6.

1.1.2 Conditions de bords numériques

Pour définir la mise à jour U_j^{n+1} dans (1.2a) pour $j \in [\![0 : J]\!]$, il est nécessaire de donner un sens aux valeurs U_{j+k}^n pour $j + k < 0$ et $j + k > J$. On utilise alors pour cela des points fantômes. Pour ceux de gauche, on pose les r équations suivantes :

$$U_j^n = \sum_{k=0}^{m-1} b_{j,k}U_k^n + g_j^n, \quad j \in [\![-r : -1]\!],\ n \geqslant 0, \tag{1.2b}$$

où m est un entier fixé, $(b_{j,k})$ des constantes réelles et (g_j^n) des données de bord numériques qui peuvent, de surcroît, dépendre de la donnée de bord physique g.

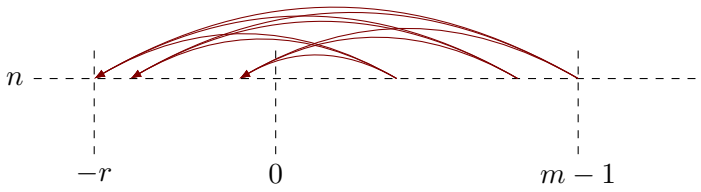
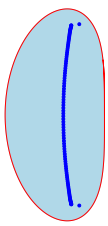


FIGURE 1.2 – Discrétisation du bord gauche (1.2b).

De même pour les points fantômes à droite, on pose les p équations suivantes :

$$U_j^n = \sum_{k=0}^{m-1} c_{j,k}U_{J-k}^n + g_j^n, \quad j \in [\![J+1 : J+p]\!],\ n \geqslant 0, \tag{1.2c}$$

où $(c_{j,k})$ sont des constantes réelles et (g_j^n) des données de bord numérique. Étant donné qu'on traite le bord droit, malgré leurs noms, les valeurs g_j^n ne font pas référence à la donnée de bord à gauche g.



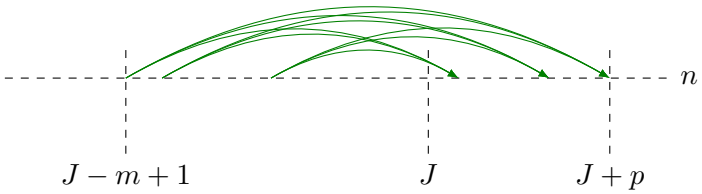


FIGURE 1.3 – Discrétisation du bord droite (1.2c).

Enfin, pour l'initialisation du schéma, on utilise la condition initiale f évaluée en tous les $j\Delta x$ pour $j \in [\![0 : J]\!]$, on pose $f_j \overset{\text{def}}{=} f(j\Delta x)$, ce qui donne

$$U_j^0 = f_j, \quad j \in [\![0 : J]\!]. \tag{1.2d}$$

Exemple 1.2. Voici quelques exemples de conditions de bord :

- les conditions de bord de Dirichlet homogène :

$$U_j^n = 0, \quad \forall j \in [\![-r : -1]\!] \cup [\![J+1 : J+p]\!], n \in \mathbb{N} \tag{1.3}$$

- les conditions de bord de Dirichlet non homogène :

$$U_j^n = g_j^n, \quad \forall j \in [\![-r : -1]\!] \cup [\![J+1 : J+p]\!], n \in \mathbb{N} \tag{1.4}$$

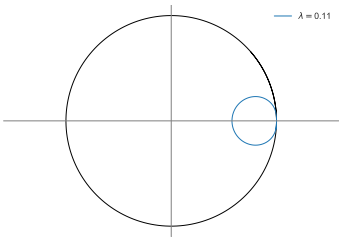
  Par exemple, on pourrait vouloir utiliser la donnée de bord g dans le schéma en posant $g_j^n = g(n\Delta t)$ pour tout $j \in [\![-r : -1]\!]$ ou même utiliser les dérivées (si elles existent) de la donnée g.

- les conditions de Neumann en entrée et en sortie : elles sont étudiées par Kreiss [Kre66] et par Goldberg [Gol77] et plus récemment par Coulombel et Lagoutière [CL20]. Elles consistent à utiliser des dérivées discrètes pour définir les points fantômes. On peut donner comme exemple $U_{-1}^n = U_0^n$, $U_{J+1}^n = 2U_J^n - U_{J-1}^n$, etc.

- la méthode Lax-Wendroff inverse en entrée, introduite par [TS10] et simplifiée par Vilar et Shu [VS15] : elle consiste à utiliser l'équation aux dérivées partielles (ici l'équation de transport (1.1)) pour transformer les dérivées spatiales en dérivées temporelles, afin d'utiliser les dérivées de la condition de bord g pour définir les points fantômes. On explique plus en détails cette méthode, ainsi que la méthode Lax-Wendroff inverse simplifiée, dans les compléments de ce chapitre (Section 1.5.2, page 39).

- la méthode de reconstruction, utilisée dans l'article de Dakin, Després et Jaouen [DDJ18] : elles vont être décrites et étudiées dans la Section 6.4.3 (page 159) au Chapitre 6.

D'autres stratégies de bord existent dans la littérature (conditions de bord transparentes ...) mais certaines n'entrent a priori pas directement dans le cadre de notre étude.



Le schéma que l'on a construit est donc de la forme (1.2a)-(1.2b)-(1.2c)-(1.2d) que l'on abrège en (1.2), on a représenté sa discrétisation en Figure 1.5 (page 37). On définit d'une deuxième manière le schéma explicite à un pas dans les compléments de ce chapitre (Section 1.5.1, page 37). On veut maintenant discuter de la convergence de la solution $(U_j^n)_{j,n}$ de ce schéma vers la solution u de (1.1).

1.2 Convergence

Définition 1.3 (Convergence)**.** On dit qu'un schéma approchant (1.1) converge pour une certaine norme $\|\cdot\|$ (qui peut dépendre du maillage $(\Delta t, \Delta x)$) si la solution $(U_j^n)_{j,n}$ du schéma converge vers la solution u de (1.1), autrement dit

$$\|(U_j^n - u(n\Delta t, j\Delta x))_{j,n}\|_{\Delta t,\Delta x} \xrightarrow[\Delta t,\Delta x\to 0]{} 0.$$

Comme norme $\|\cdot\|_{\Delta t,\Delta x}$, on peut, par exemple, utiliser la norme 2 en espace et infinie en temps, autrement dit :

$$\|U_j^n\|_{\Delta t,\Delta x} \overset{\text{def}}{=} \sqrt{\sup_{n\geqslant 0} \sum_{j=0}^{J} |U_j^n|^2 \Delta x}.$$

Cet exemple de norme est utilisé dans le livre de Strikwerda [Str04].

Comme le schéma (1.2) est un schéma dit *linéaire*, on peut utiliser le théorème de Lax, démontré par Lax et Richtmyer dans [LR56] en 1956. Le livre [Str04] donne davantage de détails sur ce théorème. Il utilise les propriétés de consistance et de stabilité que l'on va décrire dans les deux sous-sections suivantes. Informellement, on dit qu'un schéma est consistant si la solution $(u(n\Delta t, j\Delta x))_{j,n}$ est solution du schéma à une petite erreur près et on dit qu'un schéma est stable si on peut contrôler la solution du schéma par les données initiales et de bord. Par linéarité, cela permet alors de contrôler l'erreur de l'étude de consistance par les erreurs d'approximation commises sur les données initiales et de bord.

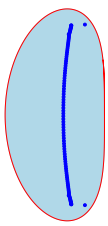
Théorème 1.4 (Lax)**.** *Un schéma linéaire est convergent si, et seulement si, il est consistant et stable.*

Remarque 1.5. Pour démontrer ce théorème, il faut introduire un terme source dans le schéma (1.2) et en étudier la consistance et la stabilité, l'étude de la stabilité avec un terme source est faite dans les travaux de Coulombel [Cou13].

Dans la suite, on définit rigoureusement la consistance et la stabilité du schéma (1.2). Ensuite, on va supposer que le schéma (1.2) est consistant et on va en étudier la stabilité.

1.2.1 Consistance

Définition 1.6 (Consistance à l'intérieur)**.** On dit que l'intérieur du schéma (1.2a) est *consistant* d'ordre $d_1 \geqslant 0$ si toute solution u assez régulière de (1.1) vérifie pour tout $j \in [\![r : J - p]\!]$ pour



tout $n \in \mathbb{N}$,

$$u((n+1)\Delta t, j\Delta x) - \sum_{k=-r}^{p} a_k u(n\Delta t, (j+k)\Delta x) = \Delta t O(\Delta x^{d_1})$$

avec d_1 maximal.

Le schéma upwind (Upw) défini à l'Exemple 1.1 est consistant d'ordre 1, car

$$\begin{aligned}
&u((n+1)\Delta t, j\Delta x) - \lambda u(n\Delta t, (j-1)\Delta x) - (1-\lambda)u(n\Delta t, j\Delta x) \\
&= u(n\Delta t, j\Delta x) + \Delta t \partial_t u(n\Delta t, j\Delta x) + O(\Delta t^2) \\
&\qquad - \lambda(u(n\Delta t, j\Delta x) - \Delta x \partial_x u(n\Delta t, j\Delta x) + O(\Delta x^2)) - (1-\lambda)u(n\Delta t, j\Delta x) \\
&= \Delta t \partial_t u(n\Delta t, j\Delta x) + a\Delta t \partial_x u(n\Delta t, j\Delta x) + O(\Delta x^2) \\
&= \Delta t O(\Delta x)
\end{aligned}$$

en utilisant l'équation (1.1) et que $\lambda \Delta x = a\Delta t$ avec a et λ fixés. On pourrait faire le même type de calcul en utilisant les développements de Taylor de u afin de trouver que le schéma de Beam-Warming (BW) est d'ordre 2, le schéma O3 (O3) est d'ordre 3 et le schéma Lax-Wendroff 5 (LW5) est d'ordre 5.

Définition 1.7 (Consistance au bord gauche)**.** On dit que le bord du schéma (1.2b) est *consistant* d'ordre $d_2 \geqslant 0$ si toute solution u assez régulière de (1.1) vérifie, pour tout $j \in [\![-r : -1]\!]$, pour tout $n \in \mathbb{N}$

$$g\left(n\Delta t - \frac{j\Delta x}{a}\right) - \sum_{k=0}^{m-1} b_{j,k} u(n\Delta t, k\Delta x) - g_j^n = O(\Delta x^{d_2})$$

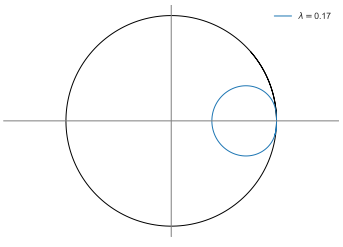
avec d_2 maximal.

Comme on a rajouté des points fantômes à l'extérieur du domaine, la solution exacte u de (1.1) posée sur le domaine $[0,1]$ ne suffit pas pour étudier la consistance de la condition de bord. Cependant, la solution u est prolongeable de manière régulière au voisinage de 0, en effet, en posant, pour tout $x \in]-\infty, 1]$,

$$f_\sharp(x) = \begin{cases} f(x) & \text{si } x \in [0,1], \\ g(-x/a) & \text{si } x < 0, \end{cases}$$

la solution $u_\sharp$ de l'équation $\partial_t u(t,x) + a\partial_x u(t,x) = 0$ sur $\mathbb{R}^+ \times]-\infty, 1]$, munie de la condition initiale $u(0,\cdot) = f_\sharp$, coïncide avec la solution u de l'équation (1.1) sur $[0,1]$. Dans la Définition 1.7, pour $j \in [\![-r : -1]\!]$ et $n \in \mathbb{N}$, l'expression $g(n\Delta t - \frac{j\Delta x}{a})$ correspond à $u_\sharp(n\Delta t, j\Delta x)$ le prolongement de la solution u sur les points fantômes de gauche.

En pratique, on étudie la consistance au bord avec des développements de Taylor, comme on le voit pour la méthode Lax-Wendroff inverse et Lax-Wendroff inverse simplifiée dont la



consistance est détaillée dans les compléments de ce chapitre à la Section 1.5.2 (page 39).

De manière similaire, on peut définir la consistance au bord droit.

Définition 1.8 (Consistance au bord droit)**.** On dit que le bord du schéma (1.2c) est *consistant* d'ordre $d_3 \geqslant 0$ si toute solution assez régulière u de (1.1) vérifie, pour tout $j \in [\![J+1 : J+p]\!]$, pour tout $n \in \mathbb{N}$ tel que $n\Delta t > \frac{j\Delta x}{a}$,

$$g(n\Delta t - \frac{j\Delta x}{a}) - \sum_{k=0}^{m-1} c_{j,k} u(n\Delta t, k\Delta x) - g_j^n = O(\Delta x^{d_3})$$

avec d_3 maximal.

Comme pour la consistance au bord gauche, la solution exacte u de (1.1) posée sur le domaine $[0,1]$ ne suffit pas pour étudier la consistance de la condition de bord de droite. Donc on prolonge la solution au voisinage de 1, en posant, pour tout $x \in [0, +\infty[$,

$$f_\flat(x) = \begin{cases} f(x) & \text{si } x \in [0,1], \\ 0 & \text{si } x > 1, \end{cases}$$

la solution $u_\flat$ de l'équation $\partial_t u(t,x) + a\partial_x u(t,x) = 0$ sur $\mathbb{R}^+ \times [0, +\infty[$, munie de la condition initiale $u(0,\cdot) = f_\flat$ et de la condition de bord g, coïncide avec la solution u de l'équation (1.1) sur $[0,1]$ et vaut, pour tout $(t,x) \in \mathbb{R}^+ \times [0, +\infty[$,

$$u_\flat(t,x) = \begin{cases} 0 & \text{si } t \leqslant \frac{x-1}{a} \\ f(x-at) & \text{si } t \in]\frac{x-1}{a}, \frac{x}{a}] \\ g(t - \frac{x}{a}) & \text{si } t > \frac{x}{a}. \end{cases}$$

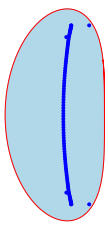
C'est pourquoi dans la Définition 1.8, on utilise l'expression $g(n\Delta t - \frac{j\Delta x}{a})$ qui correspond à $u_\flat(n\Delta t, j\Delta x)$ pour le prolongement de la solution u sur les points fantômes de droite.

Définition 1.9 (Consistance)**.** On dit que le schéma (1.2) est *consistant* d'ordre d si (1.2a) est consistant d'ordre d_1, (1.2b) est consistant d'ordre d_2, (1.2c) est consistant d'ordre d_3 et $d = \min(d_1, d_2, d_3)$.

Dans toute la suite du manuscrit, on suppose que les coefficients a_k, $b_{j,k}$, $c_{j,k}$ et g_j^n ont été choisis de sorte que le schéma soit consistant à un certain ordre. Ce qui va nous intéresser, c'est la notion de stabilité afin d'obtenir la convergence du schéma, puisque grâce au théorème de Lax (Théorème 1.4), un schéma linéaire consistant est convergent si, et seulement si, il est stable.

1.2.2 Stabilité

On va maintenant introduire la notion de stabilité qui va nous intéresser dans toute la suite du manuscrit. Informellement, on veut contrôler la solution $(U_j^n)_{j,n}$ du schéma par rapport à la



donnée initiale $(f_j)_j$ et à la donnée de bord numérique $(g_j^n)_{j,n}$, c'est-à-dire une expression de la forme

$$\text{« } \|U\| \leqslant C(\|f\| + \|g\|) \text{ »},$$

pour des normes $\|\cdot\|$ différentes que l'on va définir. Suite à la Remarque 1.5, afin de démontrer la convergence du schéma, il faut aussi contrôler la solution du schéma par rapport au terme source.

Une approche possible est d'utiliser une matrice pour représenter le schéma (comme on le fait dans le Chapitre 2) et de borner ses puissances. Par exemple, c'est la stratégie utilisée dans [DDJ18], présentée aussi dans [VS15]. Comme on le voit dans le chapitre suivant, la difficulté vient du fait qu'il faut que les puissances soient bornées indépendamment de la taille de la discrétisation J (qui est aussi la dimension de la matrice utilisée).

On va utiliser une approche alternative dans les Chapitres 3, 5 et 6 : celle de Gustafsson, Kreiss et Sundström [GKS72] de 1972. On s'appuie sur la définition de stabilité donnée dans [GKS72] et reprise par Beam, Warming et Yee [BWY82]. On explique cela dans la section suivante.

1.3 Définition de la stabilité GKS

Avant d'introduire les définitions de stabilité que l'on va utiliser, on commence par discuter informellement de la façon de voir la stabilité pour le problème (1.2) qui est défini sur $[\![0 : J]\!]$.

On va séparer ce problème de stabilité en trois autres problèmes de stabilité qui permettent d'isoler l'équation intérieure (1.2a), l'équation du bord gauche (1.2b) et l'équation du bord droit (1.2c).

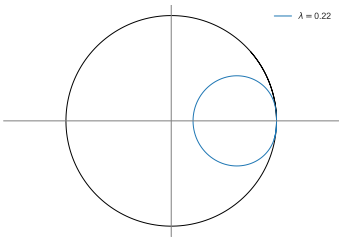
Le schéma lié à l'équation intérieure, posé sur $\mathbb{Z}$, est le suivant :

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{p} a_k U_{k+j}^n, & j \in \mathbb{Z},\ n \geqslant 0 \\ U_j^0 = f_j, & j \geqslant 0. \end{cases} \tag{1.5}$$

On va étudier la stabilité de (1.5) dans la Section 1.3.1 sur la *Cauchy-stabilité*.

Le deuxième schéma que l'on introduit et qui est lié à l'équation de bord gauche est le schéma, posé sur $\mathbb{N}$, suivant :

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{p} a_k U_{k+j}^n, & j \geqslant 0,\ n \geqslant 0 & \text{(1.6a)} \\ U_j^n = \displaystyle\sum_{k=0}^{m-1} b_{j,k} U_k^n + g_j^n, & j \in [\![-r : -1]\!],\ n \geqslant 0, & \text{(1.6b)} \\ U_j^0 = 0, & j \geqslant 0. & \text{(1.6c)} \end{cases}$$



On va étudier la stabilité de (1.6) dans la Section 1.3.2 qui introduit la notion de *stabilité forte*.

Le troisième schéma que l'on introduit et qui est lié à l'équation de bord droit est le schéma, posé sur $-\mathbb{N}\cup[\![0:J]\!]$, suivant :

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{p} a_k U_{k+j}^n, & -\infty < j \leqslant J,\ n \geqslant 0 \\ U_j^n = \displaystyle\sum_{k=0}^{m-1} c_{j,k} U_{J-k}^n + g_j^n, & j \in [\![J+1:J+p]\!],\ n \geqslant 0, \\ U_j^0 = 0, & -\infty < j \leqslant J. \end{cases} \tag{1.7}$$

L'étude de la stabilité de (1.7) est similaire à celle du schéma (1.6), il suffit d'inverser le sens du schéma en posant $U_j^{n+1} = \sum_{k=-p}^{r} a_{-k} U_{k+j}^n$, le bord droit devient alors un bord gauche et on effectue exactement la même analyse que pour le schéma (1.6) lié à l'équation de bord gauche.

Comme dans [BWY82], la stabilité du schéma borné en espace (1.2) va être définie (voir Définition 1.17) en disant que le schéma (1.5) est Cauchy-stable et les deux schémas (1.6) et (1.7) sont fortement stables.

L'intérêt de séparer le premier problème en trois autres problèmes est double. Premièrement, cela permet d'étudier séparément les deux bords, d'étudier séparément les bords et la condition initiale et de traiter le schéma posé sur $\mathbb{Z}$ avec des outils d'analyse de Fourier. Deuxièmement, le fait de poser les schémas (1.6) et (1.7) sur $\mathbb{N}$ et $-\mathbb{N}\cup[\![0:J]\!]$ permet d'étudier toutes les discrétisations $([\![0:J]\!])_{J\in\mathbb{N}^*}$ à la fois.

1.3.1 Cauchy-stabilité

Afin que le schéma (1.2) avec bords soit stable, il est naturel de supposer que le schéma (1.5) sans bord est stable. Pour étudier la stabilité du schéma (1.5), on utilise des outils d'analyse de Fourier et la notion de *symbole* que l'on va introduire dans cette section puisque le schéma est posé sur $\mathbb{Z}$.

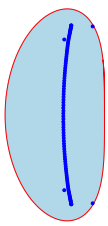
La transformée de Fourier en espace d'une solution (U_j^n) du schéma (1.5) est définie, pour tout $\xi \in \mathbb{R}$, de la manière suivante :

$$\widehat{U^n}(\xi) \overset{\text{def}}{=} \sum_{j\in\mathbb{Z}} U_j^n e^{-ij\xi}.$$

Quand on applique la transformée de Fourier à l'équation principale de (1.5), on obtient l'expression suivante :

$$\forall \xi \in \mathbb{R}, \quad \widehat{U^{n+1}}(\xi) = \sum_{k=-r}^{p} a_k e^{ik\xi} \widehat{U^n}(\xi). \tag{1.8}$$

La quantité $\sum_{k=-r}^{p} a_k e^{ik\xi}$ est le multiplicateur de Fourier du schéma, on parle de *symbole* du schéma. Dans le livre de Strikwerda [Str04], il parle de *facteur d'amplification*.



Définition 1.10 (Symbole)**.** Soit un schéma de la forme (1.5). Son *symbole* γ est défini de la manière suivante

$$\forall \xi \in \mathbb{R}, \quad \gamma(\xi) \overset{\text{def}}{=} \sum_{k=-r}^{p} a_k e^{ik\xi}. \tag{1.9}$$

En itérant l'expression (1.8), on trouve l'expression suivante :

$$\forall \xi \in \mathbb{R}, \quad \widehat{U^n}(\xi) = \gamma(\xi)^n \widehat{U^0}(\xi). \tag{1.10}$$

Pour définir la notion de Cauchy-stabilité, on introduit les normes suivantes :

$$\|U^n\|_2^2 = \sum_{j\in\mathbb{Z}} |U_j^n|^2 \quad \text{et} \quad \|\widehat{U^n}\|_{L^2}^2 = \frac{1}{2\pi}\int_0^{2\pi} |\widehat{U^n}(\xi)|^2 d\xi.$$

Par l'égalité de Parseval, pour $(U_j^n)_j$ associé à $(\widehat{U^n}(\xi))_\xi$, on a

$$\|U^n\|_2 = \|\widehat{U^n}\|_{L^2}. \tag{1.11}$$

On veut que la solution $(U_j^n)_{j,n}$ soit contrôlée par la condition initiale $(U_j^0)_j = (f_j)_j$.

Définition 1.11 (Cauchy-stabilité)**.** On dit qu'un schéma de la forme (1.5) est *Cauchy-stable* s'il existe une constante $C > 0$ telle que, pour toute condition initiale $(f_j)_j \in \ell^2(\mathbb{N})$, pour toute solution $(U_j^n)_{j,n}$ du schéma (1.5), pour tout $n \in \mathbb{N}^*$, on a

$$\|U^n\|_2 \leqslant C\|f\|_2.$$

Proposition 1.12 (Condition nécessaire et suffisante pour la Cauchy-stabilité)**.** *Un schéma de la forme* (1.5) *est Cauchy-stable si et seulement si son symbole vérifie*

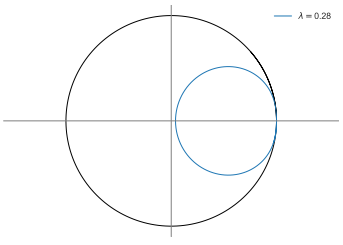
$$\forall \xi \in \mathbb{R}, \quad |\gamma(\xi)| \leqslant 1.$$

L'ingrédient principal de la preuve de ce résultat est l'égalité de Parseval (1.11) et l'utilisation de l'expression (1.10).

Proposition 1.13. *Le schéma upwind* (Upw) *est Cauchy-stable pour* $\lambda \in]0, 1]$.

La condition CFL du schéma upwind est ainsi l'intervalle $]0, 1]$. Par abus de langage, on s'autorise à dire que la condition CFL du schéma upwind est 1.

Démonstration. Le symbole du schéma upwind est $\gamma(\xi) = \lambda e^{-i\xi} + 1 - \lambda$. Ainsi, en calculant le carré de son module, on obtient $|\gamma(\xi)|^2 = 1 - 2\lambda + 2\lambda^2 + 2(1-\lambda)\lambda\cos(\xi)$. Comme $\lambda > 0$ par définition de λ, en étudiant la fonction $\xi \mapsto |\gamma(\xi)|^2$, on trouve que pour avoir $|\gamma(\xi)|^2 \leqslant 1$, il faut avoir $\lambda \leqslant 1$. □



On peut observer la propriété de Cauchy-stabilité graphiquement en traçant la courbe du symbole.

Définition 1.14 (Courbe du symbole)**.** La *courbe* Γ *du symbole* γ est la courbe complexe fermée paramétrée suivante :

$$\Gamma = \{\xi \in [0, 2\pi] \mapsto \gamma(\xi)\}. \tag{1.12}$$

On peut alors visualiser la propriété de Cauchy-stabilité en vérifiant la propriété $\Gamma \subset \overline{\mathbb{D}}$ où $\mathbb{D} \overset{\text{def}}{=} \{z \in \mathbb{C}, |z| < 1\}$ et donc $\overline{\mathbb{D}} \overset{\text{def}}{=} \{z \in \mathbb{C}, |z| \leqslant 1\}$. Dans la Figure 1.4, on représente la courbe Γ du symbole pour le schéma de Beam-Warming (BW) pour différents λ.

Proposition 1.15. *Le schéma de Beam-Warming* (BW) *est Cauchy-stable pour exactement* $\lambda \in]0, 2]$.

La condition CFL du schéma de Beam-Warming est ainsi l'intervalle $]0, 2]$ et par abus de langage, on dira que la condition CFL du schéma de Beam-Warming est 2.

Démonstration. On calcule le symbole pour tout $\xi \in \mathbb{R}$,

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

En prenant son module, on obtient

$$\begin{aligned}
|\gamma(\xi)|^2 &= (\lambda(\lambda-1)\cos\xi + \lambda(2-\lambda) - (\lambda-1)\cos\xi)^2 + (\lambda-1)^2\sin^2\xi \\
&= ((\lambda-1)^2\cos\xi + \lambda(2-\lambda))^2 + (\lambda-1)^2\sin^2\xi \\
&= (\lambda-1)^4\cos^2\xi + \lambda^2(2-\lambda)^2 + 2\lambda(2-\lambda)(\lambda-1)^2\cos\xi + (\lambda-1)^2(1-\cos^2\xi).
\end{aligned}$$

On peut remarquer que

$$(\lambda-1)^2 + \lambda^2(2-\lambda)^2 = 1 + \lambda(\lambda-2)(\lambda-1)^2 \quad \text{et} \quad (\lambda-1)^4 - (\lambda-1)^2 = \lambda(\lambda-2)(\lambda-1)^2.$$

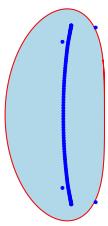
Ainsi, on obtient finalement que

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

Pour que le schéma soit Cauchy-stable, on doit avoir $|\gamma(\xi)|^2 \leqslant 1$. Ainsi, comme $\lambda > 0$, la condition nécessaire et suffisante devient $0 < \lambda \leqslant 2$. □

Ainsi, la Figure 1.4 est bien cohérente avec la Proposition 1.15 puisqu'on observe que la courbe Γ du symbole est toujours dans le disque unité fermé $\overline{\mathbb{D}}$ pour $\lambda \leqslant 2$ et en sort pour $\lambda = 2.1$.

Le schéma O3 (O3) et le schéma LW5 (LW5) sont Cauchy-stables pour $\lambda \in]0, 1]$.



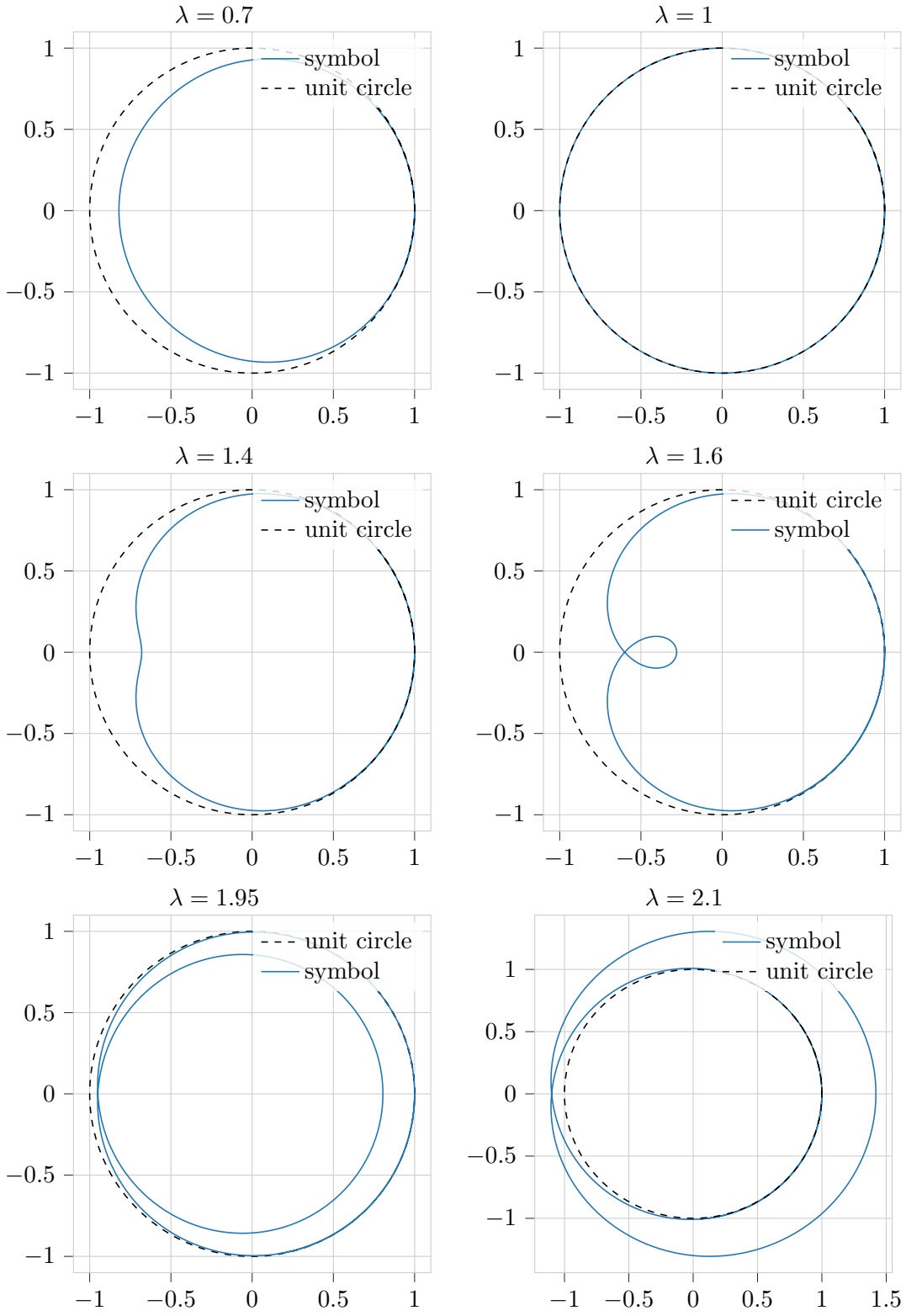
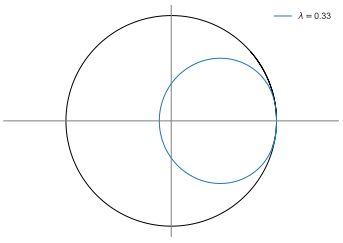


FIGURE 1.4 – Symbole du schéma Beam-Warming pour différentes valeurs de λ.



1.3.2 Stabilité forte

La *stabilité forte* est introduite par Gustafsson, Kreiss et Sundström [GKS72, Def. 3.3] en 1972. L'article [GKS72] propose différentes définitions possibles de stabilité pour le schéma (1.6) et explique les liens entre chacune de ces définitions.

Avant d'introduire la définition de stabilité forte, on donne les différentes normes qui nous seront utiles pour contrôler $(U_j^n)_{j,n}$.

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

On peut maintenant donner la définition de stabilité forte, introduite dans [GKS72] et reprise comme base dans tous les articles qui étudient la stabilité des schémas avec bord posés sur $\mathbb{N}$: [Cou13], [Gus08], [GKO13], etc.

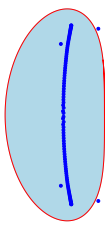
Définition 1.16 (Stabilité forte)**.** Le schéma (1.6) est *fortement stable* s'il existe une constante K telle que pour tout $\alpha > 0$, pour tout $(g_j^n)_{j,n}$, pour tout Δx, pour tout n et pour $\Delta t = \frac{\lambda \Delta x}{a}$, on ait

$$\sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2 + \frac{\alpha}{\alpha \Delta t + 1} \|e^{-\alpha n \Delta t} U\|_{\Delta x,\Delta t}^2 \leqslant K \sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} g_j\|_{\Delta t}^2. \tag{1.13}$$

Dans cette définition, on prend une condition initiale nulle et on veut contrôler la solution $(U_j^n)_{j,n}$ par la condition de bord $(g_j^n)_{j,n}$. Cette expression peut paraître compliquée, mais en comprenant qu'on veut majorer le terme de trace (premier terme du membre de gauche) et les termes intérieurs (second terme du membre de gauche) avec un facteur d'échelle pour éviter que la norme explose et que dans toutes les normes on met un poids $e^{-\alpha n \Delta t}$ afin d'écraser la croissance exponentielle des solutions du problème continu, cette expression est davantage lisible. Cette inégalité est en fait très proche de la version continue du caractère bien posé d'un problème IBVP (*Initial Boundary Value Problem*) sur le demi-espace, voir le livre de Benzoni-Gavage et Serre [BGS06]. Le lien entre l'inégalité de stabilité du cas discret et le caractère bien posé du cas continu est expliqué par Coulombel dans [Cou13]. De plus, dans l'inégalité de stabilité forte présentée dans [GKS72] et dans [Cou13], il y a aussi un contrôle par le terme source.

Enfin, pour avoir une inégalité de convergence avec la condition initiale et la condition de bord, on peut lire les travaux de Wu [Wu95] et de Coulombel [Cou13] qui passe des conditions initiales nulles aux conditions initiales non nulles. Dans la majeure partie de ce manuscrit, on cherche à obtenir la stabilité au sens de la Définition 1.16 avec condition initiale nulle.

Pour la définition de la stabilité forte pour le problème (1.7), il suffit de prendre les indices $j \in [\![J+1 : J+p]\!]$ et non $j \in [\![-r : -1]\!]$ et de changer la somme $\sum_{j=-r}^{+\infty}$ en $\sum_{j=-\infty}^{J+p}$ dans la définition de $\|\cdot\|_{\Delta t,\Delta x}$.



1.3.3 Stabilité GKS

Maintenant que les notions de Cauchy-stabilité et de stabilité forte sont définies, on peut donner la définition de stabilité GKS du schéma posé sur $[\![0 : J]\!]$.

Définition 1.17 (Stabilité GKS)**.** Un schéma de la forme (1.2) est dit GKS-stable si :

- le schéma associé (1.5) est Cauchy-stable,
- le schéma associé (1.6) est fortement stable et
- le schéma associé (1.7) est fortement stable.

Dans toute la suite du manuscrit, on fait l'hypothèse que les schémas qu'on utilise sont consistants et sont Cauchy-stables. Il ne manque donc plus que la stabilité forte pour pouvoir conclure sur la convergence du schéma grâce au Théorème 1.4 (Lax).

Dans la suite de ce chapitre, on présente une première stratégie d'étude de la stabilité forte et une première version du théorème de Kreiss qui donne des conditions nécessaires et suffisantes pour la stabilité forte.

1.4 Étude des modes propres

L'*analyse de Von Neumann* ou *étude des modes propres* consiste à injecter une solution de la forme $U_j^n = z^n \phi_j$ dans le schéma (1.6) et d'étudier la stabilité de la solution en fonction de la valeur complexe de z. Le vecteur $(\phi_j)_j$ est alors vu comme un vecteur propre du schéma et z comme une valeur propre du schéma (voir Définition 1.18).

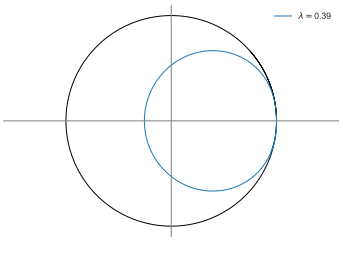
1.4.1 Condition de Godunov–Ryabenkii

Dans l'article [Tre84], Trefethen étudie les différents modes d'instabilité d'un schéma en fonction de z. Il mentionne notamment la condition de Godunov–Ryabenkii, introduite dans l'article de Godunov et Ryabenkii [GR63], dont on donne un énoncé en Proposition 1.20. Cette condition est une condition nécessaire à la stabilité du schéma mais n'est pas suffisante comme on peut le voir au Théorème 1.22.

Tout d'abord, on introduit la notion de *valeur propre* et *vecteur propre* d'un schéma. Pour cela, on va noter $\mathcal{U}$ l'ensemble des complexes de module strictement supérieur à 1, autrement dit $\mathcal{U} \overset{\text{def}}{=} \{z \in \mathbb{C}, |z| > 1\}$.

Définition 1.18 (Valeur propre)**.** On dit que $z \in \overline{\mathcal{U}}$ est une *valeur propre* du schéma (1.6a)-(1.6b) s'il existe un *vecteur propre* $(\phi_j)_j \in \ell^2(\mathbb{N})$ non nul tel que $U_j^n = z^n \phi_j$ est solution de (1.6a)-(1.6b) pour $(g_j^n)_{j,n} = 0$ et $(U_j^0)_j = 0$.

Remarque 1.19. Dans le Chapitre 2, on verra que la terminologie « valeur propre » est bien choisie car on peut voir z comme une valeur propre (au sens usuel du terme), voir Remarque 2.11 (page 50).



Proposition 1.20 (Condition de Godunov–Ryabenkii)**.** *Si le schéma* (1.6a)-(1.6b) *est fortement stable, alors il ne possède pas de valeur propre z telle que $|z| > 1$.*

Quand le schéma (1.6a)-(1.6b) *ne possède pas de valeur propre de module strictement plus grand que* 1*, on dit que la condition de Godunov–Ryabenkii est satisfaite.*

Une preuve détaillée de ce résultat peut être trouvée dans [Cou13] et dans [GKS72]. Informellement, si le schéma possède une valeur propre $|z| > 1$ alors on ne pourra pas contrôler en norme l'itération en temps car $(|z|^n)_n$ ne sera pas bornée.

Comme on le discutera dans le Chapitre 5, notamment à la Section 5.2.4 (page 115), la notion de valeur propre fluctue entre les différents articles selon notamment qu'elle comprend ou non le cas $|z| = 1$. Par exemple, l'article de Wu [Wu95] n'autorise pas les valeurs propres à être sur le cercle unité contrairement à la Définition 1.18.

1.4.2 Première version du théorème de Kreiss

Le théorème de Kreiss, introduit dans l'article de Kreiss [Kre68] pour les schémas dissipatifs et dans l'article de Gustafsson, Kreiss et Sundström [GKS72] pour le cas général, donne des conditions nécessaires et suffisantes pour assurer la stabilité forte d'un schéma. Comme la condition de Godunov–Ryabenkii n'est pas suffisante, on introduit la notion de *valeur propre généralisée*. Cette notion est couramment utilisée pour traiter la stabilité des problèmes avec bord, notamment dans les travaux de Gustafsson, Kreiss et Sundström [GKS72] et repris dans les ouvrages [Gus08] et [GKO13] ainsi que dans tous les articles qui s'appuient sur la théorie développée par Gustafsson, Kreiss et Sundström, comme par exemple [Wu95], [Cou13], [VS15], etc. Le Chapitre 3 est dédié à l'étude de cette théorie.

Définition 1.21 (Valeur propre généralisée)**.** On dit que $z_0 \in \mathbb{S}$ est une *valeur propre généralisée* du schéma (1.6a)-(1.6b) s'il existe un *vecteur propre généralisé* $(\phi_j(z_0))_j \notin \ell^2(\mathbb{N})$ non nul tel que

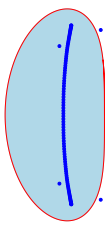
$$\phi_j(z_0) = \lim_{\substack{z \to z_0 \\ |z|>1}} \phi_j(z)$$

où $(\phi_j(z))_j \in \ell^2(\mathbb{N})$ pour $|z| > 1$ et $U_j^n = z^n \phi_j(z)$ est solution de (1.6a).

Dans le Chapitre 5, notamment à la Section 5.2.4 (page 115), on donne une catégorisation des valeurs propres et valeurs propres généralisées afin de faciliter la discussion autour de la stabilité.

Dans le théorème suivant, on donne une première condition nécessaire et suffisante de la stabilité forte. On en donnera une deuxième dans le Théorème 3.23 (page 89) quand on aura introduit la notion de *condition de Kreiss–Lopatinskii uniforme* (voir Définition 3.19, page 84).

Théorème 1.22 (Kreiss 1)**.** *Le schéma* (1.6) *est fortement stable si, et seulement si, il ne possède ni valeur propre, ni valeur propre généralisée.*



On cherche donc à étudier les propriétés spectrales des schémas afin de pouvoir conclure sur la stabilité, c'est l'objet du chapitre suivant (Chapitre 2).

1.5 Compléments

Pour ne pas alourdir le début de la discussion autour du schéma en Section 1.1, on se permet ici des précisions concernant la définition du schéma et on donne une deuxième définition de schéma équivalente à la première et qui nous sert dans la suite du manuscrit. De plus, on définit les méthodes de Lax-Wendrof inverse et de Lax-Wendroff inverse simplifiée qui déterminent des conditions de bord consistantes à un ordre d donné, on en fait d'ailleurs une étude rigoureuse de consistance.

1.5.1 Formulation alternative du schéma

Le schéma (1.6), posé sur $\mathbb{N}$, peut être visualisé de la manière suivante.

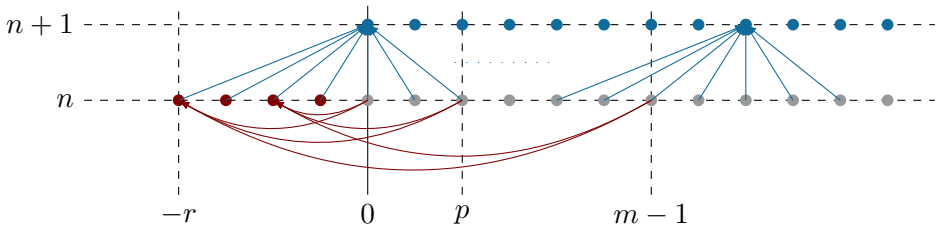


FIGURE 1.5 – Discrétisation du schéma (1.6).

Les points gris de la Figure 1.5 correspondent aux points calculés à l'étape n. Ensuite pour définir les points bleus de l'étape $n+1$, il faut définir les points fantômes (points rouges) en extrapolant les données des m premiers points.

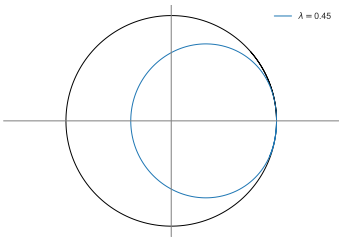
En posant les matrices

$$B = \begin{pmatrix} b_{-r,0} & \cdots & \cdots & b_{-r,m-1} \\ \vdots & & & \vdots \\ b_{-1,0} & \cdots & \cdots & b_{-1,m-1} \end{pmatrix} \in \mathcal{M}_{r,m}(\mathbb{C}) \quad \text{et} \quad G^n = \begin{pmatrix} g_{-r}^n \\ \vdots \\ g_{-1}^n \end{pmatrix}, \tag{1.14}$$

les équations de bords (1.6b) peuvent se réécrire matriciellement de la manière suivante

$$\begin{pmatrix} U_{-r}^n \\ \vdots \\ U_{-1}^n \end{pmatrix} = B \begin{pmatrix} U_0^n \\ \vdots \\ U_{m-1}^n \end{pmatrix} + G^n.$$

Cette vision est particulièrement utilisée dans le Chapitre 5, notamment à travers la définition du déterminant de Kreiss–Lopatinskii à la Définition 5.11 (page 113).



On peut vouloir intégrer les conditions de bord directement dans le schéma sans créer de points fantômes à l'extérieur du domaine, dans ce cas, les r premiers points du domaine seront affectés. Cela donne la définition de schéma suivante :

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{p} a_k U_{k+j}^n, & j \geqslant r,\ n \geqslant 0, \quad (1.15a) \\ U_j^{n+1} = \displaystyle\sum_{k=0}^{m-1} \mathcal{B}_{j,k} U_k^n + \mathcal{G}_j^n, & j \in [\![0 : r-1]\!],\ n \geqslant 0, \quad (1.15b) \\ U_j^0 = f_j, & j \geqslant 0. \quad (1.15c) \end{cases}$$

où $(\mathcal{B}_{j,k})_{j,k}$ sont des constantes réelles et $(\mathcal{G}_j^n)_{j,n}$ sont des données de bord numériques qui sont fortement liées aux constantes $(b_{j,k})_{j,k}$ et $(g_j^n)_{j,n}$ du schéma (1.6), comme on va le voir à l'équation (1.17). On peut visualiser la définition du schéma (1.15) dans la figure suivante :

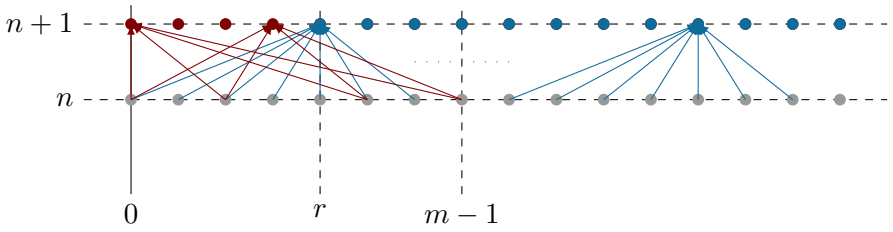


FIGURE 1.6 – Discrétisation du schéma (1.15).

La Figure 1.6 est à mettre en relation avec la Figure 1.5. De la même manière que pour la Figure 1.5, les points gris représentent les points calculés à l'étape n. Ensuite, pour définir les points à l'étape $n+1$, on peut le faire avec l'équation intérieure pour les valeurs $j \geqslant r$ (points bleus) et on définit les r premiers points (points rouges) en utilisant les données des m premiers points au temps n.

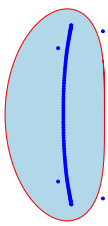
En posant les matrices

$$\mathcal{B} = \begin{pmatrix} \mathcal{B}_{0,0} & \cdots & \cdots & \mathcal{B}_{0,m-1} \\ \vdots & & & \vdots \\ \mathcal{B}_{r-1,0} & \cdots & \cdots & \mathcal{B}_{r-1,m-1} \end{pmatrix} \in \mathcal{M}_{r,m}(\mathbb{C}) \quad \text{et} \quad \mathcal{G}^n = \begin{pmatrix} \mathcal{G}_0^n \\ \vdots \\ \mathcal{G}_{r-1}^n \end{pmatrix},$$

les équations de bords (1.15b) peuvent se réécrire matriciellement de la manière suivante

$$\begin{pmatrix} U_0^{n+1} \\ \vdots \\ U_{r-1}^{n+1} \end{pmatrix} = \mathcal{B} \begin{pmatrix} U_0^n \\ \vdots \\ U_{m-1}^n \end{pmatrix} + \mathcal{G}^n. \quad (1.16)$$

Cette vision est particulièrement utilisée dans le Chapitre 2 où la matrice $\mathcal{B}$ intervient dans la définition de matrice Quasi-Toeplitz (voir Définition 2.8, page 48). Elle est aussi utilisée dans



le Chapitre 6, notamment à travers la définition du déterminant de Kreiss–Lopatinskii à la Définition 6.10 (page 151), cette vision du déterminant de Kreiss–Lopatinskii est différente de l'approche faite dans le Chapitre 5, on discute de ces deux visions dans la Section 3.3.2 (page 80).

Soulignons que les deux schémas (1.6) et (1.15) sont équivalents au sens où tout schéma écrit sous la forme (1.6) peut se réécrire sous la forme (1.15) et vice-versa. En effet, on peut faire le lien explicite entre les deux définitions de la manière suivante :

$$\mathcal{B} = \begin{pmatrix} a_{-r} & \cdots & a_{-1} \\ & \ddots & \vdots \\ 0 & & a_{-r} \end{pmatrix} B + \begin{pmatrix} a_0 & \cdots & a_p & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & & \ddots & \ddots & & & \vdots \\ a_{-r+1} & \cdots & a_0 & \cdots & a_p & 0 & \cdots & 0 \end{pmatrix} \in \mathcal{M}_{r,m}(\mathbb{C}) \tag{1.17}$$

$$\text{et} \quad \mathcal{G}^n = \begin{pmatrix} a_{-r} & \cdots & a_{-1} \\ & \ddots & \vdots \\ 0 & & a_{-r} \end{pmatrix} G^n.$$

Comme le coefficient a_{-r} est non nul, les matrices carrées sont inversibles et donc on peut exprimer B et G^n en fonction $\mathcal{B}$ et $\mathcal{G}^n$.

1.5.2 Méthode Lax-Wendroff inverse et Lax-Wendroff inverse simplifiée

Dans cette section, on présente deux façons de définir les conditions de bords numériques : la méthode Lax-Wendroff inverse (ILW) et la méthode Lax-Wendroff inverse simplifiée (SILW).

On veut créer des conditions de bord qui soient consistantes à un certain ordre et qui utilisent la donnée de bord g, pour cela on utilise la méthode *Lax-Wendroff inverse*, introduite par Tan et Shu [TS10] en 2010 et reprise dans plusieurs articles : [TWSN12], [TS13], [VS15], [LSZ16], [LSZ17], etc.

L'idée de cette méthode est d'utiliser l'équation aux dérivées partielles pour transformer les dérivées spatiales en dérivées temporelles afin de pouvoir utiliser la donnée de bord et ses dérivées, on suppose pour l'instant que la donnée de bord g est aussi régulière que l'on veut.

Pour un ordre de consistance d donné, pour $j \in [\![-r : -1]\!]$, on définit

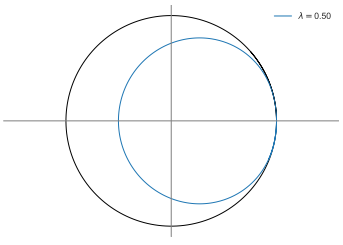
$$U_j^n = u(n\Delta t, 0) + j\Delta x \frac{\partial u}{\partial x}(n\Delta t, 0) + \frac{(j\Delta x)^2}{2} \frac{\partial^2 u}{\partial x^2}(n\Delta t, 0) + \cdots + \frac{(j\Delta x)^{d-1}}{(d-1)!} \frac{\partial^{d-1} u}{\partial x^{d-1}}(n\Delta t, 0). \tag{1.18}$$

L'équation (1.18) est bien consistante d'ordre d puisque, par développement de Taylor,

$$\forall j \in [\![-r : -1]\!], \quad u(n\Delta t, j\Delta x) - \sum_{k=0}^{d-1} \frac{(j\Delta x)^k}{k!} \frac{\partial^k u}{\partial x^k}(n\Delta t, 0) = O(\Delta x^d).$$

En dérivant, par rapport au temps et à l'espace, l'équation de transport (1.1), on obtient

$$\forall k \in \mathbb{N}^*, \quad \frac{\partial^k u}{\partial x^k} = \frac{(-1)^k}{a^k} \frac{\partial^k u}{\partial t^k}. \tag{1.19}$$



En injectant (1.19) dans (1.18) et en utilisant la condition de bord g, on obtient la méthode Lax-Wendroff inverse d'ordre d : pour tout $j \in [\![-r : -1]\!]$,

$$U_j^n = g(n\Delta t) - \frac{j\Delta x}{a} g'(n\Delta t) + \frac{(j\Delta x)^2}{2a^2} g^{(2)}(n\Delta t) + \cdots + \frac{(-j\Delta x)^{d-1}}{a^{d-1}(d-1)!} g^{(d-1)}(n\Delta t). \qquad (\mathrm{ILW}_d)$$

La condition de bord Lax-Wendroff inverse d'ordre d sera notée ILW_d ou ILWd. De plus, cette condition est bien de la forme de (1.6b). En effet, en utilisant les notations de (1.6b), on a, pour tout entier $j \in [\![-r : -1]\!]$,

$$\forall k \in [\![0 : m-1]\!], \quad b_{j,k} = 0 \quad \text{et} \quad \forall n \in \mathbb{N}, \quad g_j^n = \sum_{k=0}^{d-1} \frac{(-j\Delta x)^k}{a^k k!} g^{(k)}(n\Delta t).$$

Cependant, on n'a pas toujours accès à toutes les dérivées de la donnée de bord g, notamment lorsque la donnée de bord g est issue d'une simulation numérique, on n'a accès qu'à des valeurs ponctuelles de la fonction g. Il est alors intéressant de coupler la méthode Lax-Wendroff inverse avec de l'extrapolation des valeurs de (U_j^n) au bord du domaine. La méthode Lax-Wendroff inverse simplifiée, introduite par Vilar et Shu [VS15] en 2015, permet d'utiliser cette idée tout en conservant un ordre de consistance d donné. Pour cela, on tronque l'équation (ILW_d) à un indice k_d et pour les valeurs $k \in [\![k_d : d-1]\!]$, on approche la dérivée spatiale d'ordre k par extrapolation. La méthode Lax-Wendroff inverse simplifiée est alors définie de la manière suivante :

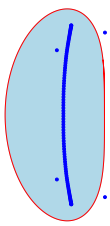
$$\forall j \in [\![-r : -1]\!], \quad U_j^n = \sum_{k=0}^{k_d-1} \frac{(-j\Delta x)^k}{k!} \frac{g^{(k)}(n\Delta t)}{a^k} + \sum_{k=k_d}^{d-1} j^k \sum_{s=0}^{d-1} p_{k,s}^{(d)} U_s^n. \qquad (\mathrm{S}_{k_d}\mathrm{ILW}_d)$$

où $\displaystyle\sum_{s=0}^{d-1} p_{k,s}^{(d)} U_{s^n}$ approche $\dfrac{\Delta x^k}{k!} \dfrac{\partial^k u}{\partial x^k}(n\Delta t, 0)$ à l'ordre d, ce qui justifie la consistance de la méthode Lax-Wendroff inverse simplifiée.

Pour trouver les coefficients $(p_{k,s}^{(d)})$, on injecte la solution exacte et on regarde les conditions d'annulation.

$$\begin{aligned}
\sum_{s=0}^{d-1} p_{k,s}^{(d)} u(n\Delta t, s\Delta x) &= \sum_{s=0}^{d-1} p_{k,s}^{(d)} \sum_{\ell=0}^{d-1} \frac{(s\Delta x)^\ell}{\ell!} \frac{\partial^\ell u}{\partial x^\ell}(n\Delta t, 0) + O(\Delta x^d) \\
&= \sum_{\ell=0}^{d-1} \underbrace{\left(\sum_{s=0}^{d-1} p_{k,s}^{(d)} s^\ell \right)}_{\delta(k-\ell)} \frac{(\Delta x)^\ell}{\ell!} \frac{\partial^\ell u}{\partial x^\ell}(n\Delta t, 0) + O(\Delta x^d)
\end{aligned}$$

où $\delta(n)$ est le symbole de Kronecker en 0.



On cherche donc à résoudre les d équations suivantes :

$$\forall \ell \in [\![0 : d-1]\!], \quad \sum_{s=0}^{d-1} p_{k,s}^{(d)} s^\ell = \delta(k-\ell). \tag{1.20}$$

On considère le polynôme $P_{(k,d)}(X) \overset{\text{def}}{=} \sum_{s=0}^{d-1} p_{k,s}^{(d)} X^s$. En utilisant l'Annexe E et notamment la Proposition E.2 (page 218), on peut réécrire (1.20) de la manière suivante :

$$\underbrace{\begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \binom{1}{1} & \ddots & & \vdots \\ \vdots & \binom{2}{1} & \binom{2}{2} & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & \binom{d-1}{1} & \binom{d-1}{2} & \cdots & \binom{d-1}{d-1} \end{pmatrix}}_{P_d} \begin{pmatrix} P_{(k,d)}(1) \\ P'_{(k,d)}(1) \\ P''_{(k,d)}(1) \\ \vdots \\ P^{(d-1)}_{(k,d)}(1) \end{pmatrix} = e_k$$

où e_k représente le k-ième vecteur de la base canonique de $\mathbb{R}^d$. Ce qui revient à la résolution du système suivant :

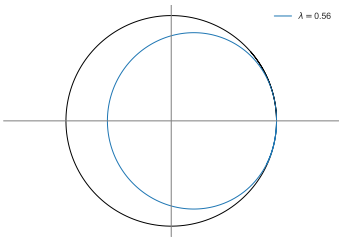
$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \binom{1}{1} & \ddots & & \vdots \\ \vdots & \binom{2}{1} & \binom{2}{2} & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & \binom{d-1}{1} & \binom{d-1}{2} & \cdots & \binom{d-1}{d-1} \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & \cdots & \cdots & \cdots & 1 \\ 0 & 1 & 2 & \cdots & \cdots & \cdots & d-1 \\ \vdots & \ddots & 2 & 6 & 12 & & \\ \vdots & & \ddots & 6 & & & \\ \vdots & & & \ddots & 24 & & \\ \vdots & & & & \ddots & \ddots & \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & (d-1)! \end{pmatrix} \begin{pmatrix} p_{k,0}^{(d)} \\ p_{k,1}^{(d)} \\ p_{k,2}^{(d)} \\ \vdots \\ p_{k,d-1}^{(d)} \end{pmatrix} = e_k \tag{1.21}$$

où le coefficient $(c_{i,j})_{i,j}$ de la matrice centrale est le coefficient devant le monôme X^j dérivé i fois.

On peut donc résoudre ces deux systèmes triangulaires afin de trouver les coefficients $p_{k,s}^{(d)}$ et définir explicitement la méthode de Lax-Wendroff inverse simplifiée, c'est ce qui est fait dans l'implémentation numérique que l'on détaille à la Section 7.3 (page 188). La condition de bord Lax-Wendroff inverse simplifiée d'ordre d et d'indice de troncature k_d sera notée $\mathrm{S}_{k_d}\mathrm{ILW}_d$ ou Sk_dILWd.

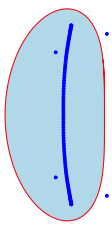
Dans toute la suite de ce manuscrit, on utilise régulièrement ce type de condition de bord, notamment l'exemple des conditions $\mathrm{S}_2\mathrm{ILW}_3$ couplées avec le schéma de Beam-Warming.

Exemple 1.23. Le schéma de Beam-Warming (BW) muni de la condition de bord $\mathrm{S}_2\mathrm{ILW}_3$ s'écrit



alors de la manière suivante :

$$\begin{cases} U_j^{n+1} = \dfrac{\lambda(\lambda-1)}{2}U_{j-2}^n + \lambda(2-\lambda)U_{j-1}^n + \dfrac{(\lambda-1)(\lambda-2)}{2}U_j^n, & j \geqslant 0, n \geqslant 0, \\ U_{-1}^n = g(n\Delta t) + \dfrac{\Delta x g'(n\Delta t)}{a} + \dfrac{1}{2}(U_2^n - 2U_1^n + U_0^n), & n \geqslant 0, \\ U_{-2}^n = g(n\Delta t) + \dfrac{2\Delta x g'(n\Delta t)}{a} + 2(U_2^n - 2U_1^n + U_0^n), & n \geqslant 0, \\ U_j^0 = 0, & j \geqslant 0. \end{cases} \tag{1.22}$$

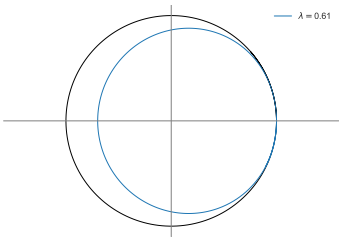


Chapitre 2

Analyse spectrale de matrices Toeplitz

Le but de ce chapitre est de décrire le lien entre les valeurs propres généralisées d'un schéma (voir Définition 1.21) et les aspects géométriques du pseudospectre de la représentation de ce schéma. Ce lien reste encore à préciser mais nous tentons ici d'établir des correspondances entre ces deux points de vue. De plus, le livre de Trefethen et Embree [TE05] de 2005 encourage cette recherche :

> « In particular, there is no tradition of looking for GKS instabilities by plotting pseudospectra. Part of the reason for this situation is undoubtedly that the routine computation of pseudospectra is a more recent development than GKS-stability theory, popular among a younger generation of researchers. It will be interesting to



> see if the use of pseudospectra for analyzing boundary conditions catches on in the future. »
>
> [TE05, Chap.34]

La suite du manuscrit peut être lue indépendamment de ce chapitre, mais celui-ci apporte un point de vue différent qui peut être mis en parallèle avec la suite. En effet, dans les chapitres suivants, on change d'approche en étudiant le déterminant de Kreiss–Lopatinskii et en utilisant l'analyse complexe afin d'identifier les GKS-instabilités.

Dans ce chapitre, nous allons résumer certains résultats déjà connus, les observations et recherches que nous avons faites sur le sujet et les questions qui restent encore ouvertes. On commence par faire le lien entre les matrices et opérateurs Toeplitz, introduits par Otto Toeplitz [Toe11] et les schémas numériques dont on a parlé au chapitre précédent. Pour plus de détails sur les matrices et opérateurs Toeplitz, le livre de Nikolski [Nik17] en fait une étude d'un point de vue fonctionnel, on va se servir de certains résultats de ce livre dans la suite.

2.1 Lien entre schémas numériques et opérateurs Toeplitz

On rappelle ici les notions de matrices Toeplitz et ses variantes qui servent en algèbre linéaire mais aussi en analyse numérique pour représenter des schémas comme on va le voir par la suite.

On se fixe deux entiers naturels $p \in \mathbb{N}$ et $r \in \mathbb{N}$, ainsi que $r+p+1$ nombres complexes : $(a_k)_{k=-r}^{p} \in \mathbb{C}^{p+r+1}$ avec $a_{-r} \neq 0$ et $a_p \neq 0$.

2.1.1 Opérateur Toeplitz sur $\mathbb{Z}$

Définition 2.1 (Opérateur Toeplitz sur $\mathbb{Z}$)**.** On définit l'*opérateur Toeplitz sur* $\mathbb{Z}$, associé aux $(a_k)_{k=-r}^{p}$, de la manière suivante :

$$T_{\mathbb{Z}} : \left\{ \begin{array}{ccc} \ell^2(\mathbb{Z}) & \to & \ell^2(\mathbb{Z}) \\ u = (u_j)_{j\in\mathbb{Z}} & \mapsto & ((T_{\mathbb{Z}}u)_j)_{j\in\mathbb{Z}} \overset{\text{def}}{=} \left(\sum_{k=-r}^{p} a_k u_{j+k} \right)_{j\in\mathbb{Z}} \end{array} \right. . \tag{2.1}$$

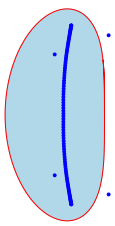
On note $\Lambda(T_{\mathbb{Z}})$ le spectre de l'opérateur $T_{\mathbb{Z}}$.

Remarque 2.2. Dans le livre [Nik17], les opérateurs Toeplitz sur $\mathbb{Z}$ sont appelés « opérateurs de Laurent », il appelle « opérateurs Toeplitz » les opérateurs Toeplitz sur $\mathbb{N}$ que l'on va voir dans la suite.

L'opérateur Toeplitz sur $\mathbb{Z}$ défini en (2.1) correspond au schéma numérique (1.5) posé sur $\mathbb{Z}$ que l'on rappelle ici

$$U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{j+k}^n, \quad n \in \mathbb{N}, j \in \mathbb{Z}$$

au sens où $U^{n+1} = T_{\mathbb{Z}} U^n$ pour tout $n \in \mathbb{N}$ en notant $U^n = (U_j^n)_{j\in\mathbb{Z}}$.



Définition 2.3 (Matrice Toeplitz circulante)**.** Soit $J \in \mathbb{N}$, on appelle *matrice Toeplitz circulante*, associée aux $(a_k)_{k=-r}^{p}$, la matrice suivante :

$$T_J^\circ \overset{\text{def}}{=} \begin{pmatrix} a_0 & \cdots & a_p & 0 & \cdots & 0 & a_{-r} & \cdots & a_{-1} \\ \vdots & a_0 & & \ddots & \ddots & & \ddots & \ddots & \vdots \\ a_{-r} & & \ddots & & \ddots & \ddots & & \ddots & a_{-r} \\ 0 & \ddots & & \ddots & & \ddots & \ddots & & 0 \\ \vdots & \ddots & \ddots & & \ddots & & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & & \ddots & & \ddots & 0 \\ a_p & \ddots & & \ddots & \ddots & & \ddots & & a_r \\ \vdots & \ddots & \ddots & & \ddots & \ddots & & a_0 & \vdots \\ a_1 & \cdots & a_p & 0 & \cdots & 0 & a_{-r} & \cdots & a_0 \end{pmatrix} \in \mathcal{M}_{J+1}(\mathbb{R}).$$

On note $\Lambda(T_J^\circ)$ le spectre de la matrice T_J°.

L'avantage de la définition des matrices Toeplitz circulantes est d'avoir un outil de dimension finie qui approche l'opérateur Toeplitz sur $\mathbb{Z}$, comme on peut le voir dans la proposition suivante.

Proposition 2.4. *On a*

$$\Lambda(T_{\mathbb{Z}}) = \left\{ \lim_{j\to\infty} \mu_j \text{ pour } \mu_j \in \Lambda(T_{J_j}^\circ) \text{ avec } \lim_{j\to\infty} J_j = +\infty \right\}.$$

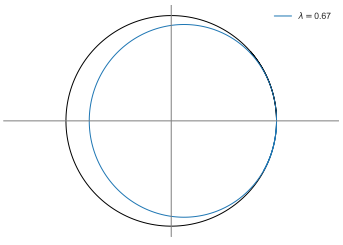
Démonstration. **Étape 1 :** *Calcul du spectre $\Lambda(T_J^\circ)$ de la matrice Toeplitz circulante T_J°.*
On veut trouver le polynôme caractéristique de T_J° donc on calcule le déterminant de la matrice :

$$XI_{J+1} - T_J^\circ = \begin{pmatrix} X-a_0 & \cdots & -a_p & 0 & -a_{-r} & \cdots & -a_{-1} \\ \vdots & X-a_0 & & & \ddots & & \vdots \\ -a_{-r} & & \ddots & & & \ddots & -a_{-r} \\ 0 & & & \ddots & & & 0 \\ -a_p & \ddots & & & \ddots & & -a_p \\ \vdots & & \ddots & & & X-a_0 & \vdots \\ -a_1 & \cdots & -a_p & 0 & -a_{-r} & \cdots & X-a_0 \end{pmatrix}.$$

C'est le déterminant d'une matrice circulante, qui s'obtient comme

$$\prod_{k=0}^{J} P(\omega^k)$$

où $\omega = \mathrm{e}^{\frac{i2\pi}{J+1}}$ et $P(t) = X - a_0 - a_1 t - a_2 t^2 - \cdots - a_p t^p - a_{-r} t^{J+1-r} - \cdots - a_{-1} t^J$.



Ainsi $\chi_{T_J^\circ}(X) = \prod_{k=0}^{J} \left(X - \sum_{j=-r}^{p} a_j e^{\frac{i2\pi jk}{J+1}} \right)$. Donc les valeurs propres de T_J° sont

$$\Lambda(T_J^\circ) = \left\{ \sum_{j=-r}^{p} a_j e^{\frac{i2\pi jk}{J+1}}, k \in [\![0 : J]\!] \right\}. \tag{2.2}$$

Étape 2 : *Calcul du spectre $\Lambda(T_\mathbb{Z})$ de l'opérateur $T_\mathbb{Z}$ sur $\mathbb{Z}$.*

On utilise la transformée de Fourier sur $\ell^2(\mathbb{Z})$.

$$\forall u \in \ell^2(\mathbb{Z}), \forall \xi \in \mathbb{R}, \quad \hat{u}(\xi) = \sum_{j \in \mathbb{Z}} u_j e^{ij\xi}.$$

On trouve alors, pour tout $\xi \in \mathbb{R}$,

$$\begin{aligned} \widehat{T_\mathbb{Z} u}(\xi) &= \sum_{j \in \mathbb{Z}} \sum_{k=-r}^{p} a_k u_{j+k} e^{ij\xi} \\ &= \sum_{k=-r}^{p} a_k e^{-ik\xi} \sum_{j \in \mathbb{Z}} u_j e^{ij\xi} \\ &= \sum_{k=-r}^{p} a_k e^{-ik\xi} \hat{u}(\xi). \end{aligned}$$

Donc pour $\mu = \sum_{k=-r}^{p} a_k e^{-ik\xi}$, $T_\mathbb{Z} - \mu$ n'est pas injectif. Ainsi, $\sum_{k=-r}^{p} a_k e^{-ik\xi}$ est dans $\Lambda(T_\mathbb{Z})$. De plus, pour tout $\mu \neq \sum_{k=-r}^{p} a_k e^{-ik\xi}$, par transformée de Fourier inverse, $T_\mathbb{Z} - \mu$ est surjectif. Donc le spectre de $T_\mathbb{Z}$ est exactement

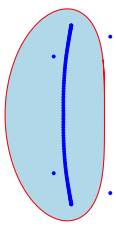
$$\Lambda(T_\mathbb{Z}) = \left\{ \sum_{j=-r}^{p} a_j e^{ij\xi}, \xi \in [0, 2\pi] \right\}. \tag{2.3}$$

On conclut la preuve par double inclusion puisque chaque $\xi \in [0, 2\pi]$ de (2.3) peut être approchée par une suite de $2\pi \frac{k}{J+1}$ pour $k \in [\![0 : J]\!]$ avec J croissant (voir (2.2)). □

Exemple 2.5. À la Figure 2.1, on trace le spectre de la matrice circulante (de taille $J = 30$) et le spectre de l'opérateur Toeplitz sur $\mathbb{Z}$ pour le schéma jouet

$$U_j^{n+1} = -\frac{1}{3} U_{j-1}^n - \frac{1}{2} U_j^n + U_{j+1}^n - \frac{1}{6} U_{j+2}^n. \tag{2.4}$$

représenté par le vecteur $(a_{-1}, a_0, a_1, a_2) = \left(-\frac{1}{3}, -\frac{1}{2}, 1, -\frac{1}{6}\right)$.



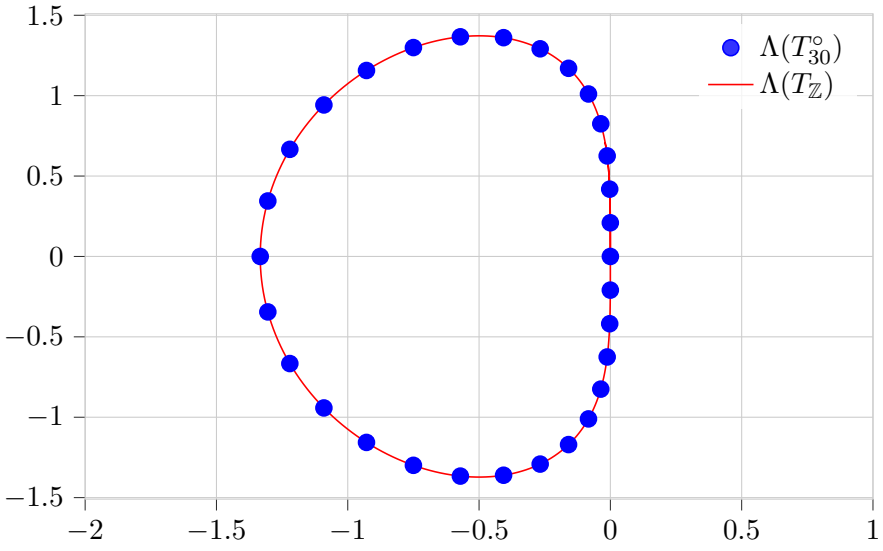


FIGURE 2.1 – Spectres de la matrice Toeplitz circulante T_{30}° et de l'opérateur Toeplitz $T_{\mathbb{Z}}$ sur $\mathbb{Z}$ liés à (2.4).

2.1.2 Matrice Toeplitz

Définition 2.6 (Matrice Toeplitz)**.** Soit $J \in \mathbb{N}$, on appelle *matrice Toeplitz*, associée aux $(a_k)_{k=-r}^{p}$, la matrice suivante :

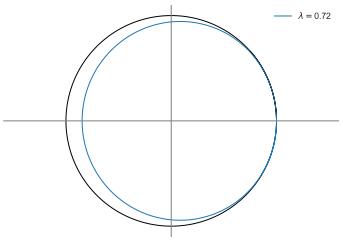
$$T_J \overset{\text{def}}{=} \begin{pmatrix} a_0 & a_1 & \dots & a_p & 0 & \dots & 0 \\ a_{-1} & a_0 & & & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & 0 \\ a_{-r} & & & \ddots & & & a_p \\ 0 & \ddots & & & \ddots & & \vdots \\ \vdots & & \ddots & & & a_0 & a_1 \\ 0 & \dots & 0 & a_{-r} & \dots & a_{-1} & a_0 \end{pmatrix} \in \mathcal{M}_{J+1}(\mathbb{R}). \tag{2.5}$$

On note $\Lambda(T_J)$ le spectre de la matrice T_J.

Une matrice Toeplitz T_J (définie en (2.5)) représente le schéma numérique (1.2a) avec condition de Dirichlet homogène à droite et à gauche (voir (1.3), page 25) que l'on rappelle ici :

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{p} a_k U_{j+k}^n & n \in \mathbb{N}, j \in [\![0 : J]\!], \\ U_j^n = 0 & n \in \mathbb{N}, j \in [\![-r : -1]\!] \cup [\![J+1 : J+p]\!]. \end{cases} \tag{2.6}$$

Le schéma (2.6) peut s'écrire $U^{n+1} = T_J U^n$ pour tout $n \in \mathbb{N}$, avec $U^n = {}^t(U_0^n, \dots, U_J^n)$.



Exemple 2.7. Le *shift* de taille J est défini par la matrice Toeplitz suivante :

$$\mathcal{S}_J \overset{\text{def}}{=} \begin{pmatrix} 0 & & & 0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix} \tag{2.7}$$

et représente le schéma $U_j^{n+1} = U_{j-1}^n$.

Le schéma de Beam-Warming (BW), défini dans l'Exemple 1.1 (page 22), est représenté par la matrice Toeplitz suivante :

$$BW_J \overset{\text{def}}{=} \begin{pmatrix} \frac{(\lambda-1)(\lambda-2)}{2} & 0 & \cdots & \cdots & 0 \\ \lambda(2-\lambda) & \frac{(\lambda-1)(\lambda-2)}{2} & \ddots & & \vdots \\ \frac{\lambda(\lambda-1)}{2} & \lambda(2-\lambda) & \frac{(\lambda-1)(\lambda-2)}{2} & \ddots & \vdots \\ & \ddots & \ddots & \ddots & 0 \\ 0 & & \frac{\lambda(\lambda-1)}{2} & \lambda(2-\lambda) & \frac{(\lambda-1)(\lambda-2)}{2} \end{pmatrix} \tag{2.8}$$

2.1.3 Matrice Quasi-Toeplitz

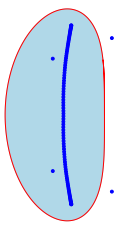
Le but est de pouvoir représenter matriciellement des schémas avec d'autres conditions de bord. C'est pour cela que l'on va introduire la notion de matrice Quasi-Toeplitz, déjà présente dans l'article de Beam et Warming [BW93].

Définition 2.8 (Matrice Quasi-Toeplitz)**.** Soit $J \in \mathbb{N}$, $r + p \leqslant m \leqslant J + 1$ et deux matrices $\mathcal{B} \in \mathcal{M}_{r,m}(\mathbb{C})$ et $\mathcal{C} \in \mathcal{M}_{p,m}(\mathbb{C})$. On appelle *matrice Quasi-Toeplitz*, associée aux $(a_k)_{k=-r}^{p}$ et aux matrices $\mathcal{B}$ et $\mathcal{C}$, la matrice de la forme suivante :

$$\widetilde{T_J} \overset{\text{def}}{=} \begin{pmatrix} \multicolumn{5}{c}{\overset{m}{\longleftrightarrow}} \\ \multicolumn{5}{c}{\boxed{\mathcal{B}}} \\ a_{-r} & \cdots & a_0 & \cdots & a_p & 0 \\ & \ddots & & \ddots & & \ddots \\ & & \ddots & & \ddots & & \ddots \\ 0 & & a_{-r} & \cdots & a_0 & \cdots & a_p \\ & & & \multicolumn{3}{c}{\boxed{\mathcal{C}}} \\ & & & \multicolumn{3}{c}{\underset{m}{\longleftrightarrow}} \end{pmatrix} \begin{matrix} \\ \updownarrow r \\ \\ \\ \\ \\ \updownarrow p \\ \\ \end{matrix} \in \mathcal{M}_{J+1}(\mathbb{C}) \tag{2.9}$$

On note $\Lambda(\widetilde{T_J})$ le spectre de la matrice $\widetilde{T_J}$.

Dans cette définition, les matrices $\mathcal{B}$ et $\mathcal{C}$ ont vocation à être fixées tandis que la dimension $J + 1$ de la matrice $\widetilde{T_J}$ va pouvoir varier et même tendre vers l'infini.



La matrice Quasi-Toeplitz définie en (2.9) représente le schéma numérique (2.10).

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{p} a_k U_{j+k}^n & n \in \mathbb{N}, j \in [\![0 : J]\!], \\ U_j^n = \displaystyle\sum_{k=0}^{m-1} b_{j,k} U_k^n & n \in \mathbb{N}, j \in [\![-r : -1]\!], \\ U_j^n = \displaystyle\sum_{k=0}^{m-1} c_{j,k} U_{J-k}^n & n \in \mathbb{N}, j \in [\![J+1 : J+p]\!], \end{cases} \tag{2.10}$$

où $\mathcal{B}$ et $\mathcal{C}$ sont définis par les deux matrices suivantes :

$$\mathcal{B} \overset{\text{def}}{=} \begin{pmatrix} a_{-r} & \cdots & a_{-1} \\ & \ddots & \vdots \\ 0 & & a_{-r} \end{pmatrix} \begin{pmatrix} b_{-r,0} & \cdots & b_{-r,m-1} \\ \vdots & & \vdots \\ b_{-1,0} & \cdots & b_{-1,m-1} \end{pmatrix} + \begin{pmatrix} a_0 & \cdots & a_p & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & & \ddots & \ddots & & & \vdots \\ a_{-r+1} & \cdots & a_0 & \cdots & a_p & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathcal{C} \overset{\text{def}}{=} \begin{pmatrix} a_p & \cdots & a_1 \\ & \ddots & \vdots \\ 0 & & a_p \end{pmatrix} \begin{pmatrix} c_{J+p,m-1} & \cdots & c_{J+p,0} \\ \vdots & & \vdots \\ c_{J+1,m-1} & \cdots & c_{J+1,0} \end{pmatrix} + \begin{pmatrix} 0 & \cdots & 0 & a_{-r} & \cdots & a_0 & \cdots & a_{p-1} \\ \vdots & & & \ddots & \ddots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & a_{-r} & \cdots & a_0 \end{pmatrix}.$$

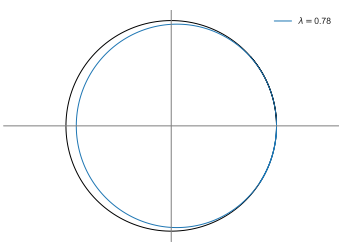
Cette reformulation d'un schéma est à mettre en relation avec la discussion de la Section 1.5.1 (page 37) puisque c'est la même vision que le schéma (1.15a)-(1.15b) pour $(g_j^n)_{j,n} = 0$ mais au lieu d'être défini sur $\mathbb{N}$, ici on est défini sur le segment $[\![0 : J]\!]$ avec une deuxième condition de bord à droite.

Exemple 2.9. Pour poursuivre l'Exemple 2.7, si on rajoute la condition de bord $U_{-1}^n = U_0^n$ à gauche et la condition de bord de Dirichlet homogène à droite au shift $\mathcal{S}_J$ défini en (2.7), on obtient la matrice Quasi-Toeplitz suivante :

$$\widetilde{\mathcal{S}}_J \overset{\text{def}}{=} \begin{pmatrix} 1 & & & 0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix}. \tag{2.11}$$

Si on ajoute la condition de bord $\mathrm{S}_2\mathrm{ILW}_3$ (définie en ($\mathrm{S}_{k_d}\mathrm{ILW}_d$) à la Section 1.5.2, page 39) à gauche et la condition de bord de Dirichlet homogène à droite sur le schéma de Beam-Warming, on obtient la matrice Quasi-Toeplitz suivante :

$$\widetilde{BW}_J \overset{\text{def}}{=} \begin{pmatrix} \lambda^2 - \frac{3\lambda}{2} + 1 & -\lambda^2 & \frac{\lambda^2}{2} & 0 & 0 \\ \frac{7\lambda - 3\lambda^2}{4} & 1 - \lambda & \frac{\lambda(\lambda-1)}{4} & & \vdots \\ \frac{\lambda(\lambda-1)}{2} & \lambda(2-\lambda) & \frac{(\lambda-1)(\lambda-2)}{2} & & \vdots \\ & \ddots & \ddots & \ddots & 0 \\ 0 & & \frac{\lambda(\lambda-1)}{2} & \lambda(2-\lambda) & \frac{(\lambda-1)(\lambda-2)}{2} \end{pmatrix}. \tag{2.12}$$



2.1.4 Opérateur Toeplitz sur $\mathbb{N}$

On veut maintenant pouvoir représenter des schémas définis sur $\mathbb{N}$ comme ceux de la Section 1.5.1.

Définition 2.10 (Opérateur Toeplitz sur $\mathbb{N}$)**.** On définit l'*opérateur Toeplitz sur* $\mathbb{N}$ de la manière suivante :

$$T_{\mathbb{N}} : \begin{cases} \ell^2(\mathbb{N}) & \to & \ell^2(\mathbb{N}) \\ u = (u_n)_{n\in\mathbb{N}} & \mapsto & ((T_{\mathbb{N}}u)_n)_{n\in\mathbb{N}} \end{cases} \tag{2.13}$$

avec $\forall n \in \mathbb{N}, \quad (T_{\mathbb{N}}u)_n \overset{\text{def}}{=} \begin{cases} \sum_{j=-r}^{p} a_j u_{n+j} & \text{si } n \geqslant r, \\ \sum_{j=-n}^{p} a_j u_{n+j} & \text{si } n < r. \end{cases}$ On note $\Lambda(T_{\mathbb{N}})$ le spectre de l'opérateur $T_{\mathbb{N}}$.

Un opérateur Toeplitz $T_{\mathbb{N}}$ (défini en (2.13)) représente un schéma numérique (2.14) posé sur le demi-espace avec condition de Dirichlet homogène à gauche :

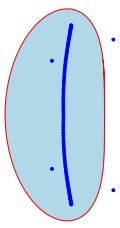
$$\begin{cases} U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{j+k}^n, & n \in \mathbb{N}, j \in \mathbb{N}, \\ U_j^n = 0 & j \in [\![-r : -1]\!]. \end{cases} \tag{2.14}$$

On peut définir de manière analogue à la matrice Quasi-Toeplitz définie en Définition 2.8 l'opérateur Quasi-Toeplitz sur $\mathbb{N}$, noté $\widetilde{T_{\mathbb{N}}}$, celui-ci représente exactement le schéma (1.15a)-(1.15b) pour $(g_j^n)_{j,n} = 0$. Le schéma posé sur $\mathbb{N}$ avec conditions de bord à gauche est le cadre dans lequel on travaille dans l'étude faite à partir du Chapitre 3.

Remarque 2.11. Pour justifier la terminologie de « valeur propre » de la Définition 1.18 (page 35) comme on le mentionne à la Remarque 1.19, il suffit de prendre une valeur propre z d'un opérateur Quasi-Toeplitz $\widetilde{T_{\mathbb{N}}}$ associée à un vecteur propre $\phi(z) = (\phi_j(z))_j$. En multipliant par z^n l'expression $\widetilde{T_{\mathbb{N}}}\phi(z) = z\phi(z)$, on obtient exactement que $U_j^n = z^n\phi_j(z)$ est solution de (1.6a)-(1.6b) pour $(g_j^n)_{j,n} = 0$. De plus, $\phi(z)$ est bien non nul car c'est un vecteur propre et dans $\ell^2(\mathbb{N})$ par définition de $\widetilde{T_{\mathbb{N}}}$. La seule différence est que, dans la Définition 1.18, on s'intéresse seulement aux valeurs propres de module supérieur à 1, car les valeurs propres de modules strictement inférieur à 1 ne représentent pas d'instabilité.

2.2 Spectre asymptotique

Pour étudier un schéma et notamment sa stabilité, on peut utiliser la matrice Quasi-Toeplitz correspondante, c'est l'approche qui est choisie dans différents articles comme dans [DDJ18] par exemple. La dimension $J+1$ de la matrice $\widetilde{T_J}$ est directement liée à la discrétisation spatiale Δx du schéma numérique. Pour avoir une inégalité de stabilité avec une constante indépendante de la discrétisation spatiale Δx, il nous faut un contrôle indépendant de J dans l'étude du spectre



de la matrice Quasi-Toeplitz. En effet, en itérant la formule $U^{n+1} = \widetilde{T_J} U^n$, on trouve l'égalité

$$U^n = \widetilde{T_J}^n U^0. \tag{2.15}$$

On veut donc borner les puissances de la matrice $\widetilde{T_J}$ indépendamment de J. Pour cela, on peut regarder le spectre de la matrice $\widetilde{T_J}$. Dans [DDJ18], on suppose que pour une dimension J relativement grande, le spectre est proche de celui de l'opérateur limite. On précisera cette idée dans la suite de ce chapitre.

Définition 2.12 (Spectre asymptotique de la matrice Quasi-Toeplitz)**.** On appelle *spectre asymptotique* de la matrice Quasi-Toeplitz $\widetilde{T_J}$ l'ensemble suivant :

$$\widetilde{\lim_{J\to\infty}} \Lambda(\widetilde{T_J}) \overset{\text{def}}{=} \left\{ \lim_{j\to\infty} \mu_j \text{ pour } \mu_j \in \Lambda(\widetilde{T_{J_j}}) \text{ avec } \lim_{j\to\infty} J_j = +\infty \right\}.$$

De la même manière, on peut définir le spectre asymptotique de la matrice Toeplitz T_J par :

$$\widetilde{\lim_{J\to\infty}} \Lambda(T_J) \overset{\text{def}}{=} \left\{ \lim_{j\to\infty} \mu_j \text{ pour } \mu_j \in \Lambda(T_{J_j}) \text{ avec } \lim_{j\to\infty} J_j = +\infty \right\}.$$

De plus, comme on l'a déjà vu à la Proposition 2.4, on a $\Lambda(T_{\mathbb{Z}}) = \widetilde{\lim_{J\to\infty}} \Lambda(T_J^\circ)$.

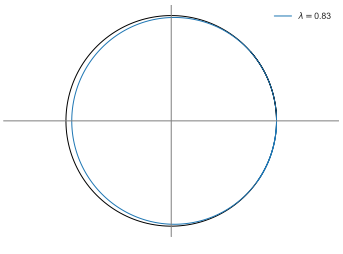
On utilise la notation « $\widetilde{\lim}$ » car c'est une limite à une sous-suite près, contrairement à la notation « lim », que l'on introduit en Notation 2.27 (page 58), où c'est une limite classique sans extraction de sous-suite, autrement dit, on a, pour $(A_n)_n \in \mathcal{P}(\mathbb{C})^{\mathbb{N}}$,

$$\widetilde{\lim_{n\to\infty}} A_n \overset{\text{def}}{=} \left\{ \lim_{n\to\infty} z_n \text{ pour } z_n \in A_{N_n} \text{ avec } \lim_{n\to\infty} N_n = +\infty \right\}$$
$$\text{et } \lim_{n\to\infty} A_n \overset{\text{def}}{=} \left\{ \lim_{n\to\infty} z_n \text{ pour } z_n \in A_n \right\}.$$

L'article de Beam et Warming [BW93] est dédié à l'étude du spectre asymptotique des matrices Toeplitz et Quasi-Toeplitz et donne un algorithme pour le calculer. Il explique notamment comment le spectre asymptotique d'une matrice Quasi-Toeplitz est constitué de deux parties. Cela s'appuie sur les résultats de Schmidt et Spitzer [SS60] qui donnent des propriétés sur les spectres asymptotiques des matrices Toeplitz et les spectres des opérateurs Toeplitz sur $\mathbb{Z}$ et sur $\mathbb{N}$.

2.2.1 Spectre asymptotique d'opérateurs Toeplitz sur $\mathbb{Z}$ et $\mathbb{N}$

On pose la fonction $\gamma : \kappa \mapsto \sum_{j=-r}^{p} a_j \kappa^j$, et on note $\gamma(\mathbb{S})$ la courbe $\{\xi \in \mathbb{R} \mapsto \gamma(e^{i\xi})\}$. La fonction γ est à relier avec la notion de symbole que l'on a définie dans la Section 1.3.1 (page 30), mais au lieu de se restreindre au cercle unité pour la définition (comme à la Définition 1.10), on étudie le symbole sur $\mathbb{C}$ tout entier. La courbe $\gamma(\mathbb{S})$ correspond à la courbe Γ définie en (1.12).



Proposition 2.13. *Les spectres des opérateurs Toeplitz sur* $\mathbb{Z}$ *et sur* $\mathbb{N}$ *sont*

$$\Lambda(T_{\mathbb{Z}}) = \{z \in \mathbb{C}, z \in \gamma(\mathbb{S})\} \quad et \quad \Lambda(T_{\mathbb{N}}) = \Lambda(T_{\mathbb{Z}}) \cup \{z \in \mathbb{C}, \mathrm{Ind}_{\gamma(\mathbb{S})}(z) \neq 0\}.$$

La première égalité a été prouvée à l'étape 2 de la preuve de la Proposition 2.4 et on peut trouver la preuve de la deuxième égalité dans l'article [SS60].

Exemple 2.14. L'opérateur *shift* sur $\mathbb{Z}$, resp. sur $\mathbb{N}$, a pour spectre le cercle unité $\mathbb{S}$, resp. le disque unité fermé $\overline{\mathbb{D}}$, comme représenté en Figure 2.2. On a représenté aussi les spectres des opérateurs $T_{\mathbb{Z}}$ et $T_{\mathbb{N}}$ liés au schéma jouet (2.4).

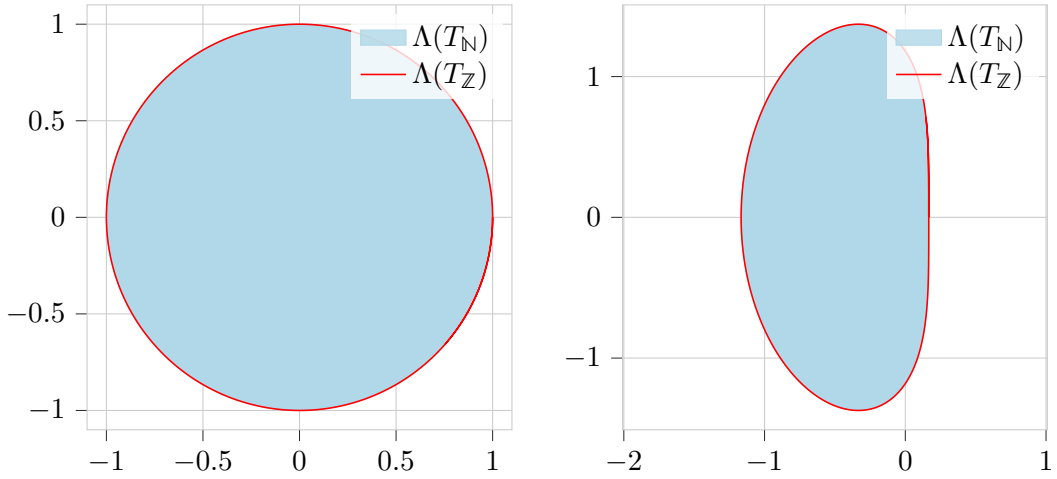


FIGURE 2.2 – Spectres des opérateurs Toeplitz sur $\mathbb{Z}$ et sur $\mathbb{N}$ pour le shift (à gauche) et pour le schéma (2.4) (à droite).

2.2.2 Spectre asymptotique de matrice Toeplitz

Proposition 2.15. *Le spectre asymptotique d'une matrice* Toeplitz T_J *s'identifie à l'ensemble suivant :*

$$\widetilde{\lim_{J\to\infty}} \Lambda(T_J) = \{z \in \mathbb{C}, |\kappa_r(z)| = |\kappa_{r+1}(z)|\}$$

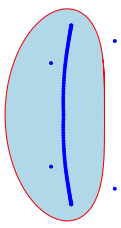
où, pour tout $z \in \mathbb{C}$, *la suite* $(\kappa_j(z))_{j=1}^{p+r}$ *correspond aux racines rangées par ordre croissant de module de l'équation (de paramètre* z*) suivante :*

$$z\kappa^r = \sum_{j=-r}^{p} a_j \kappa^{r+j}. \tag{2.16}$$

La preuve de ce résultat est l'objet des Lemmes 5.5 et 6.1 de [SS60]. De plus, en étudiant l'ensemble $\{z \in \mathbb{C}, |\kappa_r(z)| = |\kappa_{r+1}(z)|\}$, on peut voir qu'il est non vide et qu'il ne contient pas de point isolé (Corollaire 3.2 et Lemme 3.3 de [SS60]).

Proposition 2.16. *Le spectre asymptotique d'une matrice* Toeplitz T_J *est inclus dans* $\Lambda(T_{\mathbb{N}})$, *autrement dit :*

$$\widetilde{\lim_{J\to\infty}} \Lambda(T_J) \subset \Lambda(T_{\mathbb{N}}).$$



Cela vient du lemme 2.2 de [SS60].

Exemple 2.17. On représente en Figure 2.3 les différents spectres liés au shift défini dans l'Exemple 2.7 et au schéma jouet (2.4) de l'Exemple 2.5.

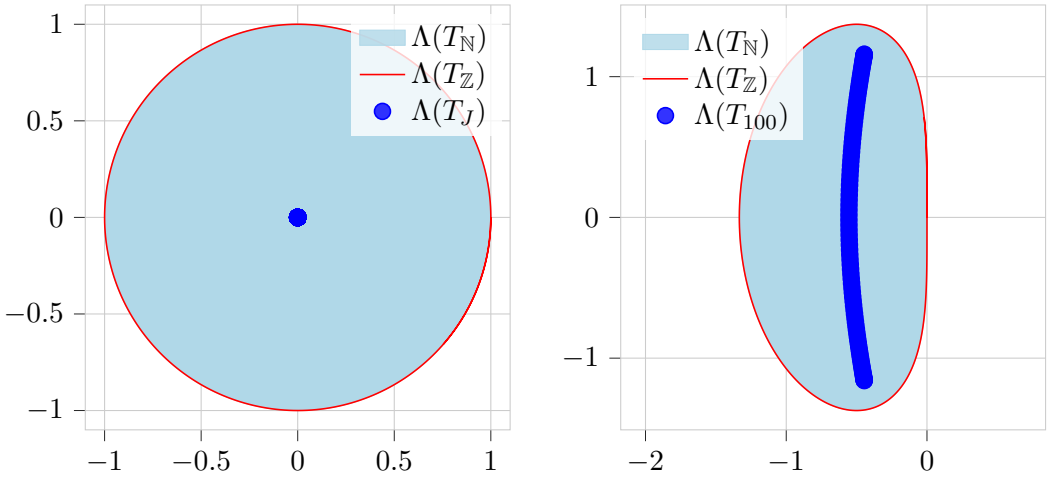


FIGURE 2.3 – Spectres de la matrice Toeplitz et des opérateurs Toeplitz sur $\mathbb{Z}$ et sur $\mathbb{N}$ pour le shift (à gauche) et pour le schéma (2.4) (à droite).

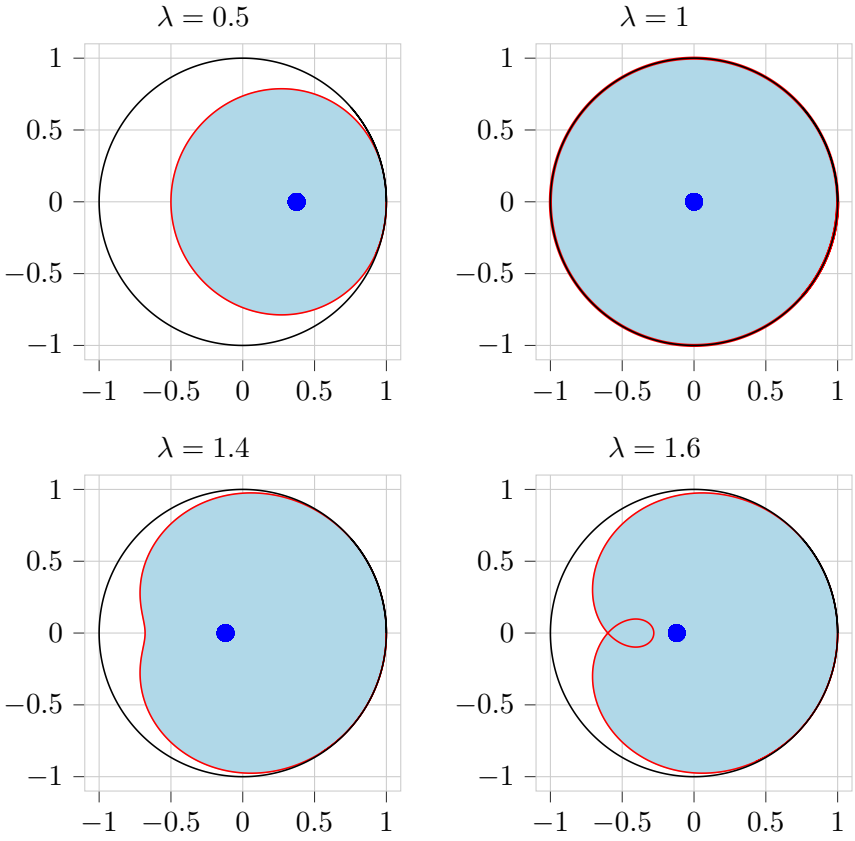
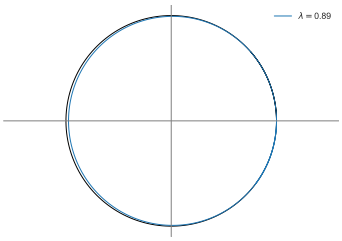


FIGURE 2.4 – Spectres de la matrice Toeplitz et des opérateurs Toeplitz sur $\mathbb{Z}$ et sur $\mathbb{N}$ pour le schéma de Beam-Warming BW_J défini en (2.8) pour différents λ (on a utilisé la même légende que la Figure 2.3 en ajoutant en noir le cercle unité).



Remarque 2.18. On peut observer sur les Figures 2.3 et 2.4 que le spectre asymptotique de T_J ne correspond pas au spectre de $T_{\mathbb{N}}$, alors qu'on aurait pu espérer que la limite du spectre de l'opérateur T_J soit le spectre de l'opérateur limite $T_{\mathbb{N}}$.

Remarque 2.19. Dans l'article de Beam et Warming [BW93], un algorithme est donné pour calculer le spectre asymptotique d'une matrice Toeplitz. Pour cela, ils s'affranchissent de l'étude d'un système de taille $J+1$ où $J+1$ est la dimension d'une matrice, ils résolvent seulement un système de taille $r+p+1$ qui est la largeur de la bande Toeplitz de la matrice Toeplitz. Cette taille est constante quand la dimension tend vers l'infini.

2.2.3 Spectre asymptotique de matrice Quasi-Toeplitz

Le spectre asymptotique d'une matrice Quasi-Toeplitz peut être séparé en deux ensembles : d'un côté, on trouve le spectre asymptotique de la matrice Toeplitz associée (sans la partie Quasi-Toeplitz) et de l'autre, des points isolés liés à la perturbation que l'on a mise aux bords.

Proposition 2.20. *Le spectre asymptotique d'une matrice Quasi-Toeplitz $\widetilde{T_J}$ se décompose de la manière suivante :*

$$\widetilde{\lim_{J\to\infty}} \Lambda(\widetilde{T_J}) = \widetilde{\lim_{J\to\infty}} \Lambda(T_J) \cup \mathcal{D}$$

où $\mathcal{D}$ est un ensemble de valeurs isolées défini dans [BW93].

Dans l'article [BW93], l'ensemble $\mathcal{D}$ est défini par

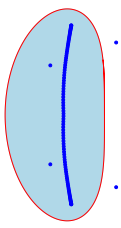
$$\{z \in \mathbb{C}, |\kappa_r(z)| < |\kappa_{r+1}(z)|, \\ \det(\mathcal{B}K^{(r)}_{0,m-1}(z) - zK^{(r)}_{0,r-1}(z)) = 0 \text{ ou } \det(\mathcal{C}K^{(p)}_{0,m-1}(z) - zK^{(p)}_{0,p-1}(z)) = 0\},$$

où $\mathcal{B}$ et $\mathcal{C}$ sont les matrices présentes dans la définition des matrices Quasi-Toeplitz (Définition 2.8) et les matrices $K^{(r)}_{0,j}(z)$ et $K^{(p)}_{0,j}(z)$ sont définies de la manière suivante :

$$K^{(r)}_{0,j}(z) \overset{\text{def}}{=} \begin{pmatrix} 1 & \cdots & 1 \\ \kappa_1(z) & & \kappa_r(z) \\ \vdots & & \vdots \\ \kappa_1(z)^j & \cdots & \kappa_r(z)^j \end{pmatrix} \text{ et } K^{(p)}_{0,j}(z) \overset{\text{def}}{=} \begin{pmatrix} 1 & \cdots & 1 \\ \kappa_{r+1}(z) & & \kappa_{r+p}(z) \\ \vdots & & \vdots \\ \kappa_{r+1}(z)^j & \cdots & \kappa_{r+p}(z)^j \end{pmatrix}.$$

Ici, on a supposé que toutes les racines $(\kappa_j(z))_{j=1}^{r+p}$ de (2.16) sont simples par soucis de lisibilité, mais si les racines de (2.16) sont multiples, il faut utiliser des matrices de Vandermonde généralisées (on discutera de cela dans la Section 3.2.3, page 78). La matrice $K^{(r)}_{0,j}(z)$ correspond à la matrice $K_{0,j}(z)$ qui est introduite dans la Section 3.2.3 qui est utile dans la suite du manuscrit quand on traite le cas avec un seul bord à gauche.

Évidemment, l'ensemble $\mathcal{D}$ ne dépend pas de la dimension J puisqu'on étudie le spectre asymptotique et cela se voit bien dans cette définition puisqu'on n'utilise que les constantes r,



p, m, les coefficients intérieurs $(a_j)_{j=-r}^{p}$, les racines $(\kappa_j(z))_{j=1}^{r+p}$ de (2.16) et les coefficients des matrices de bords $\mathcal{B}$ et $\mathcal{C}$.

L'ensemble $\mathcal{D}$ peut s'écrire $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ avec

$$\mathcal{D}_1 \overset{\text{def}}{=} \{z \in \mathbb{C}, |\kappa_r(z)| < |\kappa_{r+1}(z)|, \det(\mathcal{B}K^{(r)}_{0,m-1}(z) - zK^{(r)}_{0,r-1}(z)) = 0\}$$

$$\text{et } \mathcal{D}_2 \overset{\text{def}}{=} \{z \in \mathbb{C}, |\kappa_r(z)| < |\kappa_{r+1}(z)|, \det(\mathcal{C}K^{(p)}_{0,m-1}(z) - zK^{(p)}_{0,p-1}(z)) = 0\}.$$

Informellement, on peut y voir une ressemblance avec la Section 1.3 qui sépare le problème de stabilité du schéma borné en espace en trois problèmes de stabilité : le problème posé sur $\mathbb{Z}$, le problème posé sur $\mathbb{N}$ avec un bord à gauche et le problème posé sur $-\mathbb{N}$ avec un bord à droite. En effet, comme on l'a vu, le spectre asymptotique de la matrice Quasi-Toeplitz $\widetilde{T_J}$ se sépare de la manière suivante :

- la partie $\widetilde{\lim_{J\to\infty}} \Lambda(T_J)$ qui correspond à la Cauchy-stabilité, puisque $\widetilde{\lim_{J\to\infty}} \Lambda(T_J) \subset \Lambda(T_{\mathbb{N}})$ (Proposition 2.16) dont le contour est $\Lambda(T_{\mathbb{Z}}) = \gamma(\mathbb{S})$ la courbe du symbole.
- la partie $\mathcal{D}_1$ correspond à la stabilité du problème avec un bord à gauche posé sur $\mathbb{N}$.
- la partie $\mathcal{D}_2$ correspond à la stabilité du problème avec un bord à droite posé sur $-\mathbb{N}$.

Cependant le lien entre $\mathcal{D}_1$ et $\widetilde{T_{\mathbb{N}}}$ n'est pas forcément facile à établir rigoureusement. Comme le montre la Figure 2.5, on peut remarquer que la Proposition 2.16 n'est plus vraie pour le cas Quasi-Toeplitz, on peut avoir des valeurs propres du spectre asymptotique d'une matrice Quasi-Toeplitz qui ne sont pas dans celle de l'opérateur Toeplitz sur $\mathbb{N}$.

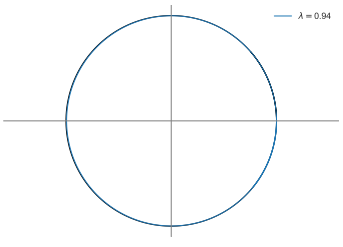
À notre connaissance, l'analyse spectrale des opérateurs Quasi-Toeplitz n'est pas étudiée d'un point de vue numérique dans la littérature. Beam et Warming [BW93] étudient les matrices Quasi-Toeplitz et Schmidt et Spitzer [SS60] les opérateurs Toeplitz sur $\mathbb{Z}$ et sur $\mathbb{N}$. De la même manière que pour le spectre des matrices Toeplitz et des matrices Quasi-Toeplitz, on peut se poser la question suivante :

Question ouverte 2.21. *Est-ce que le spectre d'un opérateur Quasi-Toeplitz $\widetilde{T_{\mathbb{N}}}$ est l'union du spectre de l'opérateur Toeplitz $T_{\mathbb{N}}$ associé et de l'ensemble $\mathcal{D}_1$?*

Exemple 2.22. On a représenté, dans la Figure 2.5, le spectre Quasi-Toeplitz du shift défini à l'Exemple 2.9 ainsi que la matrice Quasi-Toeplitz liée à (2.4) muni des deux bords

$$\mathcal{B} = \begin{pmatrix} -2.7 & 5.6 & -4.1 & 1.2 \end{pmatrix} \quad \text{et} \quad \mathcal{C} = \begin{pmatrix} 0.8 & -2.9 & 2.4 \end{pmatrix}, \tag{2.17}$$

où $\mathcal{B}$ et $\mathcal{C}$ sont les notations utilisées dans la Définition 2.8. Dans la figure de droite de la Figure 2.5, on a séparé le spectre de la matrice Quasi-Toeplitz en deux : le spectre de la matrice Toeplitz et la partie liée au bord.



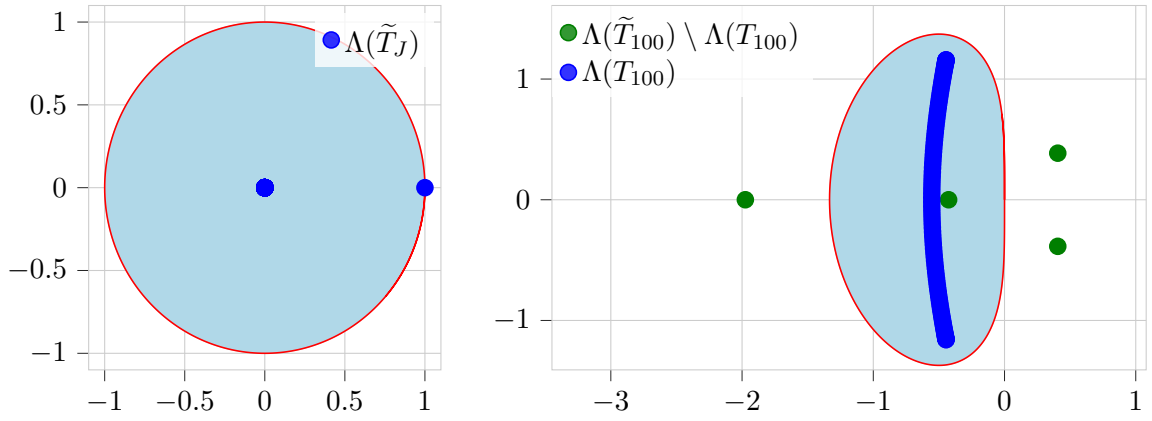


FIGURE 2.5 – Spectres de la matrice Quasi-Toeplitz et des opérateurs Toeplitz sur $\mathbb{Z}$ et sur $\mathbb{N}$ pour le shift (à gauche) et pour le schéma (2.4) muni des bords (2.17) (à droite) (on a utilisé la même légende que la Figure 2.3).

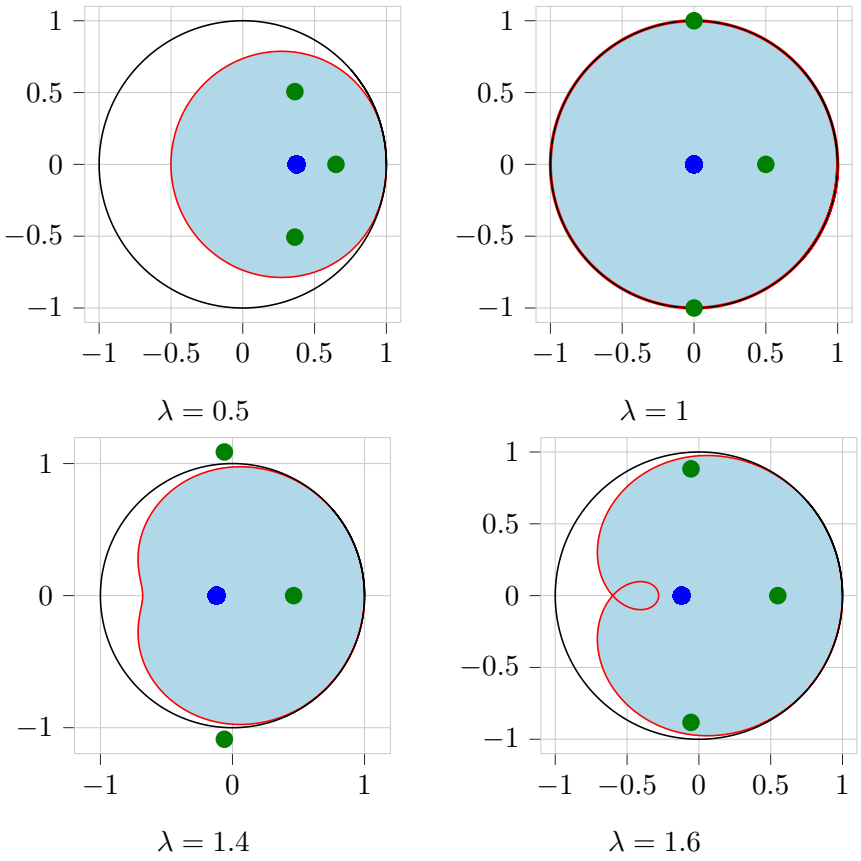
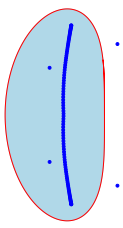


FIGURE 2.6 – Spectres de la matrice Quasi-Toeplitz et des opérateurs Toeplitz sur $\mathbb{Z}$ et sur $\mathbb{N}$ pour le schéma de Beam-Warming $\widetilde{BW}_J$ défini en (2.12) pour différents λ (on a utilisé la même légende que la Figure 2.5 en ajoutant en noir le cercle unité).

Dans toutes les figures, on se place pour J grand en supposant que le comportement de $\Lambda(\widetilde{T}_J)$ est proche de $\widetilde{\lim_{J\to\infty}} \Lambda(\widetilde{T}_J)$. Comme il a été expliqué à la Remarque 2.19, pour plus de précisions,



[BW93] donne un algorithme qui calcule le spectre asymptotique.

2.3 Pseudospectre

Comme on l'a vu à la Remarque 2.18, la limite (quand J tend vers l'infini) du spectre $\Lambda(T_J)$ de la matrice Toeplitz T_J n'est pas le spectre $\Lambda(T_{\mathbb{N}})$ de l'opérateur limite $T_{\mathbb{N}}$. Cependant, une définition plus générale de la notion de spectre, celle de *pseudospectre* permet de corriger ce « problème » de limite.

Dans toute cette section, on se fixe une norme subordonnée $\|\cdot\|$ sur l'espace des matrices associée à une norme $\|\cdot\|$ sur l'espace sous-jacent (que l'on notera de la même manière).

La majorité des résultats suivants peut être trouvée dans le livre de Trefethen et Embree [TE05].

Définition 2.23 (ε-pseudospectre)**.** Soient $\varepsilon > 0$ et $A \in \mathcal{M}_N(\mathbb{C})$ une matrice de taille $N \in \mathbb{N}^*$. Le *ε-pseudospectre* de A, noté $\Lambda_\varepsilon(A)$, est l'union des spectres des matrices $A+E$ avec $\|E\| \leqslant \varepsilon$, autrement dit :

$$\Lambda_\varepsilon(A) \overset{\text{def}}{=} \bigcup_{\|E\| \leqslant \varepsilon} \Lambda(A+E).$$

Il faut bien comprendre que le pseudospectre dépend de la norme $\|\cdot\|$ que l'on utilise. Dans toutes les figures, on représente le pseudospectre pour la norme $\|\cdot\|_2$.

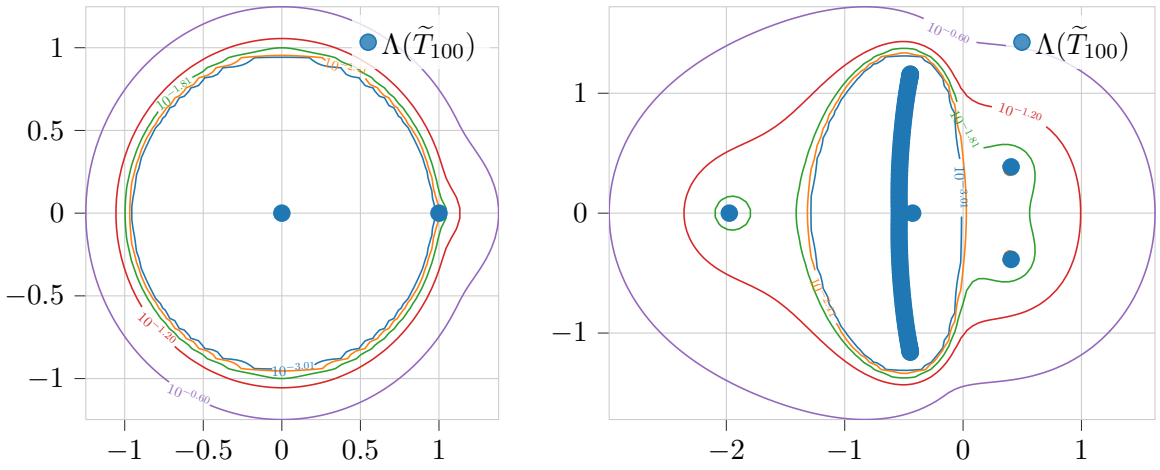


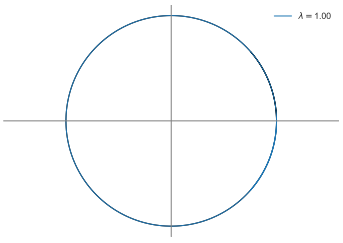
FIGURE 2.7 – Lignes de niveau du pseudospectre des matrices Quasi-Toeplitz ($J = 100$) du shift (à gauche) et du schéma (2.4) muni des bords (2.17) (à droite).

Proposition 2.24 (Caractérisation du pseudospectre)**.** *Soient $\varepsilon > 0$ et $A \in \mathcal{M}_N(\mathbb{C})$ une matrice de taille $N \in \mathbb{N}^*$. Il y a équivalence entre les assertions suivantes :*

(i) $z \in \Lambda_\varepsilon(A)$.

(ii) $\|(zI_N - A)^{-1}\| > \frac{1}{\varepsilon}$.

(iii) il existe $v \in \mathbb{C}^N$ *avec* $\|v\| = 1$ *tel que* $\|(z-A)v\| < \varepsilon$.



Cette proposition est démontrée dans [TE05, Th. 2.1].

Le pseudospectre possède les différentes propriétés suivantes.

Proposition 2.25. *Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice de taille $N \in \mathbb{N}^*$. On a*

- *si $\varepsilon_1 \leqslant \varepsilon_2$, alors $\Lambda_{\varepsilon_1}(A) \subseteq \Lambda_{\varepsilon_2}(A)$.*
- $\Lambda(A) = \bigcap_{\varepsilon>0} \Lambda_\varepsilon(A)$.

Proposition 2.26. *Soit $\varepsilon > 0$ et $A \in \mathcal{M}_N(\mathbb{C})$ une matrice de taille $N \in \mathbb{N}^*$. On a*

$$\Lambda(A) + B(0,\varepsilon) \subseteq \Lambda_\varepsilon(A).$$

De plus, si A est normale et si $\|\cdot\| = \|\cdot\|_2$, alors

$$\Lambda(A) + B(0,\varepsilon) = \Lambda_\varepsilon(A).$$

Ces propositions sont démontrées dans [TE05].

On veut étudier les résultats de pseudospectre pour les matrices Toeplitz et Quasi-Toeplitz.

2.3.1 Schéma totalement décentré sans bord

On s'appuie sur l'article [RT92] qui donne des résultats liés au pseudospectre pour les matrices Toeplitz.

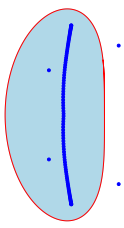
Pour les schémas totalement décentrés (schéma (1.2a) avec $p = 0$ que l'on étudiera dans le Chapitre 5 ou avec $r = 0$), on a une matrice de la forme

$$\begin{pmatrix} a_0 & & & & \\ \vdots & \ddots & & (0) & \\ a_{-r} & & \ddots & & \\ & \ddots & & \ddots & \\ (0) & & a_{-r} & \cdots & a_0 \end{pmatrix} \quad \text{ou} \quad \begin{pmatrix} a_0 & \cdots & a_p & & (0) \\ & \ddots & & \ddots & \\ & & \ddots & & a_p \\ & (0) & & \ddots & \vdots \\ & & & & a_0 \end{pmatrix}$$

On rappelle que $\gamma(\kappa) = \sum_{j=-r}^{p} a_j \kappa^j$ avec $r = 0$ ou $p = 0$ dans ce cadre.
On a les résultats suivants :

- $\Lambda(T_{\mathbb{Z}}) = \gamma(\mathbb{S})$.
- $\Lambda(T_{\mathbb{N}}) = \gamma(\overline{\mathbb{D}})$.
- $\Lambda(T_J) = \gamma(\{0\}) = \{a_0\}$, pour tout $J \in \mathbb{N}$.

On peut visualiser ces résultats sur la figure de gauche de la Figure 2.3 et sur la Figure 2.4.



Notation 2.27 (Limite d'ensembles)**.** Soit $(A_n) \in \mathcal{P}(\mathbb{C})^{\mathbb{N}}$ une suite d'ensembles de complexes. On note $\lim\limits_{n\to+\infty} A_n$ l'ensemble des limites d'éléments des A_n, autrement dit,

$$\lim_{n\to+\infty} A_n \overset{\text{def}}{=} \{z \in \mathbb{C}, z_n \xrightarrow[n\to+\infty]{} z \text{ avec } z_n \in A_n\}.$$

Proposition 2.28. *Pour (T_J) des matrices Toeplitz triangulaires, pour tout $\varepsilon > 0$, la limite du ε-pseudospectre de T_J quand J tend vers l'infini est le ε-pseudospectre de $T_{\mathbb{N}}$, autrement dit,*

$$\forall \varepsilon > 0, \quad \lim_{J\to\infty} \Lambda_\varepsilon(T_J) = \Lambda_\varepsilon(T_{\mathbb{N}}) = \Lambda(T_{\mathbb{N}}) + B(0,\varepsilon) = \gamma(\overline{\mathbb{D}}) + B(0,\varepsilon).$$

La preuve peut être trouvée dans [RT92, Théorème 2.3]. Il faut voir que la Proposition 2.28 caractérise $\Lambda_\varepsilon(T_{\mathbb{N}})$, en effet le ε-pseudospectre de $T_{\mathbb{N}}$ est simplement le spectre de $T_{\mathbb{N}}$ épaissi de ε.

La Proposition 2.28 et la Remarque 2.18 nous donne le diagramme non commutatif suivant :

$$\begin{array}{ccc} \Lambda_\varepsilon(T_J) & \xrightarrow[\varepsilon\to 0]{} & \Lambda(T_J) = \{a_0\} \\ \Big\downarrow {\scriptstyle J\to\infty} & & \Big\downarrow {\scriptstyle J\to\infty} \\ \Lambda_\varepsilon(T_{\mathbb{N}}) & \xrightarrow[\varepsilon\to 0]{} & \Lambda(T_{\mathbb{N}}) \end{array}$$

On peut visualiser cela dans la Figure 2.8 qui donne les lignes de niveaux du pseudospectre de BW_{100} (défini en (2.8)) ainsi que le spectre de T_{100} et le spectre de $T_{\mathbb{Z}}$. On peut aussi y lire le spectre de $T_{\mathbb{N}}$ grâce à la Proposition 2.13 qui montre que $\Lambda(T_{\mathbb{N}})$ est l'union de $\Lambda(T_{\mathbb{Z}})$ et des points intérieurs (*i.e.* d'indice complexe non nul) de la courbe $\Lambda(T_{\mathbb{Z}})$.

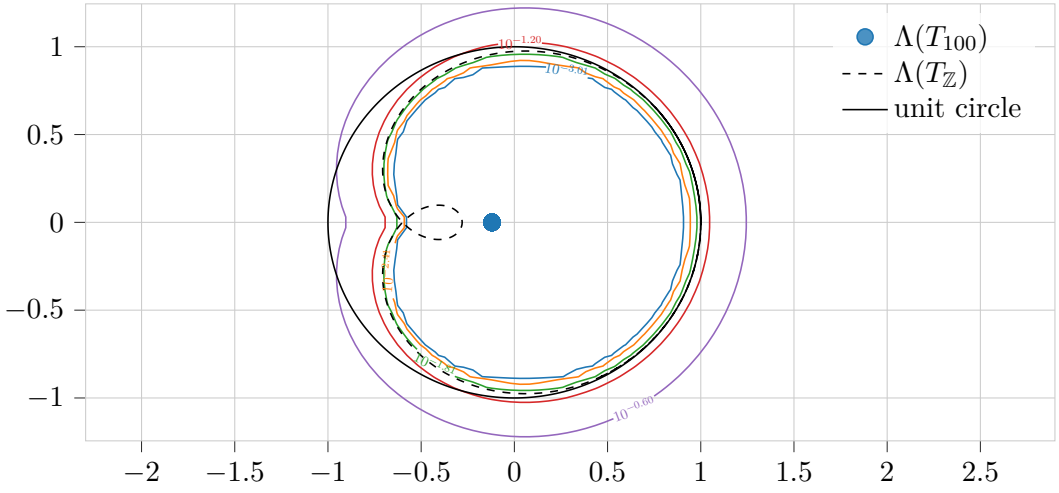
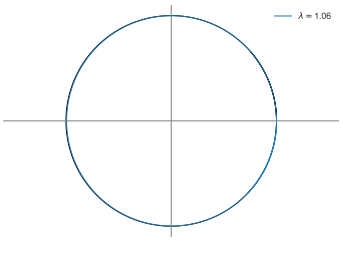


FIGURE 2.8 – Pseudospectre de la matrice Toeplitz liée au schéma de Beam-Warming pour $\lambda = 1.6$ et $J = 100$.



2.3.2 Schéma sans bord

L'article [RT92] traite aussi le cas général des matrices Toeplitz (pas seulement triangulaire comme dans la section précédente). On obtient les mêmes résultats de convergence, si ce n'est que l'on n'a pas une description aussi précise de $\Lambda_\varepsilon(T_\mathbb{N})$ en fonction de $\Lambda(T_\mathbb{N})$. Cela est expliqué plus en détails dans [RT92].

Proposition 2.29. *Pour (T_J) des matrices Toeplitz, pour tout $\varepsilon > 0$, la limite du ε-pseudospectre de T_J quand J tend vers l'infini est le ε-pseudospectre de $T_\mathbb{N}$, autrement dit,*

$$\forall \varepsilon > 0, \quad \lim_{J\to\infty} \Lambda_\varepsilon(T_J) = \Lambda_\varepsilon(T_\mathbb{N}).$$

Dans [RT92], il est expliqué que la preuve de ce résultat découle, selon [Wid89], du corollaire du Théorème II de [Wid89].

De nouveau, on a le diagramme non commutatif suivant :

$$\begin{array}{ccccc}
\Lambda_\varepsilon(T_J) & \xrightarrow[\varepsilon\to 0]{} & \Lambda(T_J) & \xrightarrow[J\to\infty]{} & \widetilde{\lim_{J\to\infty}} \Lambda(T_J) \\
\downarrow{\scriptstyle J\to\infty} & & \downarrow{\scriptstyle J\to\infty} & & \\
\Lambda_\varepsilon(T_\mathbb{N}) & \xrightarrow[\varepsilon\to 0]{} & \Lambda(T_\mathbb{N}) &&
\end{array}$$

On peut visualiser cela à la Figure 2.9 qui donne les lignes de niveaux du pseudospectre de T_{100}, le contour de $\Lambda(T_\mathbb{N})$, ainsi que le spectre asymptotique de la matrice T_J pour le schéma (2.4).

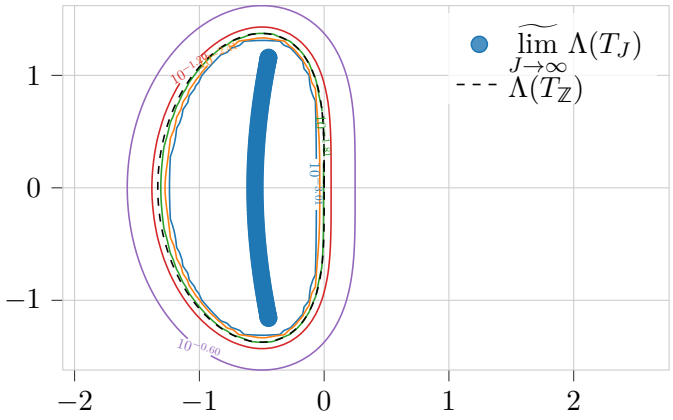
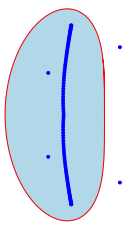


FIGURE 2.9 – Pseudospectre de la matrice Toeplitz T_{100} liée au schéma (2.4).

Question ouverte 2.30. *Comment mieux caractériser le pseudospectre d'un opérateur $T_\mathbb{N}$ défini sur $\ell^2(\mathbb{N})$? Peut-on trouver une procédure robuste pour tracer le pseudospectre d'un opérateur $T_\mathbb{N}$?*



2.3.3 Schéma avec bord

On a vu que les schémas avec bord sont représentés par des matrices Quasi-Toeplitz (voir Section 2.1.3). On ne peut pas espérer avoir convergence du spectre, mais il est naturel d'espérer avoir la propriété sur les pseudospectres

$$\forall \varepsilon > 0, \quad \lim_{J\to\infty} \Lambda_\varepsilon(\widetilde{T_J}) = \Lambda_\varepsilon(\widetilde{T_{\mathbb{N}}})$$

puisque la Proposition 2.29 montre que cela est vrai pour les matrices Toeplitz.

Il est dur de raisonner numériquement, car on ne sait pas tracer $\Lambda_\varepsilon(\widetilde{T_{\mathbb{N}}})$. En effet, il est déjà difficile de décrire $\Lambda_\varepsilon(T_{\mathbb{N}})$ pour des matrices Toeplitz, donc travailler avec une matrice Quasi-Toeplitz ne fera que compliquer la tâche.

Question ouverte 2.31. *Pour les matrices Quasi-Toeplitz, a-t-on*

$$\forall \varepsilon > 0, \quad \lim_{J\to\infty} \Lambda_\varepsilon(\widetilde{T_J}) = \Lambda_\varepsilon(\widetilde{T_{\mathbb{N}}}) \ ?$$

Une piste envisagée a été d'étudier les opérateurs Quasi-Toeplitz sur $\mathbb{N}$ comme des opérateurs de théorie spectrale avec les notions d'opérateurs de Fredholm (voir Annexe C, page 205). Le Théorème C.12 (page 207) nous permet d'étudier le spectre de $T_{\mathbb{N}}$ grâce à la notion de spectre essentiel, notée Λ_{ess}, voir Définition C.7 (page 206).

Proposition 2.32. *Pour tout opérateur Toeplitz $T_{\mathbb{N}}$ sur $\mathbb{N}$, on a $\Lambda(T_{\mathbb{N}}) = \Lambda_{ess}(T_{\mathbb{N}})$.*

Démonstration. En utilisant le Théorème C.12, on a

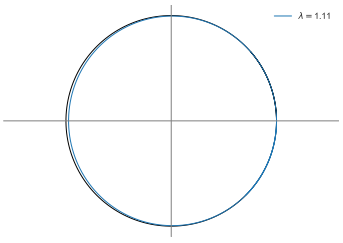
$$\begin{aligned} z \in \Lambda(T_{\mathbb{N}}) &\iff T_{\mathbb{N}} - z \text{ n'est pas bijectif} \\ &\iff T_{\mathbb{N}} - z \text{ est Fredholm d'indice non nul ou n'est pas Fredholm} \\ &\iff T_{\mathbb{N}} - z \text{ n'est pas Fredholm d'indice nul} \\ &\iff z \in \Lambda_{ess}(T_{\mathbb{N}}). \end{aligned}$$

□

Grâce à la Remarque C.15 et à la Proposition C.14, les opérateurs Quasi-Toeplitz $\widetilde{T_{\mathbb{N}}}$ peuvent être vus comme la somme d'un opérateur compact (la partie liée au bord) et un opérateur Toeplitz $T_{\mathbb{N}}$. La proposition suivante nous permet donc de décrire aussi le spectre de $\widetilde{T_{\mathbb{N}}}$.

Proposition 2.33. *Soit $\widetilde{T_{\mathbb{N}}}$ un opérateur Quasi-Toeplitz sur $\mathbb{N}$ et son opérateur Toeplitz associé $T_{\mathbb{N}}$. On a $\Lambda_{ess}(\widetilde{T_{\mathbb{N}}}) = \Lambda_{ess}(T_{\mathbb{N}})$.*

Démonstration. La Remarque C.15 permet d'écrire $\widetilde{T_{\mathbb{N}}} - z = T_{\mathbb{N}} - z + K$ pour un certain



opérateur K compact. Ainsi, en utilisant la Proposition C.6, on a

$$\begin{aligned} z \notin \Lambda_{ess}(T_{\mathbb{N}}) &\iff T_{\mathbb{N}} - z \text{ est Fredholm d'indice nul} \\ &\iff T_{\mathbb{N}} + K - z \text{ est Fredholm d'indice nul} \\ &\iff \widetilde{T_{\mathbb{N}}} - z \text{ est Fredholm d'indice nul} \\ &\iff z \notin \Lambda_{ess}(\widetilde{T_{\mathbb{N}}}). \end{aligned}$$

□

Ces propositions semblent proches de la vision du spectre asymptotique que l'on voit dans la figure de droite de la Figure 2.5 et dans [BW93]. En effet, on rappelle que le spectre asymptotique d'une matrice Quasi-Teoplitz est l'union entre le spectre asymptotique de la matrice Toeplitz associée et de valeurs isolées (voir Proposition 2.20). Cependant le lien rigoureux entre les deux visions ne semble pas évident à démontrer, car dans la vision du spectre asymptotique, on travaille avec des matrices (donc de dimension finie) et l'on a vu à la Remarque 2.18 que le spectre asymptotique et le spectre de l'opérateur limite (sur $\mathbb{N}$) ne coïncident pas.

Nous ne sommes pas allés plus loin dans cette direction. Nous avons préféré nous concentrer sur les bulbes du pseudospectre dont nous allons parler à la Section 2.4.2.

2.4 Kreiss Matrix Theorem

Comme on l'a vu au début de la Section 2.2, pour obtenir la stabilité d'un schéma, on cherche à borner les puissances de la matrice T_J associée au schéma. Pour cela, on va utiliser le Kreiss Matrix Theorem, introduit par Kreiss dans [Kre62] en 1962.

Dans toute cette section, on utilise de nouveau la norme euclidienne $\|\cdot\| = \|\cdot\|_2$.

Notation 2.34. Soit $N \in \mathbb{N}^*$. On utilise la notion de résolvante, notée $R_z(A)$, qui est définie, pour $A \in \mathcal{M}_N(\mathbb{C})$, de la manière suivante :

$$\forall z \in \mathbb{C} \setminus \Lambda(A), \quad R_z(A) \overset{\text{def}}{=} (zI_N - A)^{-1}.$$

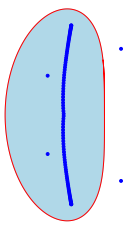
Pour tout $z \in \Lambda(A)$, on prendra comme convention que $\|R_z(A)\| = +\infty$.

Le point (ii) de la Proposition 2.24 peut donc se réécrire

$$\Lambda_\varepsilon(A) = \left\{ z \in \mathbb{C}, \|R_z(A)\| > \varepsilon^{-1} \right\}.$$

2.4.1 Kreiss Matrix Theorem

On veut relier le caractère borné des puissances de la matrice T_J et un contrôle de la résolvante grâce au *Kreiss Matrix Theorem* démontré par Kreiss [Kre62].



Théorème 2.35 (Kreiss Matrix Theorem)**.** *Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice de taille $N \in \mathbb{N}^*$. Les deux assertions suivantes sont équivalentes :*

(i) il existe $C(A) > 0$ telle que, pour tout $n \in \mathbb{N}$,

$$\|A^n\| \leqslant C(A).$$

(ii) il existe $K(A) > 0$ telle que, pour tout $|z| > 1$,

$$\|R_z(A)\| \leqslant \frac{K(A)}{|z| - 1}.$$

Derrière ce théorème, on veut aussi pouvoir avoir un contrôle sur les bornes en fonction de la dimension N de la matrice. En effet, comme la matrice T_J représente une discrétisation de taille J d'un schéma numérique, on veut une borne uniforme en J.

La proposition suivante permet de comprendre comment sont liées les constantes

$$C(A) \overset{\text{def}}{=} \sup_{n \in \mathbb{N}} \|A^n\| \quad \text{et} \quad K(A) \overset{\text{def}}{=} \sup_{|z|>1} (|z| - 1)\|R_z(A)\|$$

du Théorème 2.35.

Proposition 2.36. *Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice de taille $N \in \mathbb{N}^*$. On a*

$$K(A) \leqslant C(A) \leqslant eNK(A).$$

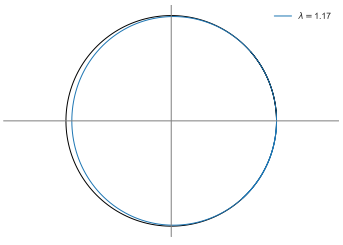
La preuve de ce résultat est un travail de longue haleine débuté en 1962. En effet, dans le travail de [Kre62], la borne devant $K(A)$ était c^{N^N}, où c est une constante, et a été raffiné au fur et à mesure des années : $\sim c^{N^N}$ par [Kre62], $\sim 6^N(N+4)^{5N}$ par [Mor64], $\sim N^N$ par [MS66], $\sim e^{9N^2}$ par [Mil67], $32eN^2/\pi$ par [Lap75], $32eN/\pi$ par [Tad81], $2eN$ par [LT84] et enfin eN par [LS91]. De plus, Spijker, Tracogna et Welfert [STW02] démontrent en 2002 que eN est la borne optimale de cette inégalité. La preuve de la Proposition 2.36 peut être trouvée dans les articles [Spi98], [TE05] et [STW02].

Si on veut une borne de la $n^{\text{ième}}$ puissance de la matrice A, on peut raffiner la Proposition 2.36 afin d'obtenir l'inégalité

$$\|A^n\| \leqslant e \min(n+1, N)K(A),$$

voir l'article de Spijker [Spi98].

Dans le livre de Trefethen et Embree [TE05], on énonce une expression explicite de $K(A)$ mettant en jeu le rayon pseudospectral. On en donnera ici une preuve qui s'inspire de l'article de Toh et Trefethen [TT99].



Proposition 2.37. *Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice de taille $N \in \mathbb{N}^*$ dont les puissances sont bornées. On a alors*

$$K(A) \overset{def}{=} \sup_{|z|>1} (|z|-1)\|R_z(A)\| = \sup_{\varepsilon>0} \frac{\rho_\varepsilon(A)-1}{\varepsilon} < +\infty$$

où $\rho_\varepsilon(A) \overset{def}{=} \sup_{z\in\Lambda_\varepsilon(A)} |z|$ est le rayon pseudospectral de A.

Démonstration. Montrons dans un premier temps qu'il y a équivalence entre les deux assertions suivantes :

(i) il existe $C > 0$ tel que pour tout $|z| > 1$, on a $\|R_z(A)\| \leqslant \dfrac{C}{|z|-1}$.

(ii) il existe $C > 0$ tel que pour tout $\varepsilon \geqslant 0$, pour tout $z \in \Lambda_\varepsilon(A)$, $\max(0, |z|-1) \leqslant C\varepsilon$.

(i) $\implies$ (ii) Comme les puissances de la matrice A sont bornées, on a $\Lambda(A) \subset \overline{\mathbb{D}}$. Si $\varepsilon = 0$, pour tout $z \in \Lambda_0(A) = \Lambda(A)$, on a $\max(0, |z|-1) = 0 \leqslant C\varepsilon$. Si $\varepsilon > 0$, pour tout $z \in \Lambda_\varepsilon(A)$, on a $\|R_z(A)\| > \varepsilon^{-1}$. Si $|z| > 1$, $\max(0, |z|-1) = |z|-1$ et $|z|-1 \leqslant \frac{C}{\|R_z(A)\|} \leqslant C\varepsilon$. Sinon $|z| \leqslant 1$ et $\max(0, |z|-1) = 0 \leqslant C\varepsilon$.

(ii) $\implies$ (i) Soit $|z| > 1$. On pose $\varepsilon_0 = \inf\{\varepsilon > 0, z \in \Lambda_\varepsilon(A)\}$. On a $0 < \varepsilon_0 < +\infty$. Par inclusion croissante des $\Lambda_\varepsilon(A)$ par rapport à ε, il existe $\varepsilon \in]0, \varepsilon_0[$ tel que $z \notin \Lambda_{\varepsilon_0-\varepsilon}(A)$ et $z \in \Lambda_{\varepsilon_0+\varepsilon}(A)$. Ainsi, on a $\|R_z(A)\| \leqslant \frac{1}{\varepsilon_0-\varepsilon}$ et $|z|-1 \leqslant C(\varepsilon_0+\varepsilon)$. On obtient alors

$$\|R_z(A)\|(|z|-1) \leqslant C\frac{\varepsilon_0+\varepsilon}{\varepsilon_0-\varepsilon}.$$

En faisant tendre ε vers 0, on obtient le résultat attendu.

On a vu que le C pour les deux assertions est le même et par le Kreiss Matrix Theorem (Théorème 2.35), la quantité $\sup_{|z|>1}(|z|-1)\|R_z(A)\|$ est bornée. En passant au sup dans les deux assertions, on obtient

$$\sup_{|z|>1} (|z|-1)\|R_z(A)\| = \sup_{\varepsilon>0} \sup_{z\in\Lambda_\varepsilon(A)} \frac{\max(0,|z|-1)}{\varepsilon} < +\infty.$$

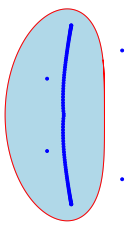
De plus, comme pour ε suffisamment large, il existe des $z \in \Lambda_\varepsilon(A)$ tels que $|z| > 1$, on peut enlever le max car $|z| > 1$ prédominera. On a alors

$$\sup_{\varepsilon>0} \sup_{z\in\Lambda_\varepsilon(A)} \frac{\max(0,|z|-1)}{\varepsilon} = \sup_{\varepsilon>0} \sup_{z\in\Lambda_\varepsilon(A)} \frac{|z|-1}{\varepsilon} = \sup_{\varepsilon>0} \frac{\rho_\varepsilon(A)-1}{\varepsilon}.$$

□

Pour avoir la stabilité d'un schéma représenté par une matrice Toeplitz/Quasi-Toeplitz on aimerait bien pouvoir contrôler les puissances de cette matrice $T_J/\widetilde{T_J}$, mais indépendamment de la dimension J de la matrice.

Question ouverte 2.38. *Est-ce que la constante $e(J+1)$ de la Proposition 2.36 est optimale pour le sous-ensemble des matrices Toeplitz/Quasi-Toeplitz ou même pour le sous-sous-ensemble des matrices Toeplitz/Quasi-Toeplitz résultant d'un schéma numérique consistant ?*



2.4.2 Bulbe du pseudospectre

On suppose que le spectre de T_J est inclus dans $\overline{\mathbb{D}}$ car c'est une condition nécessaire pour que les puissances de T_J soient bornées, on peut relier cette hypothèse avec la condition de Godunov–Ryabenkii (Proposition 1.20, page 36) à J fixé. De plus, si le spectre de T_J est inclus dans $\mathbb{D}$, les puissances de T_J sont bornées, on cherche donc à étudier l'influence des valeurs propres se trouvant sur le cercle unité.

Lorsque l'on trace le pseudospectre de matrices Toeplitz, on peut parfois observer l'apparition d'excroissances au niveau de certaines valeurs du cercle unité que l'on appelera *bulbe*.

Exemple 2.39. Dans la figure de gauche de la Figure 2.7 (page 57), pour la matrice Quasi-Toeplitz du shift, on voit qu'il y a un bulbe autour de la valeur propre 1. En fait le bulbe est lié au caractère non borné des puissances de $\widetilde{\mathcal{S}}_J$ à cause de la valeur propre 1.

Définition 2.40 (Bulbe)**.** Pour $J \in \mathbb{N}^*$ ou $J = \mathbb{N}$. On dit que le pseudospectre de T_J possède un *bulbe au point* $z_0 \in \mathbb{S}$ si

$$\sup_{\substack{z \to z_0 \\ |z|>1}} (|z| - 1)\|R_z(T_J)\| = +\infty.$$

De part cette définition et l'énoncé de Kreiss Matrix Theorem (Théorème 2.35), on voit que l'apparition de bulbe est relié au caractère non borné des puissances de T_J.

La Figure 2.10 du pseudospectre de la matrice Quasi-Toeplitz $\widetilde{BW}_{100}$ (défini en (2.12)) pour $\lambda = 1$ présente deux bulbes, en i et en $-i$.

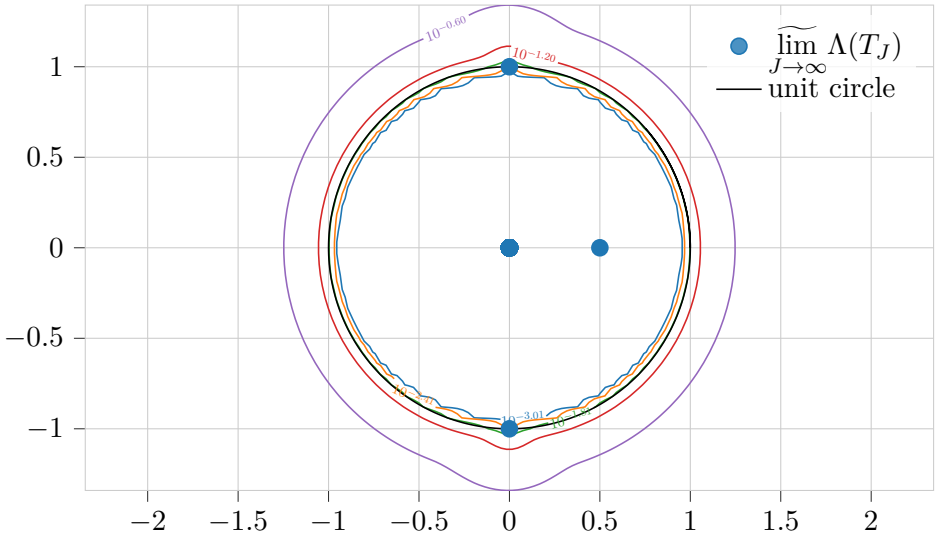
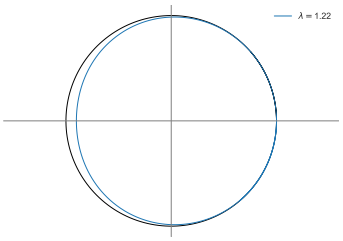


FIGURE 2.10 – Pseudospectre de la matrice Quasi-Toeplitz liée au schéma de Beam-Warming avec condition de bord S2ILW3 pour $\lambda = 1$ et $J = 100$.

La question ouverte suivante est primordiale, car elle relie la dimension fixée finie J et $\mathbb{N}$, ce que l'on ne peut pas faire facilement en étudiant le spectre usuel de T_J comme on l'a vu déjà vu à la Remarque 2.18.



Question ouverte 2.41. *Est-ce que l'apparition d'un bulbe est indépendante de la dimension J Si oui, apparaît-il aussi dans le pseudospectre de $T_{\mathbb{N}}$?*

2.4.3 Lien valeurs propres généralisées et bulbes

L'apparition de bulbes semble correspondre à l'existence de valeurs propres généralisées (voir Définition 1.21, page 36) et donc à des instabilités. En effet, on peut citer le livre de Trefethen et Embree [TE05],

> « We know from the Kreiss Matrix Theorem that although a bulge in pseudospectra guarantees instability, the lack of a bulge cannot guarantee stability under all circumstances ; there is a gap between upper and lower bounds of a factor of either n, the time step, or J, the matrix dimension. Thus a complete connection between resolvent norms or pseudospectra and stability would necessarily be delicate. »
>
> [TE05, Chap.34]

Proposition 2.42. *Si le pseudospectre de T_J possède un bulbe, alors le schéma lié à T_J est instable.*

Démonstration. Par définition du bulbe (Définition 2.40), on sait que $K(T_J) = +\infty$. Ainsi, par la Proposition 2.36, on a

$$\sup_{n\in\mathbb{N}} \|T_J^n\| \geqslant K(T_J) = +\infty.$$

Le schéma est donc instable. □

Comme Trefethen et Embree le disent (dans la citation de [TE05] au dessus de la Proposition 2.42), la réciproque n'est pas claire. En effet, si le schéma est instable, on traduit cela par $\sup_{J\in\mathbb{N}^*} \sup_{n\in\mathbb{N}} \|T_J^n\| = +\infty$ mais la Proposition 2.36 nous donne seulement

$$+\infty = \sup_{J\in\mathbb{N}^*} C(T_J) \leqslant e(J+1)K(T_J).$$

Or $\sup_{J\in\mathbb{N}^*} e(J+1)K(T_J) = +\infty$ n'implique ni que $K(T_J)$ est borné, ni qu'il n'est pas borné.

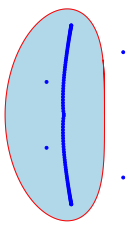
On suppose que le spectre de $T_{\mathbb{N}}$ est inclus dans $\overline{\mathbb{D}}$ afin d'être sûr que la condition de Godunov–Ryabenkii est satisfaite. On voudrait pouvoir relier l'apparition de bulbe et la notion de valeur propre généralisée définie pour un schéma lié à $T_{\mathbb{N}}$.

Question ouverte 2.43. *Peut-on faire le lien entre valeur propre généralisée et l'apparition de bulbe dans le pseudospectre ?*

Pour appréhender cette question, on a voulu représenter le pseudospectre de cas connus possédant des valeurs propres généralisées.

2.4.4 Exemple de conditions de bord sur le schéma leap-frog

Dans [Tre84], Trefethen donne quatre conditions de bord différentes pour le schéma leap-frog afin de classifier différents types d'instabilité. Le schéma leap-frog n'est pas un schéma à



un pas en temps, il ne peut pas s'écrire directement sous la forme de (1.2a), c'est un schéma multipas dont on donne plus de détails en Annexe D (page 209). On y décrit plus en détails, à la Section D.2 (page 213), les quatre conditions de bord que l'on va étudier dans cette section. Certaines de ces conditions de bord mettent en jeu des valeurs propres généralisées, on va tracer le pseudospectre pour voir s'il y a apparition de bulbe au niveau des valeurs propres généralisées.

On rappelle que le schéma leap-frog peut s'écrire de la manière suivante :

$$U_j^{n+1} = U_j^{n-1} + \lambda(U_{j+1}^n - U_{j-1}^n), \qquad j \geqslant 1, n \geqslant 1,$$

où $\lambda = \frac{a\Delta t}{\Delta x}$ vérifie la condition CFL : $0 < \lambda < 1$ (voir Annexe D, page 209).

L'équation caractéristique associée à ce schéma est

$$z - \frac{1}{z} = \lambda\left(\kappa - \frac{1}{\kappa}\right). \tag{2.18}$$

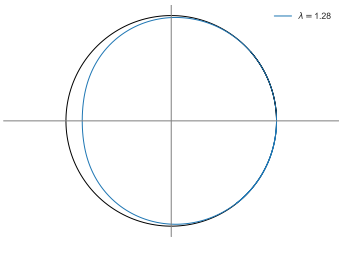
On va étudier la stabilité autour de certains modes propres $(z, \kappa(z))$. À chaque $z \in \overline{\mathcal{U}}$, on associe deux solutions $\kappa_-(z) \in \overline{\mathbb{D}}$ et $\kappa_+(z) \in \overline{\mathcal{U}}$ racines de l'équation (2.18) comme représenté à la Figure D.5 (page 212).

Les quatre conditions de bord de [Tre84] que l'on va considérer sont les suivantes :

$$\alpha : \quad U_0^{n+1} = U_1^n. \tag{α}$$
$$\beta : \quad U_0^{n+1} = U_1^{n-2}. \tag{β}$$
$$\gamma : \quad U_0^{n+1} = \frac{1}{2}(U_0^n + U_2^n). \tag{γ}$$
$$\delta : \quad U_0^{n+1} = U_1^{n+1}. \tag{δ}$$

La Figure 2.11 donne les pseudospectres du schéma leap-frog muni des différentes conditions de bord $(\alpha), (\beta), (\gamma), (\delta)$ pour $J = 50$ et $\lambda = 0.5$.

La condition de bord (α) correspond à un schéma stable, toutes les valeurs propres sont dans $\overline{\mathbb{D}}$ et visuellement, on n'observe aucun bulbe. La condition de bord (β) correspond à un schéma instable, car $z = \pm e^{\pm i\pi/6}$ sont des valeurs propres généralisées (voir [Tre84] pour plus de précisions). Cela n'est pas clair s'il y a un bulbe sur la Figure 2.11. Cela nous encourage à tracer la Figure 2.12 qui étudie la quantité $(|z| - 1)\|R_z(T_J)\|$. La condition de bord (γ) correspond à un schéma instable avec pour valeur propre généralisée $z = 1$. On observe en effet un bulbe en $z = 1$. La condition de bord (δ) correspond aussi à un schéma instable avec pour valeur propre généralisée $z = -1$ et on observe également un bulbe dans le pseudospectre.



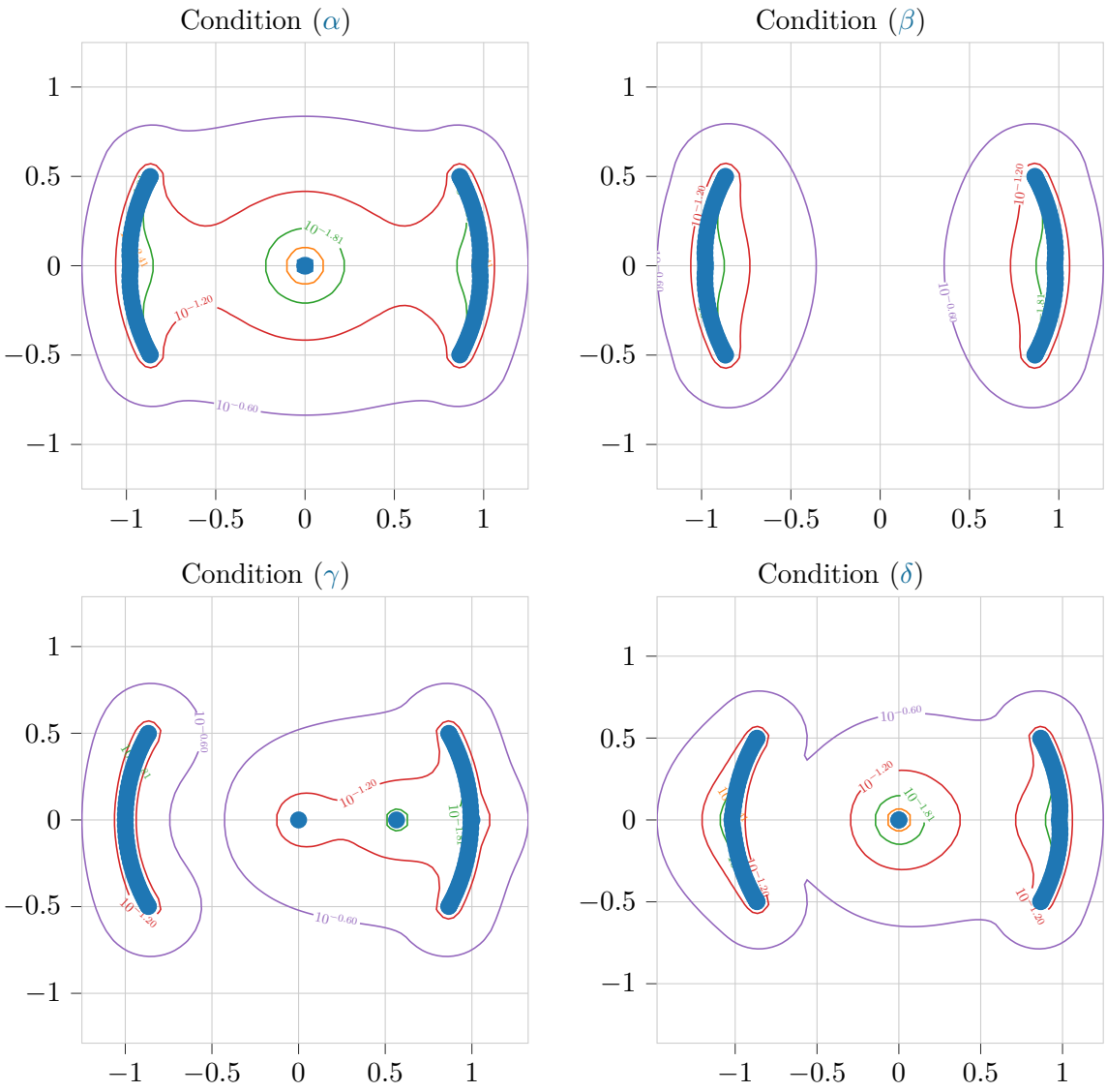


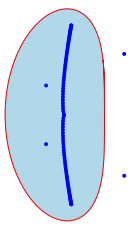
FIGURE 2.11 – Pseudospectre du schéma leap-frog avec différentes conditions de bord.

Dans la Figure 2.12, on trace la fonction $x \in [-2, 2] \mapsto (|x|-1)\|R_x(T_{100})\|$ pour les conditions de bord (α), (γ) et (δ). Cette figure confirme les observations déjà faites en Figure 2.11.

Dans la Figure 2.12, pour la condition de bord (β), on trace la fonction

$$x \in [-2, 2] \mapsto (|x| - 1)\|R_{xe^{i\pi/6}}(T_{100})\|$$

puisqu'il y a deux valeurs propres généralisées sur l'axe $[-2e^{i\pi/6}, 2e^{i\pi/6}]$. La courbe ne semble pas tendre vers l'infini au voisinage de $\pm e^{i\pi/6}$ et donc il semblerait qu'il n'y ait pas de bulbe. Cependant la Figure 2.12 est tracée pour $J = 100$ et le comportement pourrait être différent pour $J \to +\infty$. S'il n'y a pas de bulbe, on serait exactement dans le cas où le schéma est instable sans apparition de bulbe (voir citation de [TE05] au dessus de la Proposition 2.42).



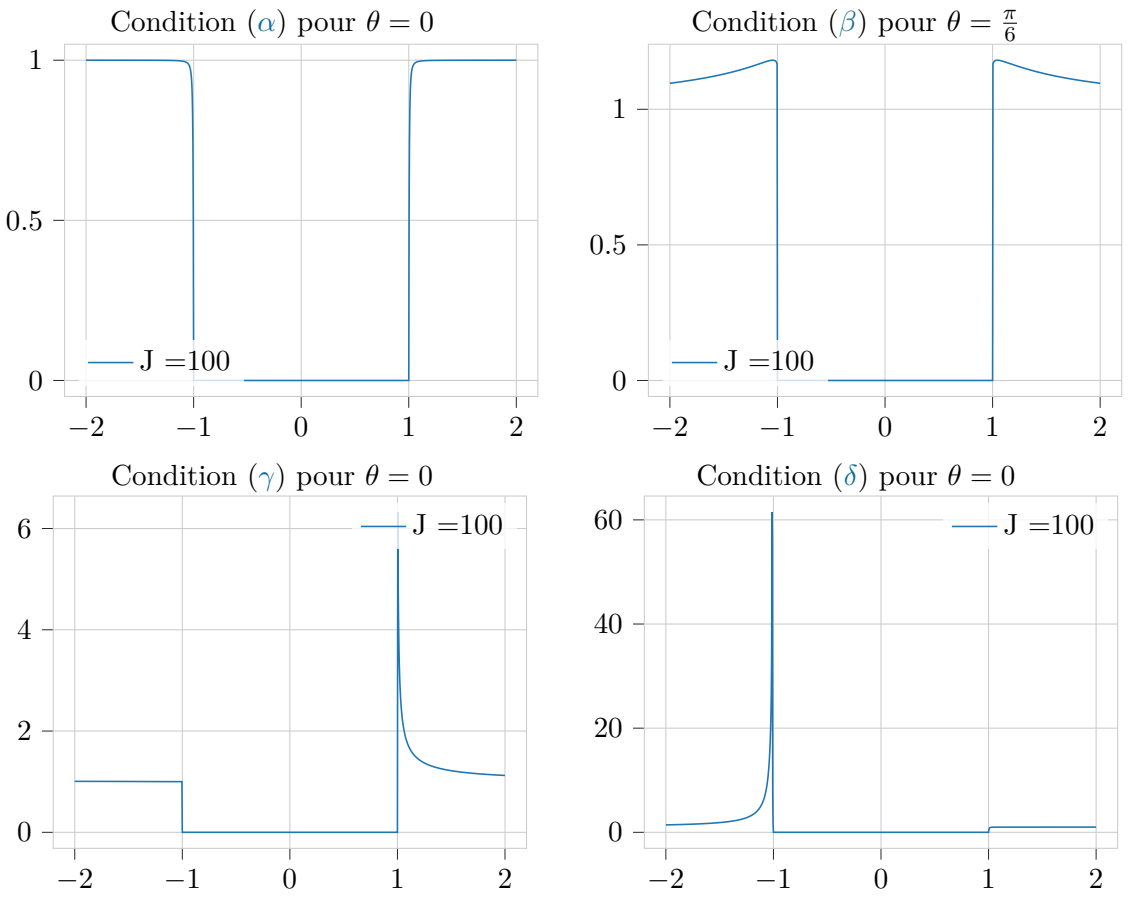
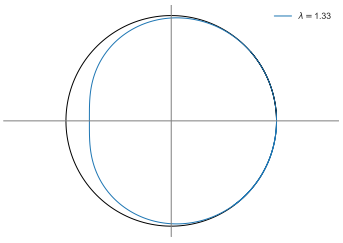


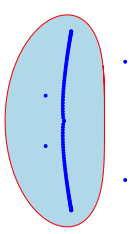
FIGURE 2.12 – Tracé de $(|x| - 1)\|R_{xe^{i\theta}}(T_{100})\|$ pour $x \in [-2, 2]$ où T_{100} est le schéma leap-frog avec différentes conditions de bord pour $\lambda = 0.5$.

Dans [Tre84], Trefethen explique que cas (β) correspond à un mode propagatif croissant avec vitesse de groupe nulle, il parle de « rightgoing steady state solution », contrairement aux cas (γ) et (δ) qui correspondent à des modes propagatifs strictement croissant (« strictly rightgoing steady state solution »).

Question ouverte 2.44. *Est-ce que les bulbes correspondent à des valeurs propres généralisées avec vitesse de groupe non nulle ?*

La difficulté pour passer de la dimension finie J de T_J à la dimension infinie de $T_{\mathbb{N}}$, ainsi que le manque d'information donnée par les bulbes des pseudospectres (et la difficulté à identifier les bulbes) nous pousse à travailler davantage sur la résolution algébrique des équations du schéma directement posé sur $\mathbb{N}$. C'est notamment l'approche de toute la suite du manuscrit. Dans le chapitre suivant, on précise davantage la vision des modes propres en utilisant la transformée en $\mathcal{Z}$. On donne alors une deuxième version du théorème de Kreiss (Théorème 1.22 pour la première version et Théorème 3.23, page 89, pour la deuxième) qui utilise la condition de Kreiss–Lopatinskii uniforme et le déterminant de Kreiss–Lopatinskii.

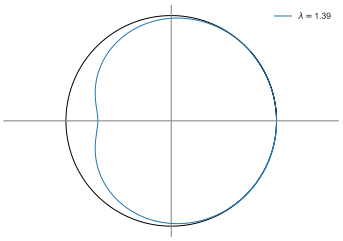




Chapitre 3

Théorie de Gustafsson, Kreiss et Sundström

Ce chapitre présente la théorie GKS, du nom de Gustafsson, Kreiss et Sundström, introduite en 1972 dans l'article [GKS72]. Elle permet d'étudier la stabilité forte et complète l'étude des modes propres présentée en Section 1.4 (page 35) grâce à l'utilisation de la transformée en $\mathcal{Z}$.



On rappelle que l'on veut étudier la stabilité forte du schéma suivant :

$$\begin{cases} U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{k+j}^n, & 0 \leqslant j < +\infty,\ n \geqslant 0, \quad (3.1a) \\ U_j^n = \sum_{k=0}^{m-1} b_{j,k} U_k^n + g_j^n, & j \in [\![-r : -1]\!],\ n \geqslant 0, \quad (3.1b) \\ U_j^0 = 0, & 0 \leqslant j < +\infty. \quad (3.1c) \end{cases}$$

Dans toute la suite, on note ℓ^2 pour l'espace $\ell^2([\![-r : -1]\!] \cup \mathbb{N})$ et on suppose que le schéma (3.1a) est Cauchy-stable (Définition 1.11, page 31).

3.1 Transformée en $\mathcal{Z}$

3.1.1 Formulation du schéma

On utilise la transformée en $\mathcal{Z}$ décrite en Annexe B (page 203) sur le schéma (3.1). On obtient alors la formulation suivante :

$$\forall z \in \mathcal{U}, \quad \begin{cases} z\widetilde{U}_j(z) = \sum_{k=-r}^{p} a_k \widetilde{U}_{k+j}(z) & 0 \leqslant j < +\infty, \quad (3.2a) \\ \widetilde{U}_j(z) = \sum_{k=0}^{m-1} b_{j,k} \widetilde{U}_k(z) + \widetilde{g}_j(z), & j \in [\![-r : -1]\!], \quad (3.2b) \\ (\widetilde{U}_j(z))_j \in \ell^2. & \quad (3.2c) \end{cases}$$

On rappelle que $\mathcal{U}$ est l'ensemble $\mathcal{U} \overset{\text{def}}{=} \{z \in \mathbb{C}, |z| > 1\}$.

3.1.2 Formulation résolvante de la stabilité forte

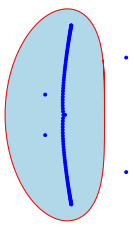
La définition de la stabilité forte (Définition 1.16, page 34) devient alors l'inégalité suivante, dite *formulation résolvante*.

Définition 3.1 (Formulation résolvante de la stabilité forte)**.** La formulation résolvante de l'inégalité (1.13) de stabilité forte est la suivante :

$$\forall |z| > 1, \quad \sum_{j=-r}^{-1} |\widetilde{U}_j(z)|^2 + \frac{|z|-1}{|z|} \sum_{j=-r}^{+\infty} |\widetilde{U}_j(z)|^2 \leqslant K \sum_{j=-r}^{-1} |\widetilde{g}_j(z)|^2. \quad (3.3)$$

La proposition suivante permet d'utiliser l'inégalité (3.3) pour obtenir la stabilité forte quand on retransforme $(\widetilde{U}_j(z))$ en (U_j^n) grâce à la tranformée en $\mathcal{Z}$ inverse (Proposition B.2, page 203).

Proposition 3.2. *Une solution $(\widetilde{U}_j(z))$ de (3.2) satisfait (3.3) si et seulement si la solution associée (U_j^n) satisfait l'inégalité (1.13) de stabilité forte.*



Démonstration. On procède par équivalence. Par l'égalité de Parseval pour la transformée en $\mathcal{Z}$ (Proposition B.3, page 204), l'inégalité

$$\forall z \in \mathcal{U}, \quad \sum_{j=-r}^{-1} |\widetilde{U}_j(z)|^2 + \frac{|z|-1}{|z|} \sum_{j=-r}^{+\infty} |\widetilde{U}_j(z)|^2 \leqslant K \sum_{j=-r}^{-1} |\widetilde{g}_j(z)|^2$$

est équivalente à l'inégalité suivante :

$$\forall R > 1, \quad \sum_{j=-r}^{-1} \sum_{n=0}^{+\infty} R^{-2n} |U_j^n|^2 + \frac{R-1}{R} \sum_{j=-r}^{+\infty} \sum_{n=0}^{+\infty} R^{-2n} |U_j^n|^2 \leqslant K \sum_{j=-r}^{-1} \sum_{n=0}^{+\infty} R^{-2n} |g_j^n|^2.$$

En posant $R = e^{\alpha \Delta t}$ pour $\alpha > 0$ et en multipliant toute l'inégalité par Δt, on obtient

$$\sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2 + \frac{e^{\alpha \Delta t} - 1}{e^{\alpha \Delta t}} \frac{1}{\Delta x} \|e^{-\alpha n \Delta t} U\|_{\Delta x, \Delta t}^2 \leqslant K \sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} g_j\|_{\Delta t}^2.$$

On utilise le fait qu'on a, pour tout $y > 0$, la suite d'inégalités suivante :

$$\frac{y}{1+y} \leqslant \frac{e^y - 1}{e^y} \leqslant \frac{2y}{1+y}.$$

En l'appliquant avec $y = \alpha \Delta t$, on obtient

$$\sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2 + \frac{\alpha}{1 + \alpha \Delta t} \frac{\Delta t}{\Delta x} \|e^{-\alpha n \Delta t} U\|_{\Delta x, \Delta t}^2 \leqslant K \sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} g_j\|_{\Delta t}^2.$$

Enfin, comme $\lambda = \dfrac{a \Delta t}{\Delta x}$ est supposé constant, en modifiant la constante K, on obtient l'inégalité de stabilité forte :

$$\sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2 + \frac{\alpha}{1 + \alpha \Delta t} \|e^{-\alpha n \Delta t} U\|_{\Delta x, \Delta t}^2 \leqslant K \sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} g_j\|_{\Delta t}^2.$$

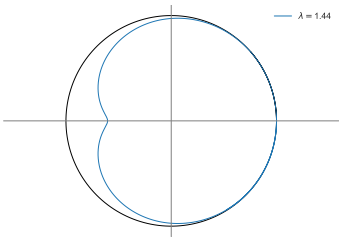
□

Dans toute la suite, on va donc uniquement se concentrer sur la résolution de (3.2) et la mise en place de l'inégalité (3.3).

3.2 Analyse de l'équation intérieure

Dans toute cette section, on veut étudier les solutions de l'équation intérieure (3.2a) qui sont dans ℓ^2. On ne s'occupe de la condition de bord (3.2b) qu'à partir de la Section 3.3.

On cherche donc à trouver les solutions $(\widetilde{U}_j(z))_j$ dans ℓ^2 de (3.2a) qui est une suite récurrente linéaire d'ordre $p + r$.



3.2.1 Équation caractéristique

Pour exprimer les solutions de la suite récurrente linéaire (3.2a), on veut résoudre l'équation caractéristique suivante :

$$z\kappa^r = \sum_{k=-r}^{p} a_k \kappa^{r+k} \tag{3.4}$$

où κ est l'inconnue et z est un paramètre. Dans toute la suite, on note $\kappa_1(z), \dots, \kappa_M(z)$ respectivement de multiplicité $\beta_1, \dots, \beta_M$ les racines de (3.4) se trouvant dans le disque unité ouvert. Parfois, on note aussi $(\kappa_k(z))_{k=1}^{r}$ les racines dans le disque unité ouvert avec redondance, enfin on note $(\kappa_k(z))_{k=1}^{r+p}$ les racines avec redondance. On omet parfois la dépendance en z pour faciliter la lecture de certaines expressions.

Les solutions de (3.4) sont alors de la forme suivante :

$$\forall j \geqslant -r, \quad \widetilde{U}_j(z) = \sum_{|\kappa|<1} P_\kappa(j)\kappa^j + \sum_{|\kappa|=1} Q_\kappa(j)\kappa^j + \sum_{|\kappa|>1} R_\kappa(j)\kappa^j, \tag{3.5}$$

où chaque polynôme P_κ, Q_κ et R_κ est de degré $\beta - 1$ pour κ de multiplicité β dans (3.4).

L'équation caractéristique (5.11) possède $p + r$ racines κ (comptées avec multiplicité) donc

$$\sum_{|\kappa|<1} (\deg P_\kappa + 1) + \sum_{|\kappa|=1} (\deg Q_\kappa + 1) + \sum_{|\kappa|>1} (\deg R_\kappa + 1) = p + r.$$

Pour faciliter la lecture, on n'a pas écrit la dépendance des κ en z, mais il faut bien comprendre que chaque racine κ de (3.4) dépend du paramètre z.

Exemple 3.3. Si l'équation caractéristique (3.4) possède quatre racines κ telles que κ_1 est de multiplicité 2 et est dans le disque unité ouvert $\mathbb{D}$, κ_2 est de multiplicité 1 et est dans le disque unité ouvert $\mathbb{D}$, κ_3 est de multiplicité 3 et est sur le cercle unité $\mathbb{S}$ et κ_4 est de multiplicité 1 et est dans $\mathcal{U}$, alors les solutions de (3.4) s'écrivent

$$\forall j \geqslant -r, \quad \widetilde{U}_j(z) = (\alpha_1 + j\alpha_2)\kappa_1^j + \alpha_3\kappa_2^j + (\alpha_4 + j\alpha_5 + j^2\alpha_6)\kappa_3^j + \alpha_7\kappa_4^j.$$

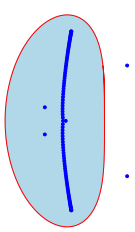
Avec les notations de (3.5), on a $P_{\kappa_1}(X) = \alpha_1 + \alpha_2 X$, $P_{\kappa_2}(X) = \alpha_3$, $Q_{\kappa_3}(X) = \alpha_4 + \alpha_5 X + \alpha_6 X^2$ et $R_{\kappa_4}(X) = \alpha_7$.

Afin de savoir si la solution $(\widetilde{U}_j(z))_j$ est dans ℓ^2, on veut localiser les racines $(\kappa_k(z))_{k=1}^{r+p}$ de (3.4). C'est l'objet de la sous-section suivante.

3.2.2 Lemme de Hersh

Le lemme de Hersh est introduit par Hersh dans [Her63] pour le cas continu. Ce lemme permet de localiser les racines de (3.4).

Lemme 3.4 (Hersh)**.** *Si le schéma* (3.1a) *est Cauchy-stable et si* $|z| > 1$*, alors l'équation caractéristique* (3.4) *:*



- *ne possède pas de racine κ sur le cercle unité $\mathbb{S}$.*
- *possède r racines (comptées avec multiplicité) dans le disque unité ouvert $\mathbb{D}$ et p racines (comptées avec multiplicité) dans $\mathcal{U}$.*

Pour que la solution $(\widetilde{U}_j(z))_j$ soit dans ℓ^2, il faut que les polynômes R_κ de l'équation (3.5) soient tous nuls. De plus, par le Lemme 3.4 (Hersh), on sait que pour $|z| > 1$ la somme sur les $|\kappa| = 1$ de l'équation (3.5) est vide. Ainsi, pour $|z| > 1$, les solutions $(\widetilde{U}_j(z))_j$ de (3.2a), qui sont dans ℓ^2, sont de la forme suivante :

$$\forall j \geqslant -r, \quad \widetilde{U}_j(z) = \sum_{k=1}^{M} P_{\kappa_k(z)}(j)\kappa_k(z)^j \tag{3.6}$$

où M est le nombre de racines distinctes de (3.4) dans $\mathbb{D}$ et $P_{\kappa_k(z)}$ est de degré $\beta_k - 1$ où β_k est la multiplicité de $\kappa_k(z)$ dans (3.4). On a, par le Lemme 3.4 (Hersh), que $\sum_{k=1}^{M} \beta_k = r$.

Démonstration du Lemme 3.4. Pour le premier point, par l'absurde, si $\kappa = e^{i\xi} \in \mathbb{S}$ est racine de (3.4) pour $|z| > 1$, on a alors

$$|z| = \left| \sum_{k=-r}^{p} a_k e^{ik\xi} \right| \leqslant 1$$

par Cauchy-stabilité du schéma, ce qui contredit le fait que $|z| > 1$.

Pour le deuxième point, il suffit de compter le nombre de racines dans le disque unité ouvert afin d'en déduire le nombre de racines dans $\mathcal{U}$ grâce au premier point.

Par continuité des racines d'un polynôme par rapport à ces coefficients et comme il n'y a aucune racine sur le cercle unité, on sait que, pour tout $z \in \mathcal{U}$, il y a un nombre constant fixé de racines dans le disque unité ouvert. Il suffit donc de compter le nombre de racines κ dans le disque unité ouvert pour un certain z qui vérifie $|z| > 1$ afin d'avoir le résultat sur tout le domaine $\mathcal{U}$.

On va utiliser le théorème de Rouché (Théorème A.5, page 202) pour compter le nombre de racines de (3.4) dans le disque unité ouvert $\mathbb{D}$. Avec les notations du Théorème A.5, on pose $B(a, r) = \mathbb{D}$, $\Gamma : \theta \in [0, 2\pi] \mapsto e^{i\theta}$,

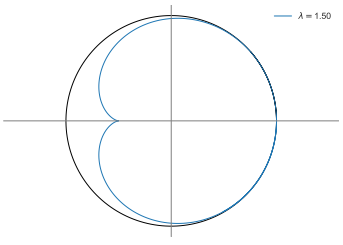
$$f(\kappa) = \kappa^r - \frac{1}{z}(a_{-r} + a_{-r+1}\kappa + \cdots + a_p\kappa^{p+r}) \quad \text{et} \quad g(\kappa) = \kappa^r - \frac{1}{z}a_{-r}.$$

Étape 1 : *Montrons que, pour $|z|$ suffisamment grand, on a $|g(\kappa)| > \frac{1}{2}$ sur $\mathbb{S}$.*

Pour $\kappa \in \mathbb{S}$, on a

$$|g(\kappa)| \geqslant \left| |\kappa^r| - \frac{|a_{-r}|}{|z|} \right| \geqslant \left| 1 - \frac{|a_{-r}|}{|z|} \right|.$$

Ainsi, pour $|z| > 2|a_{-r}|$, on a l'inégalité souhaitée.



Étape 2 : *Montrons que, pour $|z|$ suffisamment grand, on a $|f(\kappa) - g(\kappa)| \leqslant \frac{1}{2}$ sur $\mathbb{S}$.*

Pour $\kappa \in \mathbb{S}$, on a

$$|f(\kappa) - g(\kappa)| = \left|\frac{1}{z}(a_{-r+1}\kappa + \cdots + a_p\kappa^{p+r})\right| \leqslant \frac{1}{|z|}(|a_{-r+1}||\kappa| + \cdots + |a_p||\kappa|^{p+r}) \leqslant \frac{1}{|z|}\sum_{j=-r+1}^{p}|a_j|.$$

Ainsi, pour $|z| \geqslant 2\displaystyle\sum_{j=-r+1}^{p}|a_j|$, on a l'inégalité souhaitée.

Étape 3 : *Comptons le nombre de zéros de g dans $\mathbb{D}$ pour $|z|$ suffisamment grand.*

Par le théorème de D'Alembert–Gauss, on sait que le polynôme g possède r racines comptées avec multiplicité. Soit κ une racine de g, on a donc

$$\kappa^r = \frac{1}{z}a_{-r}.$$

Pour $|z| \geqslant 2\displaystyle\sum_{j=-r}^{p}|a_j|$, on a

$$|\kappa|^r = \frac{1}{|z|}|a_{-r}| < 1.$$

Ainsi, $\kappa \in \mathbb{D}$. Donc, pour $|z| \geqslant 2\displaystyle\sum_{j=-r}^{p}|a_j|$, les r racines de g sont dans $\mathbb{D}$.

Étape 4 : *Conclusion.*

Pour $|z| \geqslant 2\sum_{j=-r}^{p}|a_j|$, grâce aux étapes 1 et 2, on a

$$|f(\kappa) - g(\kappa)| \leqslant \frac{1}{2} < |g(\kappa)| \quad \text{pour } \kappa \in \mathbb{S}.$$

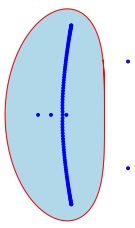
Donc, par le Théorème A.5 (Rouché), le nombre de zéros (compté avec multiplicité) de f dans le disque unité ouvert est égal à r (grâce à l'étape 3).

Ainsi, pour tout $z \in \mathcal{U}$, l'équation (3.4) possède r racines $\mathbb{D}$ et donc p racines dans $\mathcal{U}$. □

Les hypothèses de ce lemme peuvent être relaxées. En effet, on peut simplement supposer que z est dans la composante connexe infinie de $\mathbb{C} \setminus \Gamma$ où Γ est la courbe du symbole définie à la Définition 1.14 (page 32). En effet, par l'absurde, on a alors

$$\underbrace{z}_{\notin\Gamma} = \underbrace{\sum_{k=-r}^{p} a_k e^{ik\xi}}_{\in\Gamma},$$

ce qui est absurde. La preuve pour compter les racines dans $\mathbb{D}$ est la même que celle ci-dessus. La Cauchy-stabilité impose que la courbe Γ du symbole soit dans $\overline{\mathbb{D}}$, ainsi, pour $|z| > 1$, on est



dans la composante connexe infinie de $\mathbb{C} \setminus \Gamma$. On présente cette version plus générale dans les deux Lemmes 5.4 (page 109) et 6.3 (page 147).

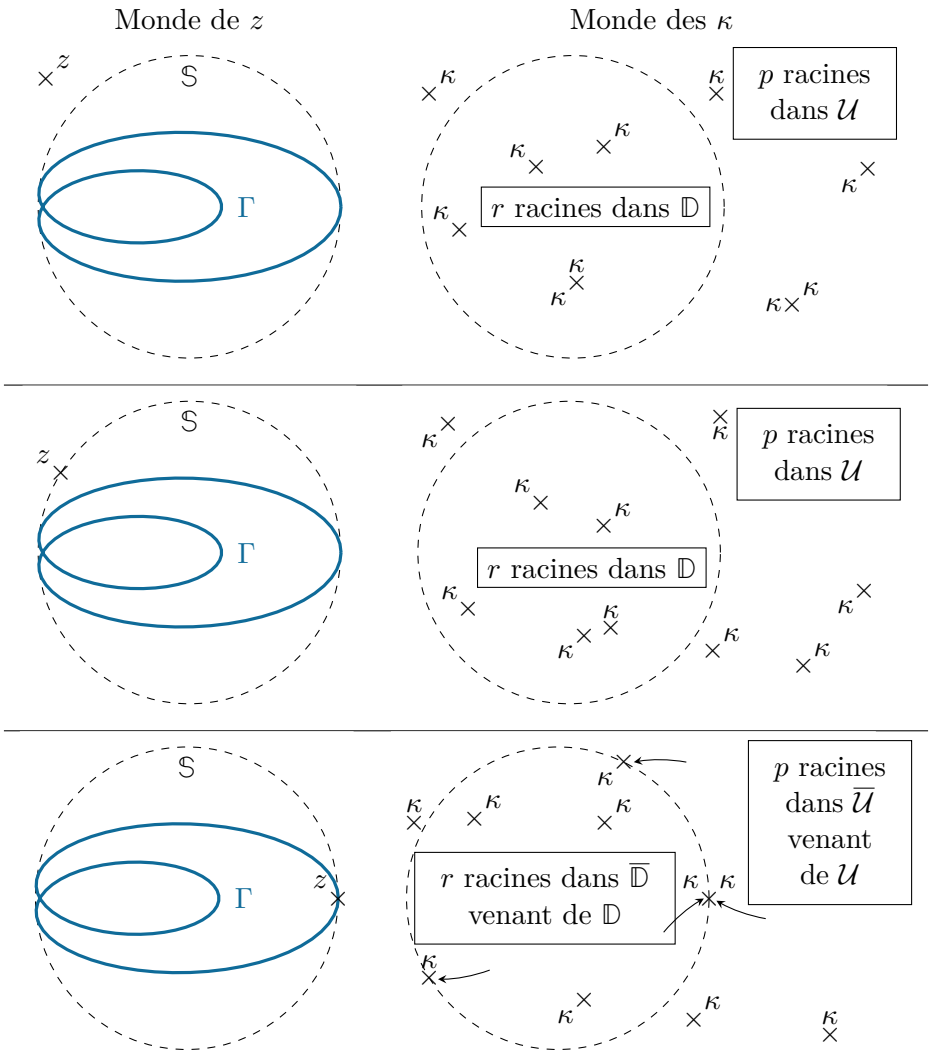
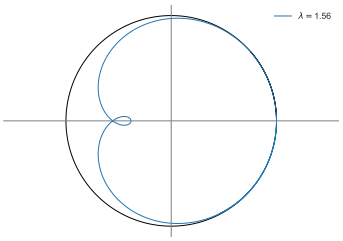


FIGURE 3.1 – Illustration du Lemme 3.4 : cas $|z| > 1$ (première ligne), cas $|z| = 1$ et $z \notin \Gamma$ (deuxième ligne) et cas $z \in \Gamma$ où le Lemme 3.4 ne s'applique pas (troisième ligne).

Dans la Figure 3.1, la première ligne illustre le Lemme 3.4 (Hersh), la deuxième illustre la relaxation d'hypothèse que l'on vient de mentionner ci-dessus et la troisième ligne illustre un cas où les hypothèses du Lemme 3.4 (Hersh) ne sont pas réunies et où il y a potentiellement des racines de (3.4) sur le cercle unité $\mathbb{S}$.

Proposition 3.5. *Sous hypothèse de Cauchy-stabilité, si z_0 est une valeur propre généralisée, alors $z_0 \in \mathbb{S} \cap \Gamma$.*

Démonstration. Par définition d'une valeur propre généralisée (Définition 1.21, page 36), on a $z_0 \in \mathbb{S}$.



Par l'absurde, si $z_0 \notin \Gamma$, par Cauchy-stabilité et comme $z_0 \in \mathbb{S}$, on sait que z_0 est dans la composante connexe infinie de $\mathbb{C} \setminus \Gamma$, donc le Lemme 3.4 (Hersh) s'applique et par l'expression (3.6) des solutions $(\widetilde{U}_j(z_0))_j$, on a $(\widetilde{U}_j(z_0))_j \in \ell^2$, donc z_0 ne peut pas être une valeur propre généralisée. □

3.2.3 Espace vectoriel des solutions dans ℓ^2 de l'équation intérieure

Pour tout $z \in \mathcal{U}$, on a vu en (3.6) qu'on peut écrire les solutions de (3.2a), qui sont dans ℓ^2, sous la forme :

$$\forall j \geqslant -r, \quad \widetilde{U}_j(z) = \sum_{k=1}^{M} P_{\kappa_k(z)}(j)\kappa_k(z)^j \tag{3.7}$$

où les $(\kappa_k)_{k=1}^M$ sont les r racines (avec multiplicité) de l'équation caractéristique (3.4) dans $\mathbb{D}$.

On introduit alors le sous-espace vectoriel $\mathcal{E}^s(z)$ de ℓ^2 des solutions de (3.2a) qui sont dans ℓ^2. La notation $\mathcal{E}^s(z)$ est la même que celle de Coulombel dans [Cou13]. Par le Lemme 3.4 (Hersh), l'espace $\mathcal{E}^s(z)$ est de dimension r.

Pour écrire l'expression (3.7) de la solution $(\widetilde{U}_j(z))_j \in \mathcal{E}^s(z)$, on a utilisé la base composée des r vecteurs suivants.

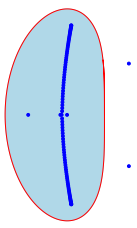
$$\begin{pmatrix} \kappa_k^{-r} \\ \vdots \\ \kappa_k^{-1} \\ 1 \\ \kappa_k \\ \kappa_k^2 \\ \kappa_k^3 \\ \vdots \end{pmatrix}, \begin{pmatrix} -r\kappa_k^{-r} \\ \vdots \\ -\kappa_k^{-1} \\ 1 \\ \kappa_k \\ 2\kappa_k^2 \\ 3\kappa_k^3 \\ \vdots \end{pmatrix}, \begin{pmatrix} r^2\kappa_k^{-r} \\ \vdots \\ \kappa_k^{-1} \\ 0 \\ \kappa_k \\ 4\kappa_k^2 \\ 9\kappa_k^3 \\ \vdots \end{pmatrix}, \ldots, \begin{pmatrix} (-r)^{\beta_k-1}\kappa_k^{-r} \\ \vdots \\ (-1)^{\beta_k-1}\kappa_k^{-1} \\ 0 \\ \kappa_k \\ 2^{\beta_k-1}\kappa_k^2 \\ 3^{\beta_k-1}\kappa_k^3 \\ \vdots \end{pmatrix}, \quad k = 1, \ldots, M. \tag{3.8}$$

Notation 3.6. Soit $-r \leqslant i < j < \infty$. On introduit la matrice $K_{i,j}(z) \in \mathcal{M}_{j-i+1,r}(\mathbb{C})$ qui contient l'extraction des lignes i à j (comprises) des r vecteurs de la base (3.8).

Exemple 3.7. Soient $r = 2$ et $z \in \mathcal{U}$. Si les deux racines présentes dans $\mathbb{D}$ de (3.4) sont distinctes $\kappa_1(z) \neq \kappa_2(z)$, alors par exemple la matrice $K_{-2,1}(z)$ s'écrit de la manière suivante :

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

Si les deux racines dans $\mathbb{D}$ de (3.4) sont en fait qu'une racine double $\kappa(z)$, alors par exemple



la matrice $K_{0,3}(z)$ s'écrit de la manière suivante :

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

On remarque, par ces deux derniers exemples, que la base (3.8) et les matrices $K_{i,j}(z)$ ne sont pas continues en z. On discute plus en détails de cela dans les Section 5.2.1 (page 109) et Section 6.2.1 (page 147).

La représentation de $\mathcal{E}^s(z)$ choisie avec la base (3.8) n'a pas de bonnes propriétés de régularité, cependant l'espace $\mathcal{E}^s(z)$ possède de bonnes propriétés de régularité. En effet, Coulombel démontre dans [Cou13, Th. 4.3] que $\mathcal{E}^s(z)$ a une structure de fibré vectoriel holomorphe sur $\mathcal{U}$. La preuve s'inspire du cas continu étudié par Métivier et Zumbrun [Mét00, MZ04]. On peut trouver des précisions sur la notion de fibré vectoriel holomorphe dans le livre de Lang [Lan95].

Dans la suite, on va avoir besoin d'obtenir des estimations sur $\overline{\mathcal{U}}$, pour cela, il va falloir étendre l'espace $\mathcal{E}^s(z)$ sur le cercle unité.

Par continuité des racines $\kappa(z)$ de (3.4) par rapport à z, on peut prolonger les racines $(\kappa_k(z))_{k=1}^{p+r}$ de (3.4) pour $z \in \mathbb{S}$. Cependant, pour prolonger la base (3.8) et la notation $K_{i,j}(z)$, il faut sélectionner les racines issues des r racines $(\kappa_k(z))_{k=1}^{r}$ venant du disque unité ouvert quand $|z| > 1$. En effet, comme le montre la troisième ligne de la Figure 3.1, les racines κ de (3.4) peuvent être sur le cercle $\mathbb{S}$ et il peut y avoir des racines multiples dont une partie de la multiplicité vient de $\mathbb{D}$ et l'autre vient de $\mathcal{U}$.

Dans les simulations numériques, on est obligé de sélectionner les racines venant de $\mathbb{D}$, on donne une stratégie numérique pour les sélectionner en Section 7.1.1 (page 177).

Dans [Cou13, Th. 4.3], Coulombel a justifié que l'on peut étendre l'espace $\mathcal{E}^s(z)$ de manière continue sur $\mathbb{S}$ et cela de manière unique. On résume cela dans l'énoncé suivant.

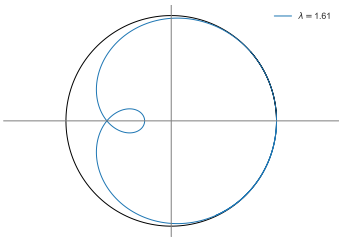
Théorème 3.8 ([Cou13])**.** *On suppose que le schéma est Cauchy-stable. L'espace $\mathcal{E}^s(z)$ est un fibré vectoriel holomorphe sur $\mathcal{U}$ et peut être étendu continûment sur $\overline{\mathcal{U}}$ et cela d'une unique manière.*

On prolonge alors la notation $K_{i,j}(z)$ de la Notation 3.6 sur $\overline{\mathcal{U}}$ tout entier.

3.3 Déterminant de Kreiss–Lopatinskii

3.3.1 Intégration des conditions de bord

En injectant les solutions $(\widetilde{U}_j(z))_j$ de la forme (3.7) de l'équation intérieure (3.2a) dans l'équation de bord (3.2b), on obtient alors un système linéaire de r équations à r inconnues. Ce



système peut s'écrire sous la forme

$$\overset{r}{\updownarrow}\underbrace{\begin{pmatrix} 1 & & 0 & \\ & \ddots & & -B \\ 0 & & 1 & \end{pmatrix}}_{\overset{r}{\longleftrightarrow}\quad\overset{m}{\longleftrightarrow}} K_{-r,m-1}(z)\begin{pmatrix}\alpha_1(z)\\ \vdots \\ \alpha_r(z)\end{pmatrix} = \begin{pmatrix}\widetilde{g}_{-r}(z)\\ \vdots \\ \widetilde{g}_{-1}(z)\end{pmatrix} \tag{3.9}$$

où $B \in \mathcal{M}_{r,m}(\mathbb{R})$ est la matrice définie en (1.14) (page 37) et les r inconnues $\alpha_1(z), \dots, \alpha_r(z)$ sont les coefficients de $(\widetilde{U}_j(z))_j$ écrits dans la base (3.8) de $\mathcal{E}^s(z)$.

Pour que ce système soit inversible, il faut vérifier que le déterminant du système (3.9) ne s'annule pas sur tout $\overline{\mathcal{U}}$. En effet, comme on va le voir à la Section 3.3.4, on veut une minoration uniforme du module du déterminant sur $\mathcal{U}$, donc on a besoin de regarder ces points d'annulation aussi sur $\mathbb{S}$. Ce déterminant est appelé *déterminant de Kreiss–Lopatinskii*, il est utilisé dans [GKS72] et plus récemment dans [Gus08] et [GKO13].

Définition 3.9 (Déterminant de Kreiss–Lopatinskii)**.** Le *déterminant de Kreiss–Lopatinskii*, noté Δ_{KL}, est une fonction à valeurs complexes définie, pour tout $z \in \overline{\mathcal{U}}$, par

$$\Delta_{\mathrm{KL}}(z) \overset{\text{def}}{=} \det(K_{-r,-1}(z) - BK_{0,m-1}(z)). \tag{3.10}$$

Pour obtenir l'équation (3.10), on a simplement effectué un produit par blocs dans l'équation (3.9). Cette vision du déterminant est celle utilisée dans le Chapitre 5 dédié à l'article [BLBS23a].

Dans la Définition 3.9, on a choisi une base arbitraire de l'espace $\mathcal{E}^s(z)$ à travers la notation $K_{-r,m-1}(z)$. On peut en fait définir de manière plus informelle un déterminant de Kreiss–Lopatinskii de la manière suivante :

$$\Delta_{\mathrm{KL}}(z) = \det(\mathfrak{B}e_1(z), \dots, \mathfrak{B}e_r(z))$$

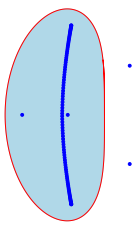
où $\mathfrak{B}$ est un encodage de la condition de bord et $e_1(z), \dots, e_r(z)$ est une base de $\mathcal{E}^s(z)$.

3.3.2 Formulation alternative du déterminant de Kreiss–Lopatinskii

Si, au lieu d'utiliser la formulation (3.1) pour le schéma, on utilise la formulation (1.15) décrite dans la Section 1.5.1 (page 37), le bord est encodé par l'expression (1.16).

En passant (1.16) à la transformée en $\mathcal{Z}$, on obtient alors l'expression suivante :

$$z\begin{pmatrix}\widetilde{U}_0(z)\\ \vdots \\ \widetilde{U}_{r-1}(z)\end{pmatrix} - \mathcal{B}\begin{pmatrix}\widetilde{U}_0(z)\\ \vdots \\ \widetilde{U}_{m-1}(z)\end{pmatrix} = \begin{pmatrix}\widetilde{\mathcal{G}}_0(z)\\ \vdots \\ \widetilde{\mathcal{G}}_{r-1}(z)\end{pmatrix}. \tag{3.11}$$



Dans cette vision du schéma, on travaille dans $\ell^2(\mathbb{N})$ et non plus dans $\ell^2 \overset{\text{def}}{=} \ell^2(\llbracket -r : -1 \rrbracket \cup \mathbb{N})$, il faut donc adapter les définitions en faisant partir les sommes à $j = 0$ et en translatant les indices $\llbracket -r : -1 \rrbracket$ à $\llbracket 0 : r-1 \rrbracket$.

En injectant les solutions $(\widetilde{U}_j(z))_j$ de (3.2a) qui sont de la forme (3.7) dans l'équation (3.11), on obtient alors le système suivant :

$$(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z)) \begin{pmatrix} \alpha_1(z) \\ \vdots \\ \alpha_r(z) \end{pmatrix} = \begin{pmatrix} \widetilde{\mathcal{G}}_0(z) \\ \vdots \\ \widetilde{\mathcal{G}}_{r-1}(z) \end{pmatrix}. \tag{3.12}$$

De cette équation, on en déduit une autre formulation pour le déterminant de Kreiss–Lopatinskii.

Définition 3.10. Une deuxième formulation du déterminant de Kreiss–Lopatinskii (3.10) est définie, pour tout $z \in \overline{\mathcal{U}}$, par

$$\Delta_{\mathrm{KL}}^{(2)}(z) \overset{\text{def}}{=} \det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z)). \tag{3.13}$$

Cette formulation du déterminant est celle utilisée dans le Chapitre 6 dédié à l'article [BLBS23b]. On justifie dans la proposition suivante l'utilisation de manière indifférente de $\Delta_{\mathrm{KL}}(z)$ ou de $\Delta_{\mathrm{KL}}^{(2)}(z)$.

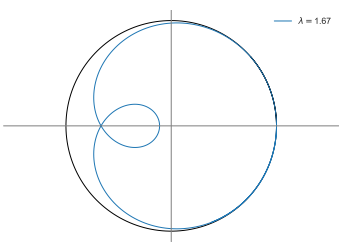
Proposition 3.11. *Pour tout $z \in \overline{\mathcal{U}}$, on a*

$$\Delta_{\mathrm{KL}}^{(2)}(z) = a_{-r}^r \Delta_{\mathrm{KL}}(z). \tag{3.14}$$

Comme par hypothèse du schéma, a_{-r} est non nul et est fixé, $\Delta_{\mathrm{KL}}(z)$ et $\Delta_{\mathrm{KL}}^{(2)}(z)$ ont les mêmes propriétés de régularité et d'annulation.

Démonstration. On rappelle l'équation (1.17) qui fait le lien entre B et $\mathcal{B}$:

$$\mathcal{B} = \underbrace{\begin{pmatrix} a_{-r} & \cdots & a_{-1} \\ & \ddots & \vdots \\ 0 & & a_{-r} \end{pmatrix}}_{\overset{\text{def}}{=} A_1} B + \underbrace{\begin{pmatrix} a_0 & \cdots & a_p & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & & \ddots & \ddots & & & \vdots \\ a_{-r+1} & \cdots & a_0 & \cdots & a_p & 0 & \cdots & 0 \end{pmatrix}}_{\overset{\text{def}}{=} A_2}.$$



On remarque aussi que par l'équation intérieure du schéma, pour tout $z \in \overline{\mathcal{U}}$, on a

$$\begin{aligned} z\begin{pmatrix} \widetilde{U}_0(z) \\ \vdots \\ \widetilde{U}_{r-1}(z) \end{pmatrix} &= \sum_{k=-r}^{p} a_k \begin{pmatrix} \widetilde{U}_k(z) \\ \vdots \\ \widetilde{U}_{r-1+k}(z) \end{pmatrix} \\ &= A_1 \begin{pmatrix} \widetilde{U}_{-r}(z) \\ \vdots \\ \widetilde{U}_{-1}(z) \end{pmatrix} + \begin{pmatrix} a_0 & \cdots & a_p & & 0 \\ \vdots & \ddots & & \ddots & \\ a_{-r+1} & \cdots & a_0 & \cdots & a_p \end{pmatrix} \begin{pmatrix} \widetilde{U}_0(z) \\ \vdots \\ \widetilde{U}_{r+p-1}(z) \end{pmatrix} \\ &= A_1 \begin{pmatrix} \widetilde{U}_{-r}(z) \\ \vdots \\ \widetilde{U}_{-1}(z) \end{pmatrix} + A_2 \begin{pmatrix} \widetilde{U}_0(z) \\ \vdots \\ \widetilde{U}_{m-1}(z) \end{pmatrix}. \end{aligned}$$

Ainsi, en utilisant la notation $K_{i,j}(z)$, on obtient

$$\begin{aligned} \Delta_{\mathrm{KL}}^{(2)}(z) &= \det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z)) \\ &= \det(zK_{0,r-1}(z) - (A_1B + A_2)K_{0,m-1}(z)) \\ &= \det(A_1K_{-r,-1}(z) + A_2K_{0,m-1}(z) - (A_1B + A_2)K_{0,m-1}(z)) \\ &= \det(A_1(K_{-r,-1}(z) - BK_{0,m-1}(z))) \\ &= \det(A_1)\det(K_{-r,-1}(z) - BK_{0,m-1}(z)) = a_{-r}^r \Delta_{\mathrm{KL}}(z). \end{aligned}$$

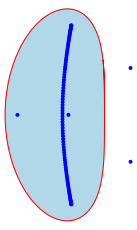
□

3.3.3 Introduction du déterminant intrinsèque de Kreiss–Lopatinskii

On souhaite que le déterminant de Kreiss–Lopatinskii possède les mêmes propriétés de régularité que l'espace $\mathcal{E}^s(z)$. Cependant, comme les racines $(\kappa_j(z))_{j=1}^{r+p}$ sont des racines d'un polynôme à coefficients holomorphes (3.4), il y a peu de chance qu'elles soient holomorphes, comme en témoigne l'exemple du polynôme $X^2 - z$. Ainsi, les coefficients des matrices $K_{i,j}(z)$ ne sont pas holomorphes. Néanmoins, on sait que les fonctions symétriques des racines $(\kappa_j(z))_{j=1}^{r+p}$ sont holomorphes grâce aux relations coefficients/racines des polynômes. Dans l'écriture de $K_{i,j}(z)$, on a imposé un ordre arbitraire. Ainsi, afin de rendre l'expression Δ_{KL} plus symétrique en les κ, on va diviser par la quantité $\det K_{0,r-1}(z)$ qui utilise la même permutation des racines κ. Cette quantité est non nulle car les colonnes de $K_{0,r-1}(z)$ forment une base.

Définition 3.12 (Déterminant intrinsèque de Kreiss–Lopatinskii)**.** Le *déterminant intrinsèque de Kreiss–Lopatinskii* Δ est défini, pour tout $z \in \overline{\mathcal{U}}$, par

$$\Delta(z) \overset{\text{def}}{=} \frac{\det(K_{-r,-1}(z) - BK_{0,m-1}(z))}{\det K_{0,r-1}(z)}.$$



De la même façon que pour le déterminant de Kreiss–Lopatinskii, on peut définir le déterminant intrinsèque de Kreiss–Lopatinskii dans la deuxième formulation (3.13).

Proposition 3.13 (Deuxième formulation du déterminant intrinsèque de Kreiss–Lopatinskii)**.** *Le déterminant intrinsèque de Kreiss–Lopatinskii peut aussi être défini, à un facteur a_{-r}^r près, par*

$$\forall z \in \overline{\mathcal{U}}, \quad \Delta(z) = \frac{\det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z))}{\det K_{0,r-1}(z)}. \tag{3.15}$$

L'étude de ce déterminant, notamment ses propriétés de régularité, sera traitée dans les deux Chapitres 5 et 6.

3.3.4 Propriétés du déterminant

Proposition 3.14. *Les deux fonctions Δ_{KL} et Δ ont les mêmes zéros.*

Démonstration. Cela vient du fait que $\Delta(z) = \frac{\Delta_{\mathrm{KL}}(z)}{\det K_{0,r-1}(z)}$. □

Les deux théorèmes suivants font partie des principaux résultats de [BLBS23b], on en fera les preuves complètes dans le Chapitre 6 à la Section 6.3 (page 153).

Théorème 3.15. *L'application $z \mapsto K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$ est holomorphe sur $\mathcal{U}$, continue sur $\overline{\mathcal{U}}$ et est bornée sur $\overline{\mathcal{U}}$.*

Théorème 3.16. *L'application $z \mapsto \Delta(z)$ est holomorphe sur $\mathcal{U}$, continue sur $\overline{\mathcal{U}}$. De plus, le déterminant intrinsèque de Kreiss–Lopatinskii vérifie :*

$$|\Delta(z)| \underset{|z|\to\infty}{\sim} |z|^r.$$

La preuve de ce résultat vient du fait qu'on peut écrire le déterminant intrinsèque de Kreiss–Lopatinskii (3.15) de la manière suivante :

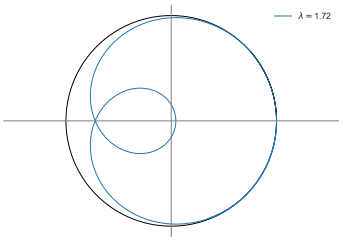
$$\Delta(z) = \det\left(zI_r - \mathcal{B}K_{0,m-1}(z)K_{0,r-1}^{-1}(z)\right) = z^r \det\left(I_r - \frac{1}{z}\mathcal{B}K_{0,m-1}(z)K_{0,r-1}^{-1}(z)\right).$$

Proposition 3.17. *Soit $R > 1$. Les deux conditions suivantes sont équivalentes :*

(i) Pour tout $z \in \{1 \leqslant |z| \leqslant R\}$, on a $|\Delta(z)| \neq 0$.

(ii) Il existe $K > 0$ tel que, pour tout $z \in \{1 < |z| \leqslant R\}$, on a $|\Delta(z)| \geqslant K$.

Démonstration. $\boxed{(i) \implies (ii)}$ L'application $z \in \{1 \leqslant |z| \leqslant R\} \mapsto |\Delta(z)| \in]0, +\infty[$ est continue sur le compact $\{1 \leqslant |z| \leqslant R\}$ par le Théorème 3.16, donc atteint ses bornes ; sa borne inférieure donne la constante K.
$\boxed{(ii) \implies (i)}$ On sait que pour $z \in \{1 < |z| \leqslant R\}$, on a $|\Delta(z)| \neq 0$. Pour tout z tel que $|z| = 1$, il existe une suite $z_n \to z$ avec $z_n \in \{1 < |z| \leqslant R\}$ pour tout $n \in \mathbb{N}$ et $|\Delta(z_n)| \geqslant K$ pour tout $n \in \mathbb{N}$. On a le résultat en passant à la limite. □



Proposition 3.18. *Si $\Delta(z) \neq 0$ pour tout $z \in \overline{\mathcal{U}}$, alors il existe une constante $c > 0$ telle que*

$$\forall z \in \mathcal{U}, \quad |\Delta(z)| \geqslant c.$$

Démonstration. Par le Théorème 3.16, on a $|\Delta(z)| \xrightarrow[|z|\to\infty]{} +\infty$. Ainsi, il existe $R > 0$ tel que pour tout $|z| > R$, on a $|\Delta(z)| > 1$. De plus, $\Delta(z) \neq 0$ sur $\{1 \leqslant |z| \leqslant R\}$, donc, par la Proposition 3.17, il existe $K > 0$ tel que $|\Delta(z)| \geqslant K$ pour tout $z \in \{1 \leqslant |z| \leqslant R\}$. Ainsi, en posant $c = \min(K, 1) > 0$, on a le résultat voulu. □

3.4 Condition de Kreiss–Lopatinskii uniforme

On veut montrer que la minoration du déterminant intrinsèque de Kreiss–Lopatinskii permet d'obtenir la formulation résolvante (3.3) de la stabilité forte. Dans un premier temps, on va introduire la condition de Kreiss–Lopatinskii uniforme. Cette condition tient son nom du cas continu comme on peut le voir dans [BGS06]. Pour le cas discret, elle est étudiée dans les livres [Gus08] et [GKO13].

3.4.1 Définition

Definition 3.19 (Condition de Kreiss–Lopatinskii uniforme)**.** Un schéma de la forme (3.2) vérifie la *condition de Kreiss–Lopatinskii uniforme* s'il existe une constante K telle que pour tout $(\widetilde{g}_j(z))_j$, pour toute solution $(\widetilde{U}_j(z))_j$ de (3.2), on a

$$\forall z \in \mathcal{U}, \quad \sum_{j=-r}^{-1} |\widetilde{U_j}(z)|^2 \leqslant K \sum_{j=-r}^{-1} |\widetilde{g_j}(z)|^2. \tag{3.16}$$

3.4.2 Théorème de la couronne uniforme

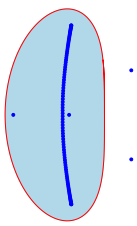
Le théorème suivant permet de localiser l'étude de la condition de Kreiss–Lopatinskii uniforme au voisinage du cercle unité. En effet, on peut traduire le théorème suivant par : pour un certain rang $R_0 > 1$, le déterminant de Kreiss–Lopatinskii ne s'annule pas sur $\{|z| > R_0\}$.

Théorème 3.20. *Il existe $R_0 > 1$ tel qu'il existe une constante K vérifiant : pour tout $(\widetilde{g}_j(z))_j$, pour toute solution $(\widetilde{U}_j(z))_j$ de (3.2), on a*

$$\forall |z| > R_0, \quad \sum_{j=-r}^{+\infty} |\widetilde{U_j}(z)|^2 \leqslant K \sum_{j=-r}^{-1} |\widetilde{g_j}(z)|^2. \tag{3.17}$$

La preuve suivante s'inspire de [Cou13, Lem. 3.3].

Démonstration. On munit l'espace $\ell^2 \overset{\text{def}}{=} \ell^2([\![-r : -1]\!] \cup \mathbb{N})$ de la norme $||W||^2 \overset{\text{def}}{=} \sum_{j=-r}^{+\infty} |W_j|^2$



et on pose les applications linéaires suivantes :

$$L(z) : W \in \ell^2 \mapsto L(z)W \in \ell^2 \quad \text{et} \quad L_\infty : W \in \ell^2 \mapsto L_\infty W \in \ell^2$$

où

$$(L(z)W)_j = \begin{cases} W_j - \sum_{k=-r}^{p} a_k \frac{1}{z} W_{k+j} & j \geqslant 0, \\ W_j - \sum_{k=0}^{m} b_{j,k} W_k & j \in [\![-r : -1]\!], \end{cases}$$

$$\text{et} \quad (L_\infty W)_j = \begin{cases} W_j & j \geqslant 0, \\ W_j - \sum_{k=0}^{m-1} b_{j,k} W_k & j \in [\![-r : -1]\!]. \end{cases}$$

On va utiliser la notation $|||\cdot|||$ pour les normes d'opérateurs. On remarque que l'application L_∞ est inversible car elle est linéaire, continue et bijective. Ainsi, on va pouvoir utiliser $|||L_\infty^{-1}|||$.

Étape 1 : *Contrôle de* $|||L_\infty - L(z)|||$.

On observe que

$$z((L_\infty - L(z))W)_j = \begin{cases} \sum_{k=-r}^{p} a_k W_{k+j} & j \geqslant 0, \\ 0 & j \in [\![-r : -1]\!]. \end{cases}$$

Ainsi,

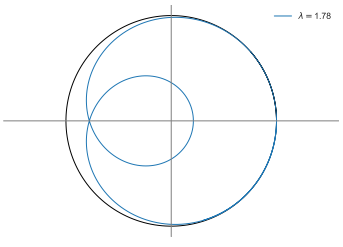
$$||z((L_\infty - L(z))W)||^2 = \sum_{j=0}^{+\infty} \left| \sum_{k=-r}^{p} a_k W_{k+j} \right|^2 \leqslant \sum_{j=0}^{+\infty} (p+r) \max_{k=-r}^{p} |a_k| \sum_{k=-r}^{p} |W_{k+j}|^2$$
$$\leqslant (p+r)^2 \max_{k=-r}^{p} |a_k| \sum_{j=-r}^{+\infty} |W_j|^2.$$

On pose alors la constante $C \overset{\text{def}}{=} (p+r)^2 \max_{k=-r}^{p} |a_k|$. On a donc

$$|||L_\infty - L(z)||| \leqslant \frac{C}{|z|}. \tag{3.18}$$

Étape 2 : *Montrons que si* $|||H||| < 1$*, alors* $I - H$ *est inversible.*

L'ensemble $\mathcal{L}(\ell^2)$ est un espace de Banach pour la norme $|||\cdot|||$. Ainsi, l'application $\sum_{n=0}^{+\infty} H^n$



est bien définie (car la série est normalement convergente). On a alors, pour tout $N \in \mathbb{N}$,

$$(I-H)\circ\sum_{n=0}^{N} H^n = I - \underbrace{H^{N+1}}_{\xrightarrow[N\to+\infty]{}0}$$

De même pour l'inverse à gauche. Ainsi, $I-H$ est inversible, d'inverse $\displaystyle\sum_{n=0}^{+\infty} H^n$.

Étape 3 : *Montrons que, pour tout $|z| > C|||L_\infty^{-1}|||$, $L(z)$ est inversible.*

On a

$$L(z) = L_\infty - (L_\infty - L(z)) = \underbrace{L_\infty}_{\text{inversible}}\ (I - L_\infty^{-1}(L_\infty - L(z))).$$

Or, pour tout $|z| > C|||L_\infty^{-1}|||$, par (3.18), on a

$$|||L_\infty^{-1}(L_\infty - L(z))||| \leqslant |||L_\infty^{-1}|||\ |||L_\infty - L(z)||| \leqslant |||L_\infty^{-1}|||\frac{C}{|z|} < 1.$$

Donc, par l'étape 2, l'application $I - L_\infty^{-1}(L_\infty - L(z))$ est inversible. L'étape 3 vient de montrer que pour tout $|z| > C|||L_\infty^{-1}|||$, les équations (3.2) admettent une unique solution dans ℓ^2.

Étape 4 : *Contrôle de la norme de l'inverse $L(z)$ pour $|z| > 2C|||L_\infty^{-1}|||$.*

Pour $|z| > 2C|||L_\infty^{-1}|||$, on a

$$\begin{aligned}
|||L(z)^{-1}||| &= |||(I - L_\infty^{-1}(L_\infty - L(z)))^{-1}L_\infty^{-1}||| \leqslant |||(I - L_\infty^{-1}(L_\infty - L(z)))^{-1}|||\ |||L_\infty^{-1}||| \\
&\leqslant \sum_{n=0}^{+\infty} |||L_\infty^{-1}(L_\infty - L(z))|||^n\ |||L_\infty^{-1}||| \leqslant \sum_{n=0}^{+\infty} \Big(|||L_\infty^{-1}|||\ |||L_\infty - L(z)|||\Big)^n\ |||L_\infty^{-1}||| \\
&\leqslant \sum_{n=0}^{+\infty} \left(|||L_\infty^{-1}|||\ \frac{C}{|z|}\right)^n\ |||L_\infty^{-1}||| \leqslant \sum_{n=0}^{+\infty} 2^{-n}\ |||L_\infty^{-1}||| \leqslant 2|||L_\infty^{-1}|||.
\end{aligned}$$

Étape 5 : *Conclusion.*

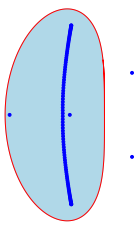
Ainsi, on pose $R_0 = 2C|||L_\infty^{-1}|||$. Soit une solution $(\widetilde{U}_j(z))_j$ de (3.2), on a

$$L(z)\widetilde{U}_j(z) = \begin{cases} 0 & j \geqslant 0, \\ \widetilde{g}_j(z) & j \in [\![-r:-1]\!]. \end{cases}$$

Donc, par l'étape 4, on a pour tout $|z| > R_0$, l'inégalité

$$\sum_{j=-r}^{+\infty} |\widetilde{U}_j(z)|^2 \leqslant 2|||L_\infty^{-1}||| \sum_{j=-r}^{-1} |\widetilde{g}_j(z)|^2.$$

□



Remarque 3.21. La constante R_0 du Théorème 3.20 a la forme explicite suivante :

$$R_0 \overset{\text{def}}{=} 2(p+r)^2 \max_k |a_k| \left(1 + (m+1)^2 \max_{j,k} |b_{j,k}|\right).$$

3.4.3 Lien avec le déterminant intrinsèque de Kreiss–Lopatinskii

Le théorème suivant permet de justifier que dans les Chapitres 5 et 6, on cherche à savoir combien de fois le déterminant intrinsèque de Kreiss–Lopatinskii s'annule sur $\overline{\mathcal{U}}$, puisque cela permet de vérifier si la condition de Kreiss–Lopatinskii uniforme est satisfaite.

Théorème 3.22. *Les deux assertions suivantes sont équivalentes :*

(i) le déterminant intrinsèque de Kreiss–Lopatinskii Δ ne s'annule pas sur $\overline{\mathcal{U}}$.

(ii) la condition de Kreiss–Lopatinskii uniforme (3.16) est satisfaite.

Ce théorème est présent dans les livres [Gus08] et [GKO13] mais démontré dans le cas semi-discret. On en donne ici une preuve dans le cas discret utilisant les notations introduites dans ce chapitre.

Démonstration. Dans cette preuve, on utilise la formulation (3.15) pour le déterminant intrinsèque de Kreiss–Lopatinskii.

$\boxed{(i) \implies (ii)}$ On rappelle que

$$z \begin{pmatrix} \widetilde{U}_0(z) \\ \vdots \\ \widetilde{U}_{r-1}(z) \end{pmatrix} = \mathcal{B} \begin{pmatrix} \widetilde{U}_0(z) \\ \vdots \\ \widetilde{U}_{m-1}(z) \end{pmatrix} + \underbrace{\begin{pmatrix} a_{-r} & \cdots & a_{-1} \\ & \ddots & \vdots \\ 0 & & a_{-r} \end{pmatrix}}_{\overset{\text{def}}{=} A_1} \begin{pmatrix} \widetilde{g}_{-r} \\ \vdots \\ \widetilde{g}_{-1} \end{pmatrix}.$$

On sait que

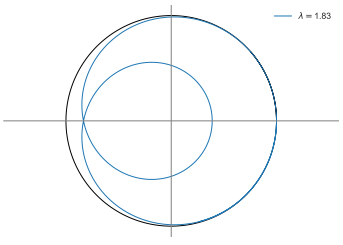
$$\Delta(z) = \frac{\det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z))}{\det K_{0,r-1}(z)} = \det((zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z))K_{0,r-1}^{-1}(z)).$$

De plus, on a

$$K_{0,r-1} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \end{pmatrix} = \begin{pmatrix} \widetilde{U}_0 \\ \vdots \\ \widetilde{U}_{r-1} \end{pmatrix} \text{ et } \begin{pmatrix} \widetilde{g}_{-r} \\ \vdots \\ \widetilde{g}_{-1} \end{pmatrix} = A_1^{-1}(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z)) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \end{pmatrix}. \tag{3.19}$$

Ainsi, on a

$$\begin{pmatrix} \widetilde{g}_{-r} \\ \vdots \\ \widetilde{g}_{-1} \end{pmatrix} = A_1^{-1}(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z))K_{0,r-1}^{-1}(z) \begin{pmatrix} \widetilde{U}_0 \\ \vdots \\ \widetilde{U}_{r-1} \end{pmatrix}.$$



En étudiant l'application $\phi_z : (\widetilde{U}_0, \dots, \widetilde{U}_{r-1}) \mapsto (\widetilde{g}_{-r}, \dots, \widetilde{g}_{-1})$, on a

$$\|\phi_z^{-1}\| \leqslant \|A_1\| \frac{\|\mathrm{Com}((zK_{0,r-1} - \mathcal{B}K_{0,m-1})K_{0,r-1}^{-1})\|}{|\Delta(z)|}.$$

Or, par le Théorème 3.15, la fonction $z \mapsto K_{0,m-1}K_{0,r-1}^{-1}$ est continue sur $\{1 \leqslant |z| \leqslant R_0\}$ où R_0 est la constante du Théorème 3.20. Ainsi, les normes du numérateurs sont majorées sur $\{1 \leqslant |z| \leqslant R_0\}$ et par la Proposition 3.18, il existe une constante $c > 0$ telle que pour tout $z \in \mathcal{U}$, on a $|\Delta(z)| \geqslant c$. Donc la quantité $\|\phi_z^{-1}\|$ est majorée par une constante uniformément sur $\{1 \leqslant |z| \leqslant R_0\}$. Appelons cette constante K_1. On a alors

$$\forall z \in \{1 < |z| \leqslant R_0\}, \quad \sum_{j=0}^{r-1} |\widetilde{U_j}(z)|^2 \leqslant K_1 \sum_{j=-r}^{-1} |\widetilde{g_j}(z)|^2. \tag{3.20}$$

Pour avoir un contrôle de $(\widetilde{U}_j(z))_{j=-r}^{-1}$, on veut étudier la matrice $K_{-r,-1}(z)K_{0,r-1}^{-1}(z)$ puisque

$$\begin{pmatrix} \widetilde{U}_{-r}(z) \\ \vdots \\ \widetilde{U}_{-1}(z) \end{pmatrix} = K_{-r,-1}(z)K_{0,r-1}^{-1}(z) \begin{pmatrix} \widetilde{U}_0(z) \\ \vdots \\ \widetilde{U}_{r-1}(z) \end{pmatrix}.$$

La preuve du Théorème 3.15, qui donne un contrôle uniforme sur $\|K_{0,m-1}(z)K_{0,r-1}^{-1}(z)\|$, peut être adaptée pour étudier la quantité $\|K_{-r,-1}(z)K_{0,r-1}^{-1}(z)\|$ et en donner un contrôle uniforme en z. Ainsi, il existe une constante K_2 telle que

$$\forall z \in \{1 < |z| \leqslant R_0\}, \quad \sum_{j=-r}^{-1} |\widetilde{U_j}(z)|^2 \leqslant K_2 \sum_{j=0}^{r-1} |\widetilde{U_j}(z)|^2. \tag{3.21}$$

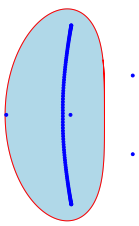
En utilisant les deux inégalités (3.20) et (3.21) et le théorème de la couronne uniforme (Théorème 3.20), on obtient la condition de Kreiss–Lopatinskii uniforme (3.16).

$\boxed{(ii) \implies (i)}$ Par contraposée, si Δ s'annule en un point $z \in \mathcal{U}$, alors Δ_{KL} s'annule aussi en z. Ainsi, pour $(\widetilde{g}_j)_j = 0$, le système (3.9) possède une solution non nulle. Il est alors impossible d'obtenir l'inégalité (3.16) car le membre de gauche est non nul et le membre de droite est nul. De plus, si Δ s'annule en un point $z \in \mathbb{S}$, le déterminant intrinsèque de Kreiss–Lopatinskii ne pourra pas être minoré uniformément au voisinage de z, donc la constante K de la condition de Kreiss–Lopatinskii uniforme ne pourra pas être bornée uniformément en z. □

3.5 Deuxième version du théorème de Kreiss

3.5.1 Énoncé

On donne une deuxième version du théorème de Kreiss (après celle du Théorème 1.22, page 36) qui présente une deuxième condition nécessaire et suffisante pour la stabilité forte,



grâce à l'utilisation de la Proposition 3.2 (page 72).

Théorème 3.23 (Kreiss 2)**.** *Les assertions suivantes sont équivalentes :*

(i) le schéma (3.2) *est stable au sens de la Définition 3.1.*

(ii) la condition de Kreiss–Lopatinskii uniforme (3.16) *est satisfaite.*

La preuve de ce théorème utilise les symétriseurs de Kreiss, comme cela est présenté par Coulombel [Cou13]. Dans la suite, on va en donner une preuve plus élémentaire dans le cas particulier où les racines $(\kappa_j(z))_{j=1}^r$ de (3.4) localisées dans le disque unité sont simples.

Le Théorème 3.22 nous donne une troisième condition nécessaire et suffisante pour avoir la stabilité forte.

Corollaire 3.24. *Le schéma* (3.1) *est fortement stable si et seulement si le déterminant intrinsèque de Kreiss–Lopatinskii Δ ne s'annule pas sur $\overline{\mathcal{U}}$.*

Ce corollaire est primordial dans l'étude faite dans les Chapitres 5 et 6 où on va étudier les zéros du déterminant intrinsèque de Kreiss–Lopatinskii.

3.5.2 Démonstration du théorème de Kreiss

Démonstration de $\boxed{(i) \implies (ii)}$ *du Théorème 3.23.* On peut remarquer que la condition de Kreiss–Lopatinskii uniforme est incluse dans l'inégalité de stabilité (3.3), donc la condition de Kreiss–Lopatinskii uniforme est immédiatement satisfaite. □

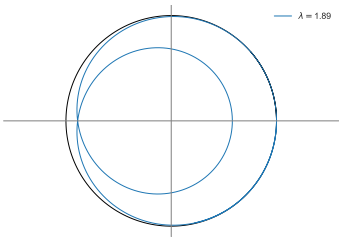
L'autre implication $\boxed{(ii) \implies (i)}$ est plus technique et peut être trouvée dans [Cou13], elle utilise les symétriseurs de Kreiss étudiés (pour le cas continu) dans [Mét00, MZ04]. Dans [Cou13], le cadre est plus général, Coulombel traite les schémas multipas pour des systèmes de taille N. Ici, on traite les schémas à un pas pour une équation scalaire. Dans toute la suite de cette sous-section, on va faire une preuve plus élémentaire de cette implication dans le cas où les racines $(\kappa_j)_{j=1}^r$ de (3.4) localisées dans le disque unité sont simples, *i.e.* les racines $(\kappa_j)_{j=1}^r$ sont distinctes.

Le Théorème 3.25 ne sera utilisé que dans le cas de racines simples mais on a préféré donner une forme générale qui traite aussi le cas de multiplicité dans l'espoir de pouvoir traiter le cas de racines multiples avec une preuve plus élémentaire.

Théorème 3.25. *Soit $R > 1$. Soit $0 < \eta < R$. Soit $z_0 \in \mathbb{C}$ tel que $|z_0| = 1 + \eta$ et $\kappa_0(z_0)$ une racine de* (3.4) *de multiplicité β. On a alors*

$$|\kappa_0(z_0)| \leqslant 1 - C\eta^{\frac{1}{\beta}}$$

avec $C > 0$ ne dépendant que de R, r, p et les coefficients $(a_j)_{j=-r}^p$ du schéma.



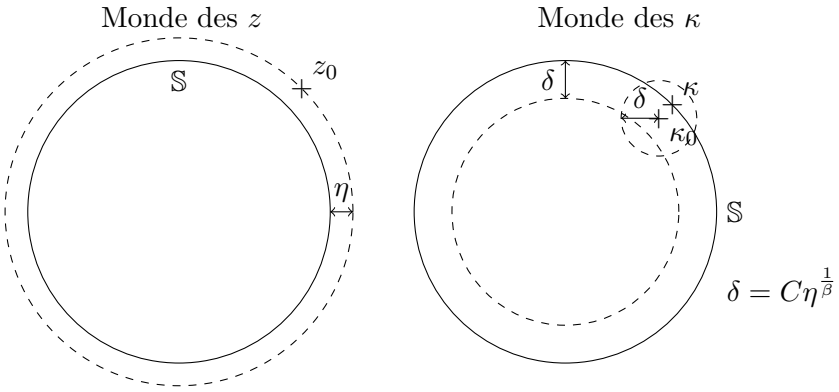


FIGURE 3.2 – Illustration de la preuve du Théorème 3.25.

Démonstration. Préliminaires : on pose, pour tout $\beta \in [\![1 : r]\!]$,

$$C_\beta = \min\left(\left(\frac{(\beta-1)!}{\sum_{j=1}^{r}|a_{-j}|\frac{(j+\beta-1)!}{(j-1)!}2^{j+\beta}+\sum_{j=1}^{p}|a_j|\frac{j!}{(j-\beta-1)!}}\right)^{\frac{1}{\beta}}, \frac{1}{2R^{\frac{1}{\beta}}}\right). \tag{3.22}$$

De plus, on pose la constante suivante :

$$C = \min_{\beta=1}^{r} C_\beta \tag{3.23}$$

qui sera la constante de l'énoncé. On utilise le symbole γ donné par $\gamma : \kappa \mapsto \sum_{j=-r}^{p} a_j\kappa^j$.

On rappelle la formule de Taylor-reste intégral :

$$\gamma(\kappa)-\gamma(\kappa_0) = \sum_{k=1}^{\beta-1}\frac{(\kappa-\kappa_0)^k}{k!}\gamma^{(\beta)}(\kappa_0) + \int_0^1 \frac{(1-t)^{\beta-1}}{(\beta-1)!}(\kappa-\kappa_0)^\beta\gamma^{(\beta)}((1-t)\kappa_0+t\kappa)dt.$$

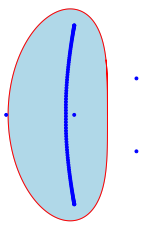
De plus, les dérivées successives de la fonction γ sont, pour tout $\beta \in [\![1 : r]\!]$, pour tout $\nu \in \mathbb{C}$,

$$\gamma^{(\beta)}(\nu) = \sum_{j=1}^{r} a_{-j}(-1)^\beta\frac{(j+\beta-1)!}{(j-1)!}\frac{1}{\nu^{j+\beta}} + \sum_{j=1}^{p} a_j\frac{j!}{(j-\beta+1)!}\nu^{j-\beta}.$$

Commençons la preuve. Raisonnons par l'absurde. Soit une racine $\kappa_0(z_0)$ de (3.4) de multiplicité β telle que $|\kappa_0(z_0)| \in]1 - C\eta^{\frac{1}{\beta}}, 1]$. On note κ_0 pour $\kappa_0(z_0)$ et δ pour $C\eta^{\frac{1}{\beta}}$ comme dans la Figure 3.2.

On choisit un élément $\kappa \in \mathbb{S} \cap B(\kappa_0, \delta)$. On pose $z = \gamma(\kappa)$.

Montrons que $z \in B(z_0, \eta)$. Pour cela, on va utiliser la formule de Taylor-reste intégral en se souvenant que $\gamma^{(k)}(\kappa_0) = 0$ pour tout $k \in [\![1 : \beta - 1]\!]$ par définition de la multiplicité β de κ_0.



De plus, comme $|\kappa_0| \in]1-\delta, 1]$ et $|\kappa| = 1$, on a

$$\forall t \in [0,1], \quad |(1-t)\kappa_0 + t\kappa| \in]1-\delta, 1]. \tag{3.24}$$

D'où

$$\begin{aligned}
|z - z_0| &= |\gamma(\kappa) - \gamma(\kappa_0)| \\
&= \left| \int_0^1 \frac{(1-t)^{\beta-1}}{(\beta-1)!} (\kappa - \kappa_0)^\beta \gamma^{(\beta)}((1-t)\kappa_0 + t\kappa) dt \right| \\
&\underset{(3.24)}{\leqslant} \frac{|\kappa - \kappa_0|^\beta}{(\beta-1)!} \left(\sum_{j=1}^{r} |a_{-j}| \frac{(j+\beta-1)!}{(j-1)!} \frac{1}{(1-\delta)^{j+\beta}} + \sum_{j=1}^{p} |a_j| \frac{j!}{(j-\beta+1)!} \right) \\
&\underset{(3.22)}{<} \frac{(C\eta^{\frac{1}{\beta}})^\beta}{(\beta-1)!} \left(\sum_{j=1}^{r} |a_{-j}| \frac{(j+\beta-1)!}{(j-1)!} 2^{j+\beta} + \sum_{j=1}^{p} |a_j| \frac{j!}{(j-\beta+1)!} \right) \\
&\underset{(3.23)}{\leqslant} \frac{C_\beta^\beta \eta}{(\beta-1)!} \left(\sum_{j=1}^{r} |a_{-j}| \frac{(j+\beta-1)!}{(j-1)!} 2^{j+\beta} + \sum_{j=1}^{p} |a_j| \frac{j!}{(j-\beta+1)!} \right) \underset{(3.22)}{\leqslant} \eta.
\end{aligned}$$

Ainsi, on a

$$|z| \geqslant ||z_0| - |z - z_0|| = 1 + \eta - |z - z_0| > 1.$$

Par le Lemme 3.4 (Hersh), comme $|z| > 1$, aucune racine de (3.4) n'est sur le cercle unité. Cela contredit le fait que $\kappa \in \mathbb{S}$ et $\gamma(\kappa) = z$. On a donc le résultat voulu. □

Théorème 3.26. *Soit $R > 1$. Dans le cas où les racines $(\kappa_j)_{j=1}^r$ sont simples, il existe une constante K telle que pour tout $z \in \mathbb{C}$ vérifiant $1 < |z| \leqslant R$, on a*

$$(|z| - 1) \sum_{j=0}^{+\infty} |\widetilde{U_j}(z)|^2 \leqslant K \sum_{j=-r}^{-1} |\widetilde{g_j}(z)|^2. \tag{3.25}$$

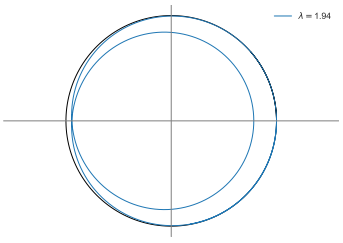
Démonstration. Soit $z \in \mathcal{U}$. Commes les racines (κ_j) de (3.4) localisées dans le disque unité sont supposées simples, les solutions $(\widetilde{U_j}(z))_j$ peuvent s'écrire

$$\forall j \geqslant -r, \quad \widetilde{U_j}(z) = \sum_{k=1}^{r} \alpha_k(z) \kappa_k^j(z).$$

où $\kappa_1(z), \ldots, \kappa_r(z)$ sont les racines de (3.4), toutes de multiplicité 1. Cette formule correspond à la forme (3.7) des solutions $(\widetilde{U_j}(z))_j$.

L'équation (3.19) nous permet de majorer la somme des carrés des $(\alpha_k(z))_k$:

$$\sum_{k=1}^{r} |\alpha_k(z)|^2 \leqslant \|A\| \frac{\|\mathrm{Com}(zK_{0,r-1} - \mathcal{B}K_{0,m-1})\|}{|\Delta_{\mathrm{KL}}(z)|} \sum_{k=-r}^{-1} |\widetilde{g_k}(z)|^2.$$



Comme les racines $\kappa_j(z)$ sont supposées simples, la quantité Δ_{KL} est continue, puisque les racines $(\kappa_j(z))_j$ le sont. La Proposition 3.18 s'applique alors aussi à Δ_{KL}. Ainsi, comme la fonction $z \mapsto zK_{0,r-1} - \mathcal{B}K_{0,m-1}$ est continue sur le compact $\{1 \leqslant |z| \leqslant R\}$, on a une constante C' telle que

$$\sum_{k=1}^{r} |\alpha_k(z)|^2 \leqslant C' \sum_{k=-r}^{-1} |\widetilde{g_k}(z)|^2.$$

On a alors

$$\begin{aligned} |\widetilde{U_j}(z)|^2 = \left|\sum_{k=1}^{r} \alpha_k(z)\kappa_k^j(z)\right|^2 &\leqslant r \sup_{k=1}^{r} |\kappa_k^j(z)|^2 \sum_{k=1}^{r} |\alpha_k(z)|^2 \\ &\leqslant rC' \sum_{k=-r}^{-1} |\widetilde{g_k}(z)|^2 \sup_{k=1}^{r} |\kappa_k^j(z)|^2. \end{aligned}$$

En utilisant le Théorème 3.25, on a pour toute racine $\kappa_k(z)$ l'inégalité suivante :

$$|\kappa_k(z)| \leqslant 1 - C(|z| - 1)$$

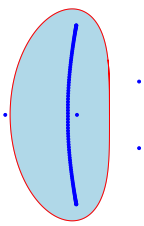
avec C la constante venant du Théorème 3.25. Ainsi, on obtient

$$\begin{aligned} (|z| - 1) \sum_{j=0}^{+\infty} |\widetilde{U_j}(z)|^2 &\leqslant rC' \sum_{k=-r}^{-1} |\widetilde{g_k}(z)|^2 (|z| - 1) \sum_{j=0}^{+\infty} (1 - C(|z| - 1))^{2j} \\ &\leqslant rC' \sum_{k=-r}^{-1} |\widetilde{g_k}(z)|^2 \frac{|z| - 1}{1 - (1 - C(|z| - 1))^2} \\ &\leqslant \frac{rC'}{C} \sum_{k=-r}^{-1} |\widetilde{g_k}(z)|^2 \end{aligned}$$

puisque $\dfrac{|z| - 1}{1 - (1 - C(|z| - 1))^2} = \dfrac{|z| - 1}{C(|z| - 1)[2 - C(|z| - 1)]} = \dfrac{1}{C + C(1 - C(|z| - 1))} \leqslant \dfrac{1}{C}$. □

Remarque 3.27. Malheureusement, la preuve ne s'étend pas facilement au cas de racines multiples. En effet, si on essaye par exemple de faire le même calcul mais avec une racine κ qui serait double. On a alors $|\kappa| \leqslant 1 - C(|z| - 1)^{1/2}$ par le Théorème 3.25 et on veut contrôler la quantité suivante :

$$(|z| - 1) \sum_{j=0}^{+\infty} j^2 |\kappa|^{2j}.$$



En posant $\eta = |z| - 1$, on obtient

$$\begin{aligned}\eta \sum_{j=0}^{+\infty} j^2 |\kappa|^{2j} &= \eta \sum_{j=0}^{+\infty} j(j-1)|\kappa|^{2j} + \eta \sum_{j=0}^{+\infty} j|\kappa|^{2j} \\ &= \frac{2\eta|\kappa|^4}{(1-|\kappa|^2)^3} + \frac{\eta|\kappa|^2}{(1-|\kappa|^2)^2} \\ &\leqslant \frac{2\eta(1-C\eta^{\frac{1}{2}})^4}{\eta^{\frac{3}{2}}(2C - C^2\eta^{\frac{1}{2}})^3} + \frac{\eta(1-C\eta^{\frac{1}{2}})^2}{\eta(2C-C^2\eta^{\frac{1}{2}})^2}\end{aligned}$$

et cette quantité tend vers l'infini quand η tend vers 0. Ce problème s'empire quand la multiplicité augmente.

On utilise maintenant le Théorème 3.26 pour finir de démontrer le Théorème 3.23 (Kreiss).

Démonstration de $\boxed{(ii) \implies (i)}$ *du Théorème 3.23 de Kreiss dans le cas de racines distinctes.* Soit $z \in \mathcal{U}$. Soit R_0 la constante du Théorème 3.20 de la couronne uniforme.

Si $|z| > R_0$, le Théorème 3.20 nous donne que

$$\sum_{j=0}^{+\infty} |\widetilde{U_j}(z)|^2 \leqslant K_1 \sum_{j=-r}^{-1} |\widetilde{g_j}(z)|^2.$$

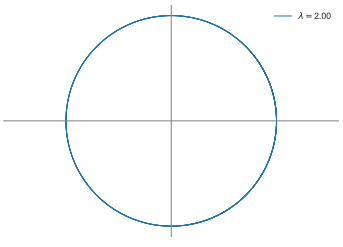
Ainsi, on a

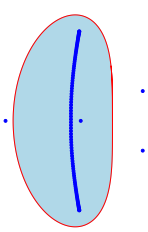
$$\frac{|z|-1}{|z|} \sum_{j=0}^{+\infty} |\widetilde{U_j}(z)|^2 \leqslant K_1 \sum_{j=-r}^{-1} |\widetilde{g_j}(z)|^2.$$

Si $1 < |z| \leqslant R_0$, on utilise le Théorème 3.26, afin d'obtenir,

$$\frac{|z|-1}{|z|} \sum_{j=0}^{+\infty} |\widetilde{U_j}(z)|^2 \leqslant (|z|-1) \sum_{j=0}^{+\infty} |\widetilde{U_j}(z)|^2 \leqslant K_2 \sum_{j=-r}^{-1} |\widetilde{g_j}(z)|^2$$

En prenant $K = K_0 + \max(K_1, K_2)$ où K_0 est la constante de la condition de Kreiss–Lopatinskii uniforme (ii), on a alors l'inégalité de stabilité (3.3) souhaitée. □





Chapitre 4

Revue de la stabilité et présentation des contributions

Le but de la thèse est d'étudier la stabilité des schémas d'ordres élévés, théoriquement et numériquement.

4.1 Bilan de la discussion sur la stabilité

Dans le premier chapitre, on a présenté l'inégalité de la stabilité forte (Définition 1.16 à la page 34) ainsi qu'une première condition nécessaire et suffisante pour obtenir cette stabilité utilisant les notions de valeurs propres et valeurs propres généralisées qui ont été utiles dans le Chapitre 2 et dans le Chapitre 3. Dans le Chapitre 2, on a voulu faire le lien entre ces valeurs propres et valeurs propres généralisées avec les matrices Toeplitz et quasi-Toeplitz associées aux schémas. Cela a permis d'identifier plusieurs questions ouvertes de la littérature. Les difficultés à y répondre que l'on présente dans le Chapitre 2 nous ont menés vers l'étude de la théorie GKS présentée dans le troisième chapitre. On y introduit notamment une nouvelle notion : le déterminant *intrinsèque* de Kreiss–Lopatinskii qui s'appuie sur le déterminant de Kreiss–Lopatinskii déjà existant dans la littérature. On a donné une deuxième condition nécessaire et suffisante pour obtenir la stabilité forte utilisant la condition de Kreiss–Lopatinskii uniforme et, grâce au Corollaire 3.24, on sait que le schéma est fortement stable si et seulement si le déterminant intrinsèque de Kreiss–Lopatinskii ne s'annule pas sur $\overline{\mathcal{U}}$. On veut donc étudier les potentiels zéros de ce déterminant.

Une première stratégie numérique serait de tracer le module du déterminant et d'essayer de voir où celui-ci s'annule. On obtient alors la Figure 4.1.

Il y a plusieurs problèmes avec la représentation de la Figure 4.1 :

- on ne peut pas représenter tout $\overline{\mathcal{U}}$, cependant en utilisant le Théorème 3.20 (Couronne uniforme), il suffit d'étudier les zéros du déterminant dans une couronne autour du cercle unité.
- on ne peut pas dire avec précision si le déterminant s'annule ou non.

De plus, on va comprendre que le schéma de Beam-Warming avec S2ILW3 est instable pour $\lambda = 1.4$ et stable pour $\lambda = 1.6$ (voir Figure 4.3), ce qui n'est pas flagrant sur la Figure 4.1.

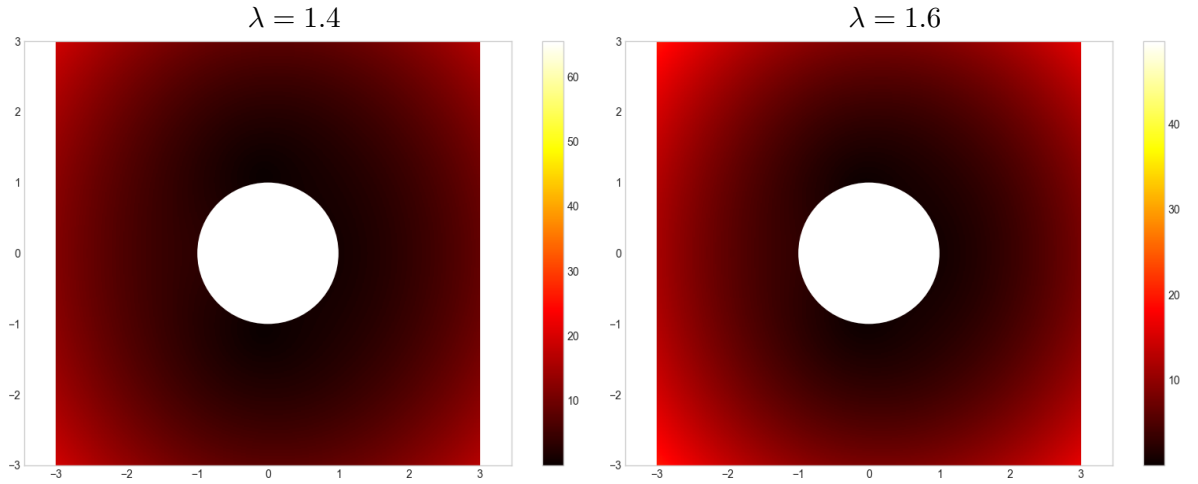


FIGURE 4.1 – Module du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma Beam-Warming (1.22) muni de S2ILW3.

Pour étudier les zéros du déterminant intrinsèque de Kreiss–Lopatinskii, on va d'abord en étudier la régularité.

4.2 Holomorphie et continuité du déterminant intrinsèque de Kreiss–Lopatinskii

Dans le cas particulier des schémas totalement décentrés (*i.e.* lorsque $p = 0$), on a une formule explicite du déterminant intrinsèque de Kreiss–Lopatinskii.

Théorème 4.1 ([BLBS23a])**.** *Sous de bonnes hypothèses, le déterminant intrinsèque de Kreiss–Lopatinskii Δ a la forme suivante :*

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

où $\det C(z)$ est un polynôme en z constructible et dépendant uniquement des coefficients $(a_j)_{j=-r}^{0}$ et de la matrice de bord B.

Ce résultat est démontré en Section 5.3 (page 116). Tout le Chapitre 5, dédié à l'article [BLBS23a], se place dans le cadre des schémas totalement décentrés. Le Théorème 4.1 permet de comprendre que Δ est holomorphe sur $\mathcal{U}$ et continu sur $\overline{\mathcal{U}}$, ce qui correspond aux propriétés de $\mathcal{E}^s(z)$ (voir Théorème 3.8, page 79).

Dans le cas général, on a aussi un théorème de régularité du déterminant intrinsèque de Kreiss–Lopatinskii.

Théorème 4.2 ([BLBS23b])**.** *Sous hypothèse de Cauchy-stabilité, le déterminant intrinsèque de Kreiss–Lopatinskii Δ est holomorphe sur $\mathcal{U}$ et continu sur $\overline{\mathcal{U}}$.*

Ce résultat est démontré en Section 6.3 (page 153). Tout le Chapitre 6, dédié à l'article [BLBS23b], traite ce cas. Cette fois-ci, on n'a pas de formule explicite du déterminant intrinsèque de Kreiss–Lopatinskii mais on a une reformulation qui permet de pouvoir tracer numériquement le déterminant (voir (6.27), page 157). Cela est utile d'un point de vue numérique pour pouvoir conclure sur la stabilité du schéma, comme on le voit dans la section suivante.

4.3 Stratégie numérique pour conclure sur la stabilité

Grâce à la régularité du déterminant de Kreiss–Lopatinskii, on peut utiliser le théorème des résidus afin de compter le nombre de zéros d'une fonction (voir principe de l'argument en Théorème A.3, page 200).

On utilise la courbe $\Delta(\mathbb{S})$ définie par :

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

Théorème 4.3 ([BLBS23a, BLBS23b]). *Sous hypothèse de Cauchy-stabilité, si* $0 \notin \Delta(\mathbb{S})$*, alors l'équation* $\Delta(z) = 0$ *possède* $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ *zéros dans* $\mathcal{U}$*.*

Il faut donc tracer la courbe $\Delta(\mathbb{S})$ du déterminant intrinsèque de Kreiss–Lopatinskii pour pouvoir vérifier que Δ ne s'annule pas sur $\mathbb{S}$, puis compter le nombre de zéros de Δ dans $\mathcal{U}$. C'est l'objet des Méthode 5.19 et Méthode 6.15 (pages 116 et 153).

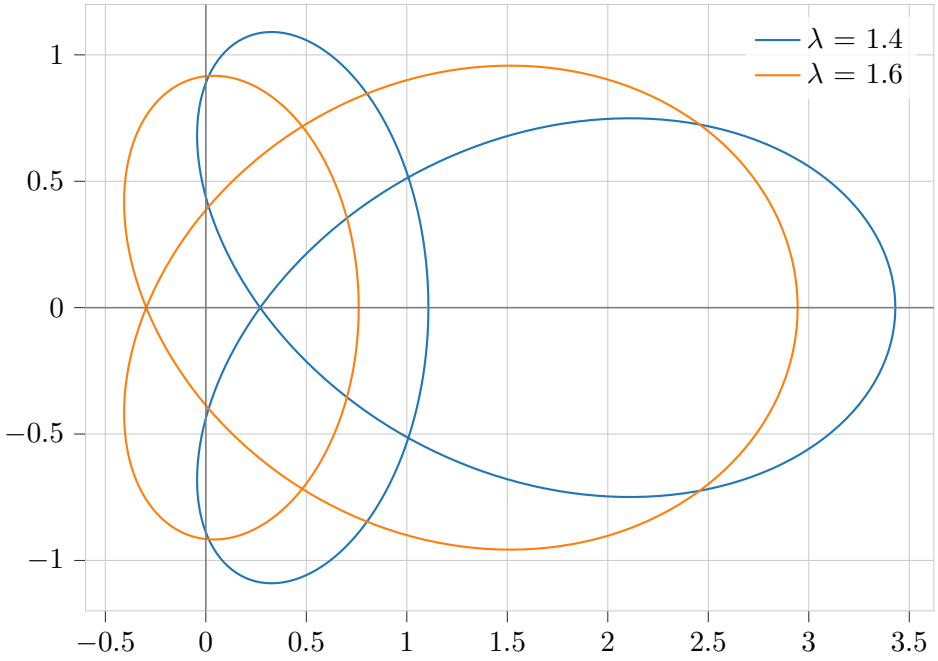


FIGURE 4.2 – Courbe $\Delta(\mathbb{S})$ du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma Beam-Warming (1.22) muni de S2ILW3.

La Figure 4.2 permet de voir que le schéma de Beam-Warming muni de S2ILW3 n'est pas fortement stable pour $\lambda = 1.4$. En effet, la courbe $\Delta(\mathbb{S})$ ne fait aucun tour autour de l'origine, donc $\mathrm{Ind}_{\Delta(\mathbb{S})}(0) = 0$, ainsi par le Théorème 4.3, le déterminant intrinsèque de Kreiss–Lopatinskii s'annule deux fois (puisque $r = 2$ pour le schéma de Beam-Warming). A contrario pour $\lambda = 1.6$, la courbe $\Delta(\mathbb{S})$ tourne deux fois autour de l'origine, donc le déterminant intrinsèque de Kreiss–Lopatinskii ne s'annule ni sur $\mathbb{S}$ ni dans $\mathcal{U}$. Ainsi, le schéma est fortement stable pour $\lambda = 1.6$.

Le Chapitre 7 donne des outils numériques pour automatiser l'étude de la courbe $\Delta(\mathbb{S})$, notamment en utilisant un algorithme pour calculer l'indice complexe d'une courbe. Ainsi, on peut calculer la quantité $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ en fonction de λ, ce qui est représenté en Figure 4.3 pour le schéma Beam-Warming muni de S2ILW3.

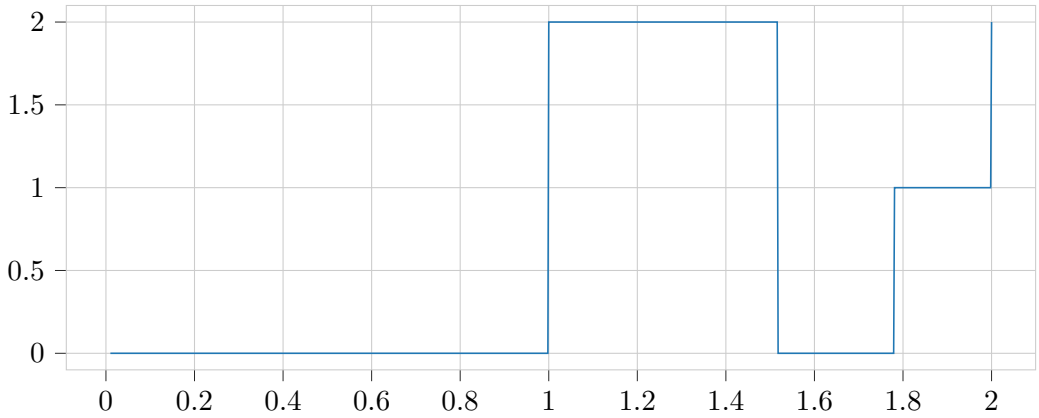


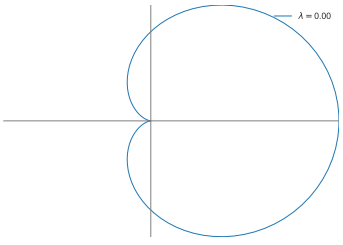
FIGURE 4.3 – Nombre de zéros de Δ dans $\mathcal{U}$ pour le schéma Beam-Warming (1.22) muni de S2ILW3.

Deuxième partie

Etude approfondie du déterminant intrinsèque de Kreiss-Lopatinskii

Chapitre 5

Stability of one-step explicit totally upwind schemes

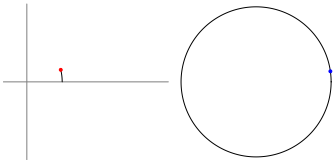


Ce chapitre est dédié à l'article [BLBS23a] : *On the stability of totally upwind schemes for the hyperbolic initial boundary value problem* écrit avec Benjamin Boutin et Nicolas Seguin et accepté dans le journal *IMA Journal of Numerical Analysis*. On a mentionné dans la partie précédente comment le contenu de ce chapitre s'insère dans la littérature.

Résumé.

Dans cet article, on présente une stratégie numérique permettant de conclure sur la stabilité forte d'un schéma explicite à un pas totalement décentré avec conditions de bord numérique. Toute l'étude est faite sur le modèle jouet de l'équation de transport en dimension 1. La stabilité forte est étudiée sous le prisme de la théorie GKS et du théorème de Kreiss. On introduit et on donne une formule explicite d'un nouvel outil : le déterminant intrinsèque de Kreiss–Lopatinskii qui possède de meilleures propriétés de régularité que le déterminant de Kreiss–Lopatinskii classique. En utilisant des résultats standards d'analyse complexe et d'algèbre linéaire, on peut relier le caractère stable du schéma au calcul de l'indice complexe d'une courbe, ce qui est numériquement robuste et peu coûteux. Cette étude est illustrée par le schéma de Beam-Warming muni de bords définis par une procédure Lax-Wendroff inverse simplifiée. Enfin, on traite le cas où le maillage ne coïncide pas avec la condition de bord physique.

La Section 5.6 (page 135) fournit des preuves plus complètes des résultats de l'article. Notamment, on mentionne les preuves du principe de l'argument utilisé dans la démonstration du Corollaire 5.15 et du théorème de Rouché utile pour la démonstration du Lemme 5.4 (Hersh). On fait explicitement les démonstrations des deux Lemmes 5.25 et 5.26 et on donne une preuve plus détaillée du Lemme 5.27. Pour conclure le chapitre, on donne une généralisation du Lemme 5.21 qui a, à la fois, motivé l'étude du cas particulier $p = 0$ (totalement décentré) et, à la fois, donné une piste pour l'étude du cas général dans l'article [BLBS23b] qui fait l'objet du Chapitre 6.



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

Abstract.

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

5.1 Introduction

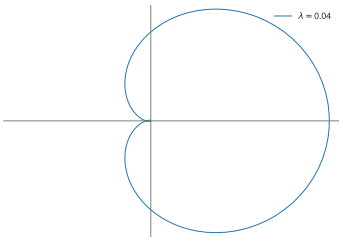
5.1.1 Motivations

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

$$\begin{cases} \partial_t u + a\partial_x u = 0, & t \geqslant 0, x \geqslant 0, \\ u(t,0) = g(t), & t \geqslant 0, \\ u(0,x) = f(x), & x \geqslant 0, \end{cases} \tag{5.1}$$

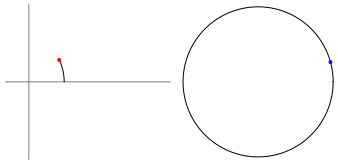
where $u(t,x) \in \mathbb{R}$. The velocity is assumed to be positive $a > 0$ so that at the inflow boundary located at the point $x = 0$, a physical boundary datum g is prescribed.

Let us first recall some general ideas and historical context. As a central idea in numerical analysis, the Lax equivalence theorem [LR56] asserts that a consistent scheme is convergent if and only if it is stable. Therefore, all along the paper only consistent numerical schemes are considered, and the discussion concentrates only on their stability issues. While the Cauchy-stability for the space-periodic problem is easily handled with the Fourier symbolic analysis and the so-called Von-Neumann stability analysis, the case with boundaries is significantly trickier. Indeed, the presence of (unphysical) numerical boundary conditions forms another kind of instabilities. The normal mode analysis, directly related to the work by Godunov and Ryabenkii [GR63], is the classic way to comprehend those kinds of instabilities. Deepening this analysis with resolvent estimates and Laplace transform leads to the notion of *GKS-stability* [GKS72]



(sometimes called *strong stability*, see Definition 5.2 hereafter). This notion is actually the most robust one concerning the stability of initial boundary value numerical methods, since this stability property is stable by perturbations and makes use of the same norms for the solution and for the data itself. These features make possible further extensions to more general cases (e.g. nonlinearities), as it is done for the initial boundary value problem in the case of partial differential equations [BGS06]. In this setting, the Kreiss theorem (see Theorem 5.3 later) expresses a necessary and sufficient condition for strong stability by the use of the so-called Uniform Kreiss-Lopatinskii Condition. When this condition fails, the corresponding instabilities may be interpreted as numerical wave packets with exponential growth in time and/or bad group velocities (see Trefethen [Tre83, Tre84]). Some sketches of the strong stability theory will be unfolded later on, but we refer the interested reader to the monograph [Gus08] by Gustafsson and [GKO13] by Gustafsson, Kreiss and Oliger for a more complete overview of the GKS-stability theory.

The GKS-stability theory is not used so often in the numerical analysis literature. The reason is that the Uniform Kreiss-Lopatinskii Condition requires the search for the vanishing points of the Kreiss-Lopatinskii determinant, which is a complex-valued function defined on $\{|z| \geqslant 1\}$. Except for some particular numerical schemes and boundary conditions, this determinant is not known explicitly. Indeed, the complexity of the underlying algebra rapidly increases as the size of stencil increase. As an example, Thuné develop in [Thu86] a software system for investigating the GKS-stability. Nevertheless, the method requires the numerical approximate computation of the roots of some parameterized characteristic polynomial equation, and may be expensive in terms of CPU time. In order to tackle the stability properties of the discrete initial boundary value problem, some other strategies are available in the literature. Among them, the most natural approach is based on the spectral properties of the operator corresponding to the time-iteration in the numerical scheme. For a large but finite grid of size J, it is represented by a matrix T_J of size J. It is a banded Toeplitz or a quasi-Toeplitz matrix depending on the boundary conditions under consideration. Beam and Warming [BW93] study the asymptotic spectra of such matrices in the limit of large J. Roughly speaking, the stability properties are then related to the uniform boundedness of the powers of the matrix T_J, known as the Kreiss matrix Theorem [TE05, Chap 18]. Nevertheless, the main difficulty is to also guarantee another uniform boundedness property, with respect to the dimension J. The uniform boundedness is not easy to characterize by spectral properties. Some specialized tools exist to that aim: resolvent estimates and ϵ-pseudospectrum. For a wide overview of the Kreiss matrix Theorem and its relationship with resolvent estimates and with the central notion of ϵ-pseudospectrum [BS00, STW02], we refer the reader to the book by Trefethen and Embree [TE05]. Nonetheless, to our knowledge, the link between GKS-instabilities and the pseudospectrum of the family of quasi-Toeplitz matrices associated to a given scheme is still not completely understood. In the numerical analysis literature, a first attempt thus consists in considering only grids with a large



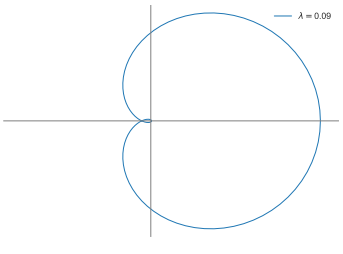
but fixed size J. The postulate is that the asymptotic spectral properties are then already available. This strategy has been used by Dakin, Despres and Jouen [DDJ18] for analyzing some specific boundary conditions that we will again consider with our own method in the present paper.

In the present work, the selected strategy is based on the Uniform Kreiss-Lopatinskii Condition and the search of the vanishing points of the corresponding Kreiss-Lopatinskii determinant, that is a function of the complex parameter z defined for $|z| \geqslant 1$. Instead of using the Kreiss-Lopatinskii determinant, we define the *intrinsic Kreiss-Lopatinskii determinant* that shares the same zeros with the Kreiss-Lopatinskii determinant. The main result of the paper (Theorem 5.13) yields an explicit formula for the intrinsic Kreiss-Lopatinskii determinant, showing that it is holomorphic on $\{|z| > 1\}$. Moreover, the formula does not require the numerical computation of the roots of the associated characteristic equation. Thus, this new theoretical result is particularly useful for numerical applications. Indeed, Corollary 5.15 presents a strategy to find the number of zeros of the intrinsic Kreiss-Lopatinskii determinant on the domain $\{|z| > 1\}$ using a numerical computation of winding numbers. Hence, this corollary enables the Method 5.19 to tackle the stability of the scheme. The whole study in this paper is restricted to totally upwind schemes, so the consistency order is limited to 2 (see Iserles [IS83]). As typical examples, we therefore focus on the classic first-order upwind and Beam-Warming schemes, while the generality of the study comes from the fact that we can take any extrapolation boundary condition using some points of the domain (the precise form of the considered boundary conditions will be set later at equation (5.5)). In the paper, the numerical examples deal with the inverse Lax-Wendroff boundary condition, and the simplified variants of it, as introduced by Tan, Shu and Vilar in [TS10, VS15] and used by Li, Shu and Zhang in [LSZ16, LSZ17, LLS22] to solve advection and diffusion equations. These authors consider a stability analysis based either on the Godunov-Ryabenkii algebraic condition, or by the so-called eigenvalue spectrum visualization method. This last method again requires the use of a finite grid and the computation of the eigenvalues for a large banded matrix.

The outline of the paper is as follows. In the sequel of this introductive section, we describe the main assumptions and the notion of stability into play. In Section 5.2, we set up the main tool for our study that is the Kreiss-Lopatinskii determinant and the intrinsic Kreiss-Lopatinskii determinant, then we state our main results. In Section 5.3, we prove these results relying on linear algebra tools and complex analysis results. Section 5.4 gathers several examples and numerical experiments for illustrating the efficiency of the proposed strategy.

5.1.2 Notations and assumptions

Throughout this paper we denote $\mathbb{S} = \{z \in \mathbb{C}, |z| = 1\}$ the unit circle, $\mathbb{D} = \{z \in \mathbb{C}, |z| < 1\}$ the open unit disk, $\mathcal{U} = \{z \in \mathbb{C}, |z| > 1\}$ the exterior domain and $\overline{\mathcal{U}} = \{z \in \mathbb{C}, |z| \geqslant 1\}$ its



closure. For $n < m$, the notation $[\![n : m]\!]$ is for the set $\{k \in \mathbb{N}, n \leqslant k \leqslant m\}$.

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

$$U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{j+k}^n, \tag{5.2}$$

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

The *symbol* associated to the scheme (5.2) is defined, for $\xi \in \mathbb{R}$, by

$$\gamma(\xi) = \sum_{k=-r}^{p} a_k e^{ik\xi}. \tag{5.3}$$

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

Assumptions. The scheme (5.2) is

(H0) *non-degenerate*, in the sense that $a_{-r} \neq 0$,

(H1) *totally upwind*, in the sense that $p = 0$,

(H2) *Cauchy-stable*, meaning that the symbol γ satisfies $|\gamma(\xi)| \leqslant 1$ for all $\xi \in \mathbb{R}$.

(H3) *consistent* and at least first order, meaning that

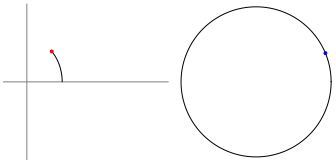
$$\gamma(0) = \sum_{k=-r}^{p} a_k = 1 \text{ and } -i\gamma'(0) = \sum_{k=-r}^{p} k a_k = -\lambda.$$

When dealing with the discrete schemes set over the full line $j \in \mathbb{Z}$, the algebraic characterization of the Cauchy-stability classically follows from the Fourier analysis and makes use of the symbol γ. This method is known as the Von Neumann analysis (see [CFL28] and [CN47]). In the scalar case, it reduces to a geometric property concerning the following closed complex curve.

Definition 5.1. The *symbol curve* Γ is the closed complex parametrized curve

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

This definition enables a geometric interpretation of the Cauchy-stability assumption (H2) reformulated equivalently as the inclusion $\Gamma \subset \overline{\mathbb{D}}$ (see later Figure 5.2 for the Beam-Warming scheme). In the same vein, the consistency assumption (H3) admits a geometric form through a first order tangency property of Γ to the vertical axis at the parameter point $\theta = 0$.



The stability condition (H2) can be easily illustrated graphically in the complex plane. In some sense, our goal is to extend this kind of graphical study when including the numerical boundary conditions.

For solving the Initial Boundary Value Problem (IBVP) (5.1) with the discrete scheme (5.2), r additional ghost points are needed to take into account the left boundary condition and to fully define the discrete approximation. In the theoretical results of the paper, we assume that the values at these ghost points are obtained from a linear combination of the first values of the solution close to the boundary and at the same time step, as follows.

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{0} a_k U_{k+j}^n, & j \in \mathbb{N},\ n \in \mathbb{N}, \quad (5.4) \\ U_j^n = \displaystyle\sum_{k=0}^{m-1} b_{j,k} U_k^n + g_j^n, & j \in [\![-r:-1]\!],\ n \in \mathbb{N}, \quad (5.5) \\ U_j^0 = f_j, & j \in \mathbb{N}, \quad (5.6) \end{cases}$$

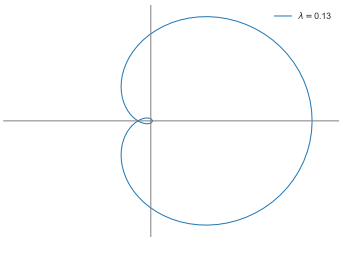
where m, r are integers, f_j are approximations of the initial condition $f(x_j)$ and g_j^n are numerical data related to the boundary datum g. With the vector notation $U = (U_{-r}^n \cdots U_{m-1}^n)^\top$ and $G = (g_{-r}^n \cdots g_{-1}^n)^\top$, the boundary equation (5.5) reads also equivalently as $BU = G$ with the following matrix

$$B \overset{\text{def}}{=} \begin{pmatrix} 1 & & 0 & -b_{-r,0} & \cdots & -b_{-r,m-1} \\ & \ddots & & \vdots & & \vdots \\ 0 & & 1 & -b_{-1,0} & \cdots & -b_{-1,m-1} \end{pmatrix} \in \mathcal{M}_{r,r+m}(\mathbb{C}). \quad (5.7)$$

This class of boundary conditions encompasses the Dirichlet and Neumann extrapoliation procedures [Gol77], but also the more general simplified inverse Lax-Wendroff procedure (see [VS15], [LSZ17], [DDJ18] and Section 5.4.3). We will focus on these boundary conditions in our numerical examples. More specific treatments at the boundary exist, as for example absorbing boundary conditions [EM77] and [Ehr10], or transparent boundary conditions [AES03] and [Cou19], however, in general, they do not enter the present framework.

5.1.3 Classic results about strong stability

The GKS-stability theory (see the seminal paper by Gustafsson, Kreiss and Sundström [GKS72]) handles the discrete IBVP (5.4)-(5.5)-(5.6) with a zero initial data. We refer the reader to the work by Wu [Wu95] and Coulombel [Cou13] for more recent development on semigroup estimates. They extend a stability result for the discrete IBVP (5.4)-(5.5)-(5.6), available for zero initial data, to the case of non-zero initial data. The corresponding notions of stability for



the boundary problem makes use of the following discrete norms:

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

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

Definition 5.2 (Strong stability)**.** The scheme (5.4)-(5.5)-(5.6) is strongly stable if, taking $(f_j) = 0$, there exist $C > 0$ and α_0, such that for all $\alpha > \alpha_0$, for all boundary data (g_j^n), for all $\Delta x > 0$, for all $n \in \mathbb{N}$, the approximate solution (U_j^n) satisfies

$$\sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2 + \left(\frac{\alpha - \alpha_0}{\alpha \Delta t + 1}\right) \|e^{-\alpha n \Delta t} U\|_{\Delta x,\Delta t}^2 \leqslant C \sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} g_j\|_{\Delta t}^2. \tag{5.8}$$

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

The following Kreiss theorem provides two necessary and sufficient conditions for the strong stability. We provide hereafter a condensed formulation of this theorem, obtained from [GKS72, Thm 5.1] combined with [GKO13, Lem 13.1.4] or with [Gus08, Def 2.23]. It makes use of the notions of *eigenvalue* and *generalized eigenvalue* that will be defined later in Definition 5.16 and Definition 5.17.

Theorem 5.3 (Kreiss)**.** *The following statements are equivalent:*

(i) The scheme (5.4)-(5.5)-(5.6) is strongly stable in the sense of Definition 5.2.

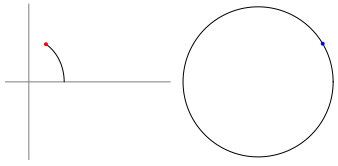
(ii) The scheme (5.4)-(5.5)-(5.6) has neither eigenvalue nor generalized eigenvalue.

(iii) The Uniform Kreiss-Lopatinskii Condition is satisfied.

The Uniform Kreiss-Lopatinskii Condition corresponds to the absence of zeros for the so-called Kreiss-Lopatinskii determinant (see later Definition 5.11 and [GKO13]). These zeros are identified to eigenvalues or to generalized eigenvalues in the sense of Definitions 5.16 and 5.17 and correspond to modal instabilities. Our numerical analysis of the strong stability of the discrete IBVP will be based on a geometrical study of the Kreiss-Lopatinskii determinant.

5.2 Kreiss-Lopatinskii determinants

In this section, we introduce the Kreiss-Lopatinskii determinant, define the intrinsic Kreiss-Lopatinskii determinant and construct an algebraic reformulation of it (see Theorem 5.13 later). This explicit formula shows that it is holomorphic on $\{|z| > 1\}$ and is independent of the roots of the associated characteristic equation. At last, by Corollary 5.15, a numerical procedure based on the Theorem 5.3 (Kreiss) gives a strategy to tackle the stability of the scheme.



5.2.1 Stable subspace $\mathcal{E}^s(z)$ and matrix representation

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

$$U_j^{n+1} = \sum_{k=-r}^{0} a_k U_{k+j}^n, \ j \in \mathbb{N}, \ n \in \mathbb{N}. \tag{5.9}$$

To study this equation, the $\mathcal{Z}$-transform (see [GW99, Lesson 40]) is applied. This transformation is defined for $(x_n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$ such that $x_0 = 0$ and $z \in \mathcal{U}$ by $\widetilde{x}(z) = \sum_{n\geqslant 0} z^{-n} x_n$. The previous equation then becomes

$$z\widetilde{U}_j(z) = \sum_{k=-r}^{0} a_k \widetilde{U}_{j+k}(z), \ j \in \mathbb{N}, \ z \in \mathcal{U}. \tag{5.10}$$

To solve the linear recurrence equation (5.10), let us introduce the following characteristic equation where z plays the role of a parameter and κ is the indeterminate:

$$z\kappa^r = \sum_{k=-r}^{0} a_k \kappa^{r+k}. \tag{5.11}$$

This equation is nothing but the discrete dispersion relation of the finite difference scheme (5.9), with frequency parameter κ in space and z in time. It is formally obtained by looking for solutions to the interior equation (5.9) having the form $U_j^n = z^n \kappa^j$.

In the spirit of a classic result by Hersh [Her63], the following lemma indicates a property of separation for the roots with respect to the unit circle.

Lemma 5.4 (Hersh)**.** *Assume (H0) and (H1). For z in the unbounded connected component of $\mathbb{C} \setminus \Gamma$, all the roots of the characteristic equation (5.11) are in $\mathbb{D}$.*

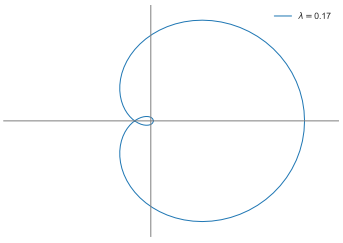
The proof of this result is omitted but may be found in [Her63].

Remark 5.5. Under the Cauchy-stability assumption (H2), the inclusion $\Gamma \subset \overline{\mathbb{D}}$ is known. From there, it follows that the unbounded connected component of $\mathbb{C} \setminus \Gamma$ contains the whole set $\mathcal{U}$ so that a weaker form of the lemma is available for considering $z \in \mathcal{U}$ only. If in addition, the considered scheme is also *dissipative*, that is if its symbol γ satisfies

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

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

Lemma 5.4 (Hersh) is illustrated in Figure 5.1. The first two columns correspond to the Hersh lemma and the third one describes the possible configuration for $z \in \Gamma \cap \mathbb{S}$, typically not meeting the assumptions. This case will be the object of a subsequent discussion.



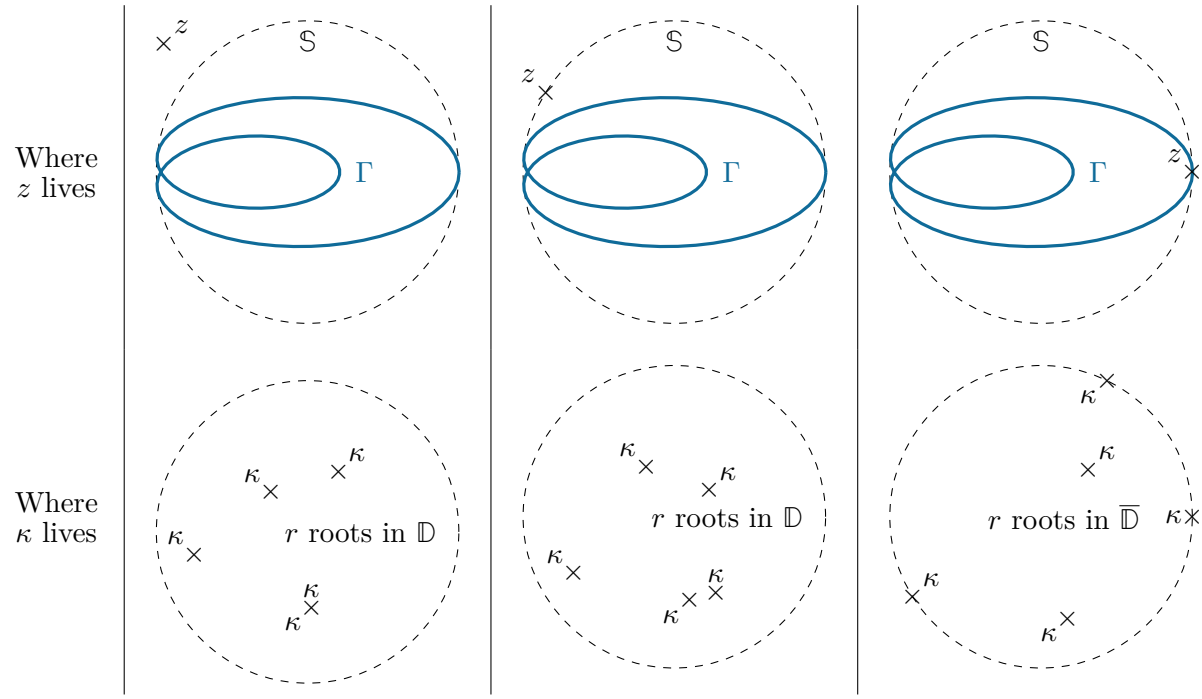


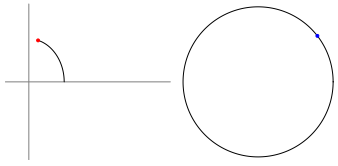
Figure 5.1 – Illustration of Lemma 5.4: case $|z| > 1$ (first column), case $|z| = 1$ and $z \notin \Gamma$ (second column) and case $z \in \Gamma$ where Lemma 5.4 does not hold (third column).

Remark 5.6. Setting the assumption (H1) aside, meaning with a nonzero number p of right points, the more general form of the Hersh lemma states that for any convenient value of z, there are exactly r roots (with multiplicity) inside the open unit disk, exactly p roots (with multiplicity) outside the unit disk and no root on the unit circle. The result can be proved by using Rouché's theorem.

For $|z| > 1$, we denote $\mathcal{E}^s(z)$ the linear subspace of solutions to (5.10) living in ℓ^2 (the ℓ^2 space with indices between $-r$ and $+\infty$). By Lemma 5.4 (Hersh), the space $\mathcal{E}^s(z)$ is generated by the following r vectors:

$$\begin{pmatrix} \kappa_i^{-r} \\ \vdots \\ \kappa_i^{-1} \\ 1 \\ \kappa_i \\ \kappa_i^2 \\ \kappa_i^3 \\ \vdots \end{pmatrix}, \begin{pmatrix} -r\kappa_i^{-r} \\ \vdots \\ -\kappa_i^{-1} \\ 0 \\ \kappa_i \\ 2\kappa_i^2 \\ 3\kappa_i^3 \\ \vdots \end{pmatrix}, \ldots, \begin{pmatrix} (-r)^{\beta_i-1}\kappa_i^{-r} \\ \vdots \\ (-1)^{\beta_i-1}\kappa_i^{-1} \\ 0 \\ \kappa_i \\ 2^{\beta_i-1}\kappa_i^2 \\ 3^{\beta_i-1}\kappa_i^3 \\ \vdots \end{pmatrix}, \quad i = 1, \ldots, M \tag{5.12}$$

where $\kappa_1, \ldots, \kappa_M$ of multiplicity $\beta_1, \ldots, \beta_M$ are the solutions to (5.11), with $\beta_1 + \cdots + \beta_M = r$. (We omit the z-dependence of $\kappa(z)$ for the sake of readability.)



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

Remark 5.7. For $r = 2$, if the solutions to (5.11) are $\kappa_1(z) \neq \kappa_2(z)$, then there are exactly two roots with multiplicity 1. The solutions to (5.10) can be written $\widetilde{U}_j(z) = \alpha_1\kappa_1(z)^j + \alpha_2\kappa_2(z)^j$, and we have

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

Remark 5.8. Still for $r = 2$, if the solution to (5.11) now is $\kappa(z)$ with multiplicity 2, then the solutions to (5.10) can be written $\widetilde{U}_j(z) = (\alpha_1 + \alpha_2 j)\kappa(z)^j$, and we have

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

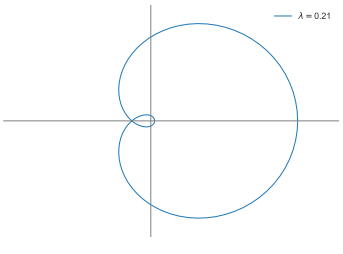
We raise awareness of the dependence on z and of the continuity issues because the map $z \mapsto K_{i,j}(z)$ is not continuous whereas the set of roots of (5.11) is a continuous mapping with respect to z. Indeed, the root curves $(\kappa_j(z))_j$ can intersect, when a multiple root occurs. For example, for $r = 2$, if there is $(z_n)_{n\in\mathbb{N}} \subset \mathcal{U}$ with $\kappa_1(z_n) \neq \kappa_2(z_n)$ which converge to $z_\infty \in \mathcal{U}$ such that $\kappa_1(z_\infty) = \kappa_2(z_\infty)$ a double root, then we have, for $j = 1$ and $j = 2$,

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

but

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

Consequently, the considered basis (5.12) of $\mathcal{E}^s(z)$ does not generally define a continuous mapping with respect to z. Nevertheless, $\mathcal{E}^s(z)$ is a continuous and even holomorphic vector bundle over $\mathcal{U}$ as it is discussed in [Cou13, Thm 4.3]. This author proves in addition that this vector bundle $\mathcal{E}^s(z)$ can even be continuously extended over $\overline{\mathcal{U}}$, thus considering $z \in \mathbb{S}$ as well (see also [MZ04] for a similar property for the hyperbolic-parabolic PDE case). The main point therein is that for some $z_0 \in \mathbb{S}$, there may exist one (or several) root $\kappa_0(z_0)$ of (5.11) on $\mathbb{S}$, because Hersh lemma does not hold anymore. This situation is depicted on the third column of Figure 5.1 and the different cases that may occur will be explained in Section 5.2.4.



In the case of a totally upwind scheme, it is easy to extend the space $\mathcal{E}^s(z)$ because it is the linear space generated by the r roots of (5.11) with polynomial terms for multiplicity. Indeed, $\kappa(z)$ can be defined for all $z \in \overline{\mathcal{U}}$ by continuity of $\kappa(z)$ for $z \in \mathcal{U}$. The space $\mathcal{E}^s(z)$ still is of dimension r and we extend the notation $K_{i,j}(z)$ for z on $\mathbb{S}$. But the difficulty is to prove the continuity of $\mathcal{E}^s(z)$ after the extension, it follows from the existence of a K-symmetrizer and is obtained e.g. in [Cou13, Thm 4.3]. As previously observed, $K_{i,j}(z)$ is generally not continuous with respect to z.

We can summarize the discussion in the following theorem.

Theorem 5.9 ([Cou13])**.** *Under assumptions (H0), (H1) and (H2), the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over $\mathcal{U}$ and can be extended in a unique way as a continuous vector bundle over $\overline{\mathcal{U}}$.*

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

5.2.2 Intrinsic Kreiss-Lopatinskii determinant

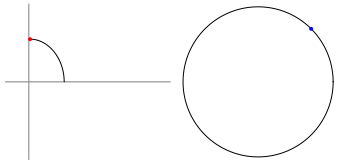
Now, let us consider the $\mathcal{Z}$-transformed version of the boundary condition (5.5), that is, for j between $-r$ and -1,

$$\widetilde{U}_j(z) = \sum_{k=0}^{m-1} b_{j,k}\widetilde{U}_k(z) + \widetilde{g}_j(z). \tag{5.13}$$

Injecting the fundamental solutions to $\mathcal{E}^s(z)$ into (5.13), we obtain a system of r equations with r scalar unknowns. They are the coefficients of the solution to (5.13) written in the basis (5.12) of $\mathcal{E}^s(z)$.

Remark 5.10. For $r = 2$ and a given value of z (we skip for convenience the dependence in z hereafter), if $\kappa_1 \neq \kappa_2$ so that the solution to (5.13) has the form $\alpha_1\kappa_1^j + \alpha_2\kappa_2^j$, then that solution is constrained by the following two scalar equations:

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



The matricial form of that system reads

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

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

Definition 5.11 (Kreiss-Lopatinskii determinant)**.** The *Kreiss-Lopatinskii determinant* is the complex-valued function defined for $|z| \geqslant 1$ by

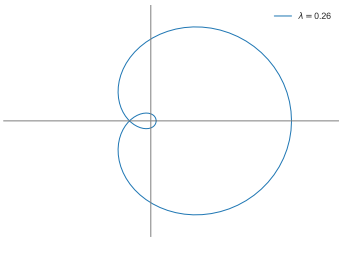
$$\Delta_{\mathrm{KL}}(z) \stackrel{\mathrm{def}}{=} \det BK_{-r,m-1}(z).$$

Before giving the definition let us motivate the *intrinsic Kreiss-Lopatinskii determinant* Δ by the following informal discussion. The above Kreiss-Lopatinskii determinant is actually not well defined until we order in some way the roots $(\kappa_j(z))_{j=1,\dots,r}$ of (5.11). There are two points to notice. The first one is related to crossing roots, already discussed after Remark 5.8. The second one is that, outside crossing cases, being given any choice for the ordering of the roots (and thus of the vectors of the basis (5.12) for the vector bundle), there is in general no chance to obtain a holomorphicity property for the components of the matrix $K_{-r,m-1}(z)$ over $\mathcal{U}$. For example, even the roots of $X^2 - z$ are not holomorphic w.r.t z because of the logarithm determination. On the other side, any symmetric functions of the roots $(\kappa_j(z))_{j=1,\dots,r}$ however are holomorphic because they can be obtained directly in terms of the coefficients of the polynomial (5.11). Therefore, except for crossing roots, the same holds for the quantity $\Delta_{\mathrm{KL}}(z)$ since the matrix B is constant and the determinant itself is a symmetric function.

It is now known that the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over $\mathcal{U}$, continuous over $\overline{\mathcal{U}}$, and thus we should expect the same for Δ_{KL}. A very natural way to reach that property and go beyond the last difficulties consists in dividing Δ_{KL} by the quantity $\det K_{0,r-1}(z)$. In this manner, the same permutation or combination of the vectors of the basis (5.12) is involved for both computations.

Definition 5.12 (Intrinsic Kreiss-Lopatinskii determinant)**.** The *intrinsic Kreiss-Lopatinskii determinant* is the complex-valued function defined for $|z| \geqslant 1$ by:

$$\Delta(z) = \frac{\Delta_{\mathrm{KL}}(z)}{\det K_{0,r-1}(z)}. \tag{5.14}$$



To conclude with these definitions, let us state a little more about the *Uniform Kreiss-Lopatinskii Condition*. With the above notations and additionally to the invertibility of $BK_{-r,m-1}(z)$, it corresponds to the existence of a constant $C > 0$ such that for any $z \in \overline{\mathcal{U}}$, any $U \in \mathcal{E}^s(z)$ solution to (5.13) satisfies the uniform estimate

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

From the Parseval identity for the $\mathcal{Z}$-transform, this inequality yields directly the first necessary half-part of the strong stability estimate (5.8). We refer the reader to [GKO13] for a more detailed presentation.

5.2.3 Main results

Theorem 5.13 is our main theoretical result. It yields an explicit formulation of the intrinsic Kreiss-Lopatinskii determinant and therefore describes its properties. Namely, as a function of z, this determinant Δ is holomorphic on $\mathcal{U}$, is continuous on $\overline{\mathcal{U}}$ and depends on $\mathcal{E}^s(z)$ but not on the choice of a basis (what justifies the *intrinsic* denomination of that quantity).

Theorem 5.13 (Explicit formula of the intrinsic Kreiss-Lopatinskii determinant)**.** *Assume (H0), (H1), (H2) and (H3). The intrinsic Kreiss-Lopatinskii determinant is given, for $z \in \overline{\mathcal{U}}$, by*

$$\Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a_{-r}}{a_0 - z}\right)^{m-r} \tag{5.15}$$

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

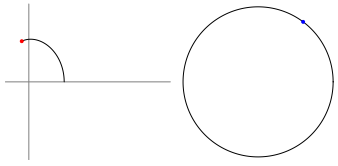
By "constructible polynomial", we mean here that we establish a computable algorithm to get a matrix $C(z)$ and then the polynomial $\det C(z)$. This algorithm, based on a gaussian elimination, is fully described in the proof of Lemma 5.21. In the proof of Theorem 5.13, we will explicitly see the holomorphic property of Δ. Another property, important for the forthcoming applications, lies in the next Corollary 5.15 and involves the following important geometrical object:

Definition 5.14. The *Kreiss-Lopatinskii curve* $\Delta(\mathbb{S})$ is the closed complex parameterized curve

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

Corollary 5.15 (Number of zeros of the intrinsic Kreiss-Lopatinskii determinant)**.** *Assume (H0), (H1), (H2) and (H3). If $0 \notin \Delta(\mathbb{S})$ then the equation $\Delta(z) = 0$ has exactly $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ zeros in $\mathcal{U}$.*

Here above and in all the paper, $\mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ denotes the winding number of the origin with respect to the closed oriented curve $\Delta(\mathbb{S})$ (see [Lan99] for a definition of the winding num-



ber). This previous corollary is the fundamental piece to the following numerical procedure to tackle stability. Indeed, by the definition (5.14), the function Δ shares the same zeros with the Kreiss-Lopatinskii determinant Δ_{KL}, which in turn characterizes the stability with Theorem 5.3 (Kreiss).

5.2.4 Numerical procedure

As already seen in the Theorem 5.3 (Kreiss), the strong stability can be characterized by the notion of eigenvalue and generalized eigenvalue for the boundary problem. The definition of generalized eigenvalue is not universal, the following one comes from [GKO13, Def.12.2.2] but one can also find a slightly different one in [Gus08, Def 2.2]. The difference will be discussed afterwards.

Definition 5.16 (Eigenvalue)**.** Let z be a complex number. If $|z| \geqslant 1$, $\Delta(z) = 0$ and the solution $(\widetilde{U}_j(z))_j$ to (5.10) and (5.13) with $(\widetilde{g}_j(z)) = 0$ is in ℓ^2 then z is called an *eigenvalue*.

Definition 5.17 (Generalized eigenvalue)**.** Let z_0 be a complex number with $|z_0| = 1$. If $\Delta(z_0) = 0$ and the solution $(\widetilde{U}_j(z_0))_j$ to (5.10) and (5.13) with $(\widetilde{g}_j(z_0)) = 0$is not in ℓ^2 then z_0 is called a *generalized eigenvalue*.

If $|z| > 1$ and $\Delta(z) = 0$, it is not possible to have $(\widetilde{U}_j(z)) \notin \ell^2$, because by Lemma 5.4 (Hersh), the r roots of (5.11) that are used to construct $(\widetilde{U}_j(z))$ are in the open unit disk. That's why the definition of generalized eigenvalue concerns only complex values on the unit circle.

Therefore, we can split all cases in four types:

(i) z such that $\Delta(z) = 0$ and $|z| > 1$.

(ii) z such that $\Delta(z) = 0$, $|z| = 1$ and $z \notin \Gamma$.

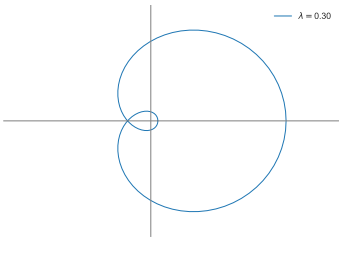
(iii) z such that $\Delta(z) = 0$, $|z| = 1$, $z \in \Gamma$ and $(\widetilde{U}_j(z)) \in \ell^2$.

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

The types (i), (ii) and (iii) describe all the eigenvalues. Indeed, for type (i) and (ii), by Lemma 5.4 (Hersh), we have $(\widetilde{U}_j(z)) \in \ell^2$, because every root κ of (5.11) is in the open unit disk. Type (i) corresponds to the first column of Figure 5.1 and type (ii) corresponds to the second column.

Moreover the non-existence of eigenvalue of type (i) is a necessary condition to have stability. It is called the *Godunov-Ryabenkii* condition, introduced in [GR63] and described in [Tre84].

If z is of type (iii) or (iv), there exists a $\kappa_0(z)$ root of (5.11) on the unit circle because $z \in \Gamma$. This is the situation depicted on the third column of Figure 5.1. The distinction between (iii) and (iv) is more subtle and comes from the expression of $(\widetilde{U}_j(z))$ in the basis of $\mathcal{E}^s(z)$, where the coefficient(s) in front of the vector(s) related to $\kappa_0(z)$ can be zero or not.



By the way, let us mention that our definition of generalized eigenvalue, from [GKO13, Def 12.2.2], corresponds, as we already said, to type (iv) whereas the definition from [Gus08, Def 2.2] combines type (iii) and (iv).

Now, Corollary 5.15 can be reformulated as follows.

Corollary 5.18. *Assume (H0), (H1), (H2) and (H3). If* $0 \notin \Delta(\mathbb{S})$ *then the scheme has* $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ *eigenvalues in* $\mathcal{U}$ *(type (i)).*

This corollary enables us to establish an efficient and practical method to study the stability of a given IBVP through Theorem 5.3 (Kreiss). In particular, the low computational cost of the following procedure is very appealing for the study of parameterised IBVP's, see Section 5.4.

Method 5.19 (Uniform Kreiss-Lopatinskii Condition check)**.** *There are two different cases:*

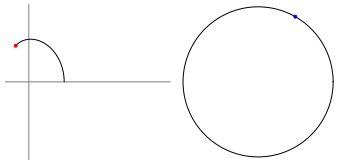
- *if* $0 \notin \Delta(\mathbb{S})$*, there is neither generalized eigenvalue (type (iv)) nor eigenvalue on the unit circle (type (ii) and (iii)) and there are* $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ *zeros of* Δ *in* $\mathcal{U}$ *by Theorem 5.15 (type (i)). It follows that if the scheme has no eigenvalue in* $\mathcal{U}$ *then the scheme is stable. Otherwise there exists an eigenvalue and the scheme is unstable.*
- *if* $0 \in \Delta(\mathbb{S})$*, then there exists* $z_0 \in \mathbb{S}$ *such that* $\Delta(z_0) = 0$*.*
  - → *If* $z_0 \in \Gamma$*, then* z_0 *is a generalized eigenvalue (of type (iv)) or an eigenvalue of type (iii).*
  - → *If* $z_0 \notin \Gamma$*, then there are two possibilities:*
    - ➢ *first,* z_0 *is in the unbounded connected component of* $\mathbb{C}\backslash\Gamma$*. By Lemma 5.4 (Hersh), there is no* κ *on the unit circle, so* z_0 *is an eigenvalue on the unit circle (type (ii)).*
    - ➢ *second,* z_0 *is in a bounded connected component of* $\mathbb{C} \setminus \Gamma$*. Contradiction with the Cauchy-stability because* $\Gamma \subset \overline{\mathbb{D}}$ *and* $z_0 \in \mathbb{S}$*.*

This method does not distinguish between types (iii) and (iv). In fact, we only study the presence or absence of instabilities, we do not attempt to determine which type of instability mode is met (see Trefethen [Tre84]).

In summary, by Theorem 5.3 (Kreiss), if $0 \in \Delta(\mathbb{S})$ then the scheme is not stable, and if $0 \notin \Delta(\mathbb{S})$, Theorem 5.15 can be used to conclude that the scheme is stable or not, depending on the value of $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$. Some illustrations for the Beam-Warming scheme follow in Section 5.4.

5.3 Proof of Theorem 5.13 and Corollary 5.15

In order to use the residue theorem, the holomorphy of the Kreiss-Lopatinskii determinant is needed. To this end, we want a nicer expression of $\det BK_{-r,m-1}(z)$ the Kreiss-Lopatinskii determinant. Clearly the multiplicativity of the determinant does not apply in the expression



$\Delta_{\mathrm{KL}}(z) = \det BK_{-r,m-1}(z)$ since B and $K_{-r,m-1}(z)$ are non-square matrices. A first step consists of reducing the problem to a linear algebra formulation with square matrices to use the multiplicativity of the determinant. All along the current section, the assumptions (H0), (H1) and (H2), required to define the matrices $K_{i,j}$, the vector bundle $\mathcal{E}^s$ as well as its extension over $\overline{\mathcal{U}}$, are supposed to be fulfilled.

5.3.1 Reduction to a square formulation

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

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

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

Definition 5.20. Let E be a linear subspace of $\ell^2(\mathbb{N})$. Two matrices $B, D \in \mathcal{M}_{r,N}(\mathbb{C})$ (with $N \in \mathbb{N} \setminus \{0\}$ be any nonzero integer) are said to be equivalent, which we denote $B \sim_E D$, if and only if for all $U \in E$, one has $B\pi(U) = D\pi(U)$, where π is the canonical projection from ℓ^2 onto $\mathbb{C}^N$, keeping the N first components of U.

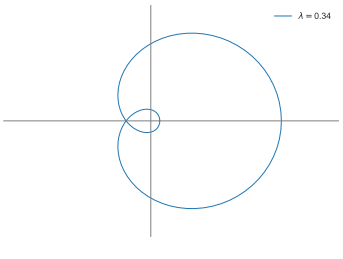
To act conveniently with elementary Gaussian operations, we use some specific notations in the following discussions. We denote $M[i : j, k : \ell]$ the matrix obtained by the extraction of the lines between i and j and the columns between k and ℓ of the matrix M (all indices are included). Similarly,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

Lemma 5.21. *Let $N \geqslant r$ be an integer. Let $B \in \mathcal{M}_{r,N}(\mathbb{C})$ be a constant complex matrix such that $B[1 : r, 1 : r] \in \mathrm{GL}_r(\mathbb{C})$. Assume moreover that $|a_0| < 1$.*
For any $z \in \overline{\mathcal{U}}$, consider the associated linear subspace $\mathcal{E}^s(z)$. There exists a unique square matrix $C(z) \in \mathcal{M}_r(\mathbb{C})$ such that

$$B \sim_{\mathcal{E}^s(z)} \left(\begin{array}{ccc|c} 0 & \cdots & 0 & \\ \vdots & & \vdots & C(z) \\ 0 & \cdots & 0 & \end{array}\right) \begin{matrix} \\ \\ \\ \end{matrix} \!\!\updownarrow r$$
$$\underbrace{\qquad\qquad}_{N-r} \quad \underbrace{\qquad}_{r}$$

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

Remark 5.22. Let us highlight our use of this lemma. Let $\ell \geqslant -r$ and $z \in \overline{\mathcal{U}}$ be fixed. From the basis (5.12), the columns of the matrix $K_{\ell,\ell+N-1}(z)$ take the form $\pi(U)$ for some $U \in \mathcal{E}^s(z)$ and π the canonical projection from $\ell^2(\mathbb{N})$ onto $\mathbb{C}^N$. Therefore, for any convenient matrices B and D with $B \sim_{\mathcal{E}^s(z)} D$, one has then $BK_{\ell,\ell+N-1}(z) = DK_{\ell,\ell+N-1}(z)$.



Now for the boundary matrix B defined in (5.7) and the matrix $D(z) = (0 \mid C(z))$ obtained by the lemma, the following computation by block is possible

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

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

Proof of Lemma 5.21.

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

(a) $B \sim_{\mathcal{E}^s(z)} B^{(j)}$.

(b) the j first columns of $B^{(j)}$ are zero.

(c) every component of $B^{(j)}$ is polynomial of z.

(d) every component of $B^{(j)}[r+1+j : N]$ are independent of z.

(e) the degree of $\det B^{(j)}[j+1 : j+r]$ is j.

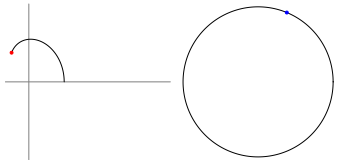
Initialization: we define $B^{(0)} \overset{\text{def}}{=} B$ which satisfies the five induction hypotheses. The induction hypotheses from (a) to (d) are trivially satisfied. The induction hypothesis (e) is satisfied because, $\det B[1 : r, 1 : r] \in \mathbb{C}^*$ which is a non zero constant polynomial.

Induction: we suppose true the induction hypotheses for some $j \in [\![0 : N-r-1]\!]$ and we want to prove it for $j+1$.

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

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

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



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

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

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

Then, by alternating property of the determinant, we have

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

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

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

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

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

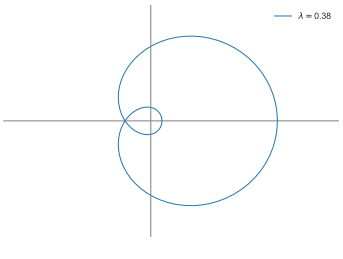
$$B \sim_{\mathcal{E}^s(z)} \underbrace{\left(\begin{array}{ccc|c} 0 & \cdots & 0 & \\ \vdots & & \vdots & C(z) \\ 0 & \cdots & 0 & \end{array}\right)}_{=D} \sim_{\mathcal{E}^s(z)} \underbrace{\left(\begin{array}{ccc|c} 0 & \cdots & 0 & \\ \vdots & & \vdots & C'(z) \\ 0 & \cdots & 0 & \end{array}\right)}_{=D'}.$$

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

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

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

We have $F \cap \mathrm{Vect}(e_1,\ldots,e_{N-r}) = \{0\}$. Indeed if $x \in F \cap \mathrm{Vect}(e_1,\ldots,e_{N-r})$, then $Ax = 0$



and $x = (x_1, \dots, x_{N-r}, 0, \dots, 0)^\top$. By solving the triangular system

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

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

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

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

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

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

Remark 5.23. The uniqueness result is actually not needed for the next results.

In Section 5.4.2 (resp. Section 5.4.5), we perform the explicit algorithmic computations described above for the classic first-order upwind scheme (5.21) (resp. Beam-Warming scheme (5.25)).

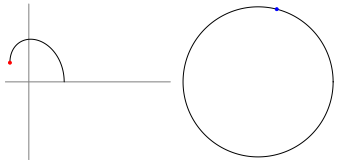
5.3.2 Holomorphy

Lemma 5.24. *Assume $|a_0| < 1$. For all $\ell \in \mathbb{N}$ and for all $z \in \overline{\mathcal{U}}$, we have*

$$\frac{\det K_{\ell,\ell+r-1}(z)}{\det K_{0,r-1}(z)} = (-1)^{\ell r}\left(\frac{a_{-r}}{a_0 - z}\right)^{\ell}. \tag{5.16}$$

Proof. **Case with only one root κ of multiplicity β.** By Lemma 5.4 (Hersh), we know that $\beta = r$, but let keep β because it will be useful for the next step. We recall that

$$\det K_{0,r-1} = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ \kappa & \kappa & \cdots & \kappa \\ \kappa^2 & 2\kappa^2 & \cdots & 2^{\beta-1}\kappa^2 \\ \vdots & & & \vdots \\ \kappa^{r-1} & (r-1)\kappa^{r-1} & \cdots & (r-1)^{\beta-1}\kappa^{r-1} \end{vmatrix}. \tag{5.17}$$



We want to work on

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

We do some operations on columns to recover (5.17). For n from $\beta-1$ to 0, we replace the column C_n by $\sum_{k=0}^{n}(-\ell)^{n-k}\binom{n}{k}C_k$. After the transformation, the component in position (i,n), with $i \in [\![0 : r-1]\!]$ and $n \in [\![0 : \beta-1]\!]$, is

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

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

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

Conclusion.

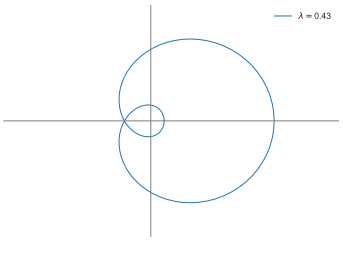
We proved

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

Observe that $a_0 - z \neq 0$ because $z \in \overline{\mathcal{U}}$ and $a_0 \in \mathbb{D}$. Therefore, by Vieta's formulas for the polynomial (5.11), we finally have

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

□



Lemma 5.24 implies the holomorphy on $\mathcal{U}$ and continuity on $\overline{\mathcal{U}}$ of the function in (5.16).

For the sake of completeness in the forthcoming proofs, we state hereafter two elementary lemmas. Both are easily deduced from classic properties of the winding number in complex analysis (see [Lan99]).

Lemma 5.25. *Let P and Q be two polynomials with $\deg P > \deg Q$. If the function $z \mapsto P(z)Q(z)^{-1}$ is holomorphic on $\mathcal{U}$ then $z \mapsto P(1/z)Q(1/z)^{-1}$ is meromorphic on $\mathbb{D}$ with only one pole, at the origin of order $\deg P - \deg Q$.*

Lemma 5.26. *Let f be a holomorphic function on $\mathcal{U}$ and continuous on $\overline{\mathcal{U}}$ and g be the function defined on $\overline{\mathbb{D}}^*$ by $g : z \mapsto f(1/z)$. Then, one has $\mathrm{Ind}_{g(\mathbb{S})}(0) = -\mathrm{Ind}_{f(\mathbb{S})}(0)$.*

5.3.3 Explicit form of the intrinsic Kreiss-Lopatinskii determinant

In the previous Lemmas 5.21 and 5.24, the assumption $|a_0| < 1$ is made. This is not a restriction since this is a consequence of the supplemented consistency assumption.

Lemma 5.27. *Let the scheme (5.4) be Cauchy-stable (H2) and consistent (H3), then $|a_0| < 1$.*

Proof. We have

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

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

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

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

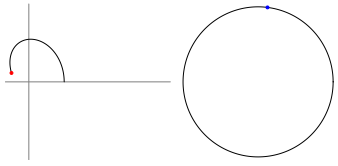
$$0 = \sum_{k=-r}^{p} k a_k = -\lambda \neq 0.$$

By this contradiction, the proof is complete. □

Now every piece can be put together to prove Theorem 5.13.

Proof of Theorem 5.13. Let us recall the function

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



By Lemma 5.27, we will be able to use Lemma 5.21 and Lemma 5.24. With Lemma 5.21 and Remark 5.22, we express Δ as

$$\Delta(z) = \frac{\det C(z) \det K_{m-r,m-1}(z)}{\det K_{0,r-1}(z)} \tag{5.18}$$

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

By Lemma 5.24, we have

$$\frac{\det K_{m-r,m-1}(z)}{\det K_{0,r-1}(z)} = (-1)^{r(m-r)} \left(\frac{a_{-r}}{a_0 - z}\right)^{m-r}. \tag{5.19}$$

By Lemma 5.27, a_0 cannot be a pole of Δ, then the function Δ can be written, for $z \in \overline{\mathcal{U}}$, as

$$\Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a_{-r}}{a_0 - z}\right)^{m-r}, \tag{5.20}$$

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

The proof of Theorem 5.15 relies on the residue theorem to count the zeros of a holomorphic function.

Proof of Theorem 5.15. By Theorem 5.13, the function Δ is holomorphic on $\mathcal{U}$ and continuous on $\overline{\mathcal{U}}$.

Let take the function

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

The function $\widetilde{\Delta}$ is meromorphic on $\mathbb{D}$ with a pole in 0 of order r.

By Lemma 5.21, we have $\deg \det C(z) = m$ because the r first columns of B form the identity matrix of size r which is invertible. By Lemma 5.25 with $P = \det C(z)$ and $Q = (-1)^{r(m-r)} \frac{(a_0 - z)^{m-r}}{a_{-r}^{m-r}}$, the only pole of $\widetilde{\Delta}$ is in 0 and of order

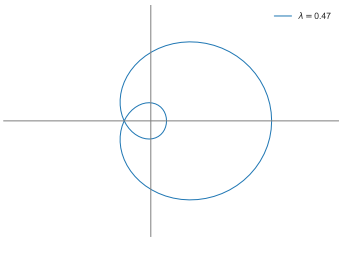
$$\deg \det C(z) - (m - r) = m - (m - r) = r.$$

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

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

Conclusion We have $\#\mathrm{zeros}_{\Delta}(\mathcal{U}) = \#\mathrm{zeros}_{\widetilde{\Delta}}(\mathbb{D})$ and, by Lemma 5.26, we have $\mathrm{Ind}_{\widetilde{\Delta}(\mathbb{S})}(0) = -\mathrm{Ind}_{\Delta(\mathbb{S})}(0)$. It follows that

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



This concludes the proof. □

5.4 Numerical results

In this section, we first explain the numerical computation of the winding number of the origin in order to use Corollary 5.15 and Method 5.19. The simplest first order upwind scheme is then quickly treated, but for a general three-points boundary condition. Next, a main class of high-order boundary conditions, known as the simplified Lax-Wendroff procedure, is presented. They will be used together with the Beam-Warming scheme. After introducing the Beam-Warming scheme, we present computations of Kreiss-Lopatinskii determinant and numerical illustrations. Finally, we study the stability of discretizations where the physical boundaries are not aligned with the mesh.

5.4.1 Computation of the winding number

In the forthcoming numerical illustrations, the interest of Method 5.19 is showcased. Indeed, Theorem 5.15 makes the link between the number of zeros of a holomorphic function and the winding number of a curve which is easy to compute. In fact, as an integer is expected, the approximation of the winding number is generally more reliable, contrary to a real or complex computation because of machine precision.

When the origin is not on the curve, there are different ways to compute the winding number of the origin with respect to the curve. Either we can apply the definition and compute approximately a complex integral, or we can count the number of paths around the origin by using a polygonal approximation of the curve. The second approach is studied by Garcia-Zapata and Martin [GZDM12, GZDM14] with a careful numerical treatment that consists in detecting the possible proximity of the curve to the origin. To that aim, the discretization of the curve is locally refined by an so-called "insertion procedure with control of singularity". Indeed, by the explicit formula (5.15) of the intrinsic Kreiss-Lopatinskii determinant, the curve $\Delta(\mathbb{S})$ is clearly parameterized by the lipschitz function Δ, thus satisfies the required assumptions from the result in [GZDM12] and [GZDM14].

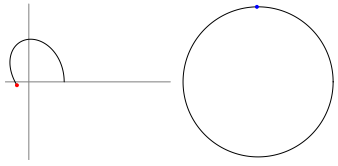
5.4.2 Upwind scheme

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

$$U_j^{n+1} = \lambda U_{j-1}^n + (1-\lambda) U_j^n \tag{5.21}$$

and supplemented at the boundary, for example, by

$$U_{-1}^n = b_0 U_0^n + b_1 U_1^n + b_2 U_2^n \tag{5.22}$$



with arbitrary coefficients b_0, b_1 and b_2. For that scheme, we have $r = 1$, $m = 3$ and the characteristic equation (5.11) reads

$$z\kappa(z) = \lambda + (1-\lambda)\kappa(z). \tag{5.23}$$

The scheme is Cauchy-stable for $\lambda \in]0,1]$ and one can check that for $0 < \lambda \leqslant 1$ and for $|z| > 1$, the root $\kappa(z)$ of (5.23) is in $\mathbb{D}$ by Lemma 5.4 (Hersh).

Let us now execute the computation of the intrinsic Kreiss-Lopatinskii determinant, as presented in Lemma 5.21 (here $N = 4$):

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

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

Hence, the explicit formula (5.15) reads as follows:

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

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

5.4.3 Simplified inverse Lax-Wendroff procedure

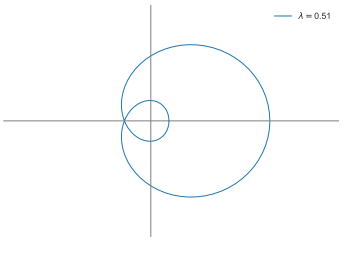
As explained in [TS10] and [VS15], the inverse Lax-Wendroff procedure is used to improve the consistency at the boundary by using the PDE to transform space derivative into time derivative. Namely, for the advection equation (5.1), the following relation holds, for $k \in \mathbb{N}^*$,

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

By a Taylor expansion at order d to approximate $u(n\Delta t, j\Delta x)$ for $n \in \mathbb{N}$ and $j \in [\![-r : -1]\!]$, one can then define the ghost points used in the boundary condition (5.5) by

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

However, many derivatives of the datum g are required to obtain a high order approximation and the complexity then severely increases for multidimensional situations. As explained in [VS15],



the simplified inverse Lax-Wendroff procedure of order d with simplified order k_d that we call "Sk_dILWd" may be used when derivatives of g are not known. Therefore, the first $k_d - 1$ derivatives of g are considered and then for the next terms between order k_d and d, an extrapolation procedure is used. Finally the general formula is, for $j \in [\![-r : -1]\!]$, the following one

$$U_j^n = \sum_{k=0}^{k_d-1} \frac{(-j\Delta x)^k}{k!} \frac{g^{(k)}(n\Delta t)}{a^k} + \sum_{k=k_d}^{d-1} \frac{j^k}{k!} \sum_{s=0}^{d-1} p_{k,s}^{(d)} U_s^n. \tag{5.24}$$

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

5.4.4 Beam-Warming scheme

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

$$\begin{cases} U_j^{n+1} = \dfrac{\lambda(\lambda-1)}{2} U_{j-2}^n + \lambda(2-\lambda) U_{j-1}^n + \dfrac{(\lambda-1)(\lambda-2)}{2} U_j^n, \\ U_{-1}^n = g(t^n) + \dfrac{\Delta x g'(t^n)}{a} + \dfrac{1}{2}(U_2^n - 2U_1^n + U_0^n), \\ U_{-2}^n = g(t^n) + \dfrac{2\Delta x g'(t^n)}{a} + 2(U_2^n - 2U_1^n + U_0^n), \\ U_j^0 = 0. \end{cases} \tag{5.25}$$

This scheme satisfies Assumptions (H1) and (H3). To have the Cauchy-stability assumption (H2), we study the symbol with respect to the CFL condition λ. From (5.25), the symbol is

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

In the Figure 5.2, this symbol is represented for $\lambda = 1.8$.

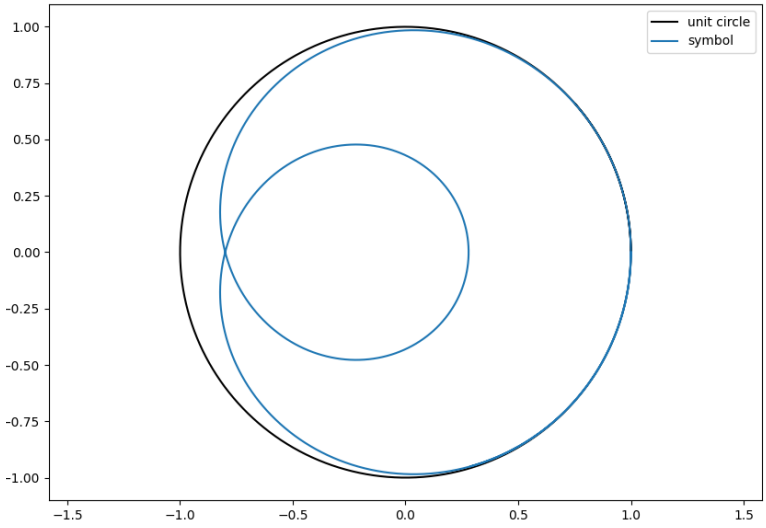
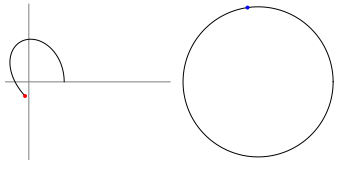


Figure 5.2 – Symbol of Beam-Warming scheme for $\lambda = 1.8$.



Proposition 5.28. *The Beam-Warming scheme is Cauchy-stable if and only if* $0 < \lambda \leqslant 2$.

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

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

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

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

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

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

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

5.4.5 Kreiss-Lopatinskii determinant computation for Beam-Warming scheme

First, we compute the Kreiss-Lopatinskii determinant Δ_{KL} from Definition 5.11 for the Beam-Warming scheme with S2ILW3 boundary condition as in (5.25). Assuming that the the roots of (5.11) are distinct for a given $|z| \geqslant 1$, we have

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

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

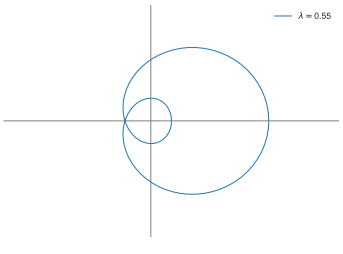
$$\Delta_{\mathrm{KL}}(z) = \det\begin{pmatrix} \kappa_1^{-2} - 2 + 4\kappa_1 - 2\kappa_1^2 & -2\kappa_1^{-2} + 4\kappa_1 - 4\kappa_1^2 \\ \kappa_1^{-1} - \frac{1}{2} + \kappa_1 - \frac{\kappa_1^2}{2} & -\kappa_1^{-1} + \kappa_1 - \kappa_1^2 \end{pmatrix}.$$

In the rest of this section, we continue the example of the Beam-Warming scheme (5.25) so as to illustrate practically the algebraic transformation set up in Lemma 5.21.

For that scheme, the corresponding $\mathcal{Z}$-transformed equation (5.10) is, for $j \in \mathbb{N}$,

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

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



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

$$\widetilde{U}_{j-2}(z) = \alpha \widetilde{U}_{j-1}(z) + \beta \widetilde{U}_j(z). \tag{5.26}$$

The considered boundary condition involves the following matrix:

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

with dimensions $r = 2$ and $N = 5$. With the notations in the proof of Lemma 5.21, let us now construct the matrix $C(z) = B^{(3)}[1:2, 4:5]$. To that aim, we transform successively the matrix B so as to keep unchanged the vector $B\left(\widetilde{U}_{j-2}(z)\widetilde{U}_{j-1}(z)\widetilde{U}_j(z)\widetilde{U}_{j+1}(z)\widetilde{U}_{j+2}(z)\right)^\top$ thanks to the recurrence relation (5.26). Hereafter are the steps:

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

From there, it follows that

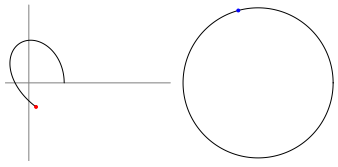
$$C(z) = \begin{pmatrix} 4+\alpha\beta+\alpha(-2+\beta+\alpha^2) & -2+\beta(-2+\beta+\alpha^2) \\ 1+\beta+\alpha(-\frac{1}{2}+\alpha) & -\frac{1}{2}+\beta(-\frac{1}{2}+\alpha) \end{pmatrix},$$

and thus

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

The intrinsic Kreiss-Lopatinskii determinant explicit formula (5.20) (with here $m = 3$ and $r = 2$) is the following:

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



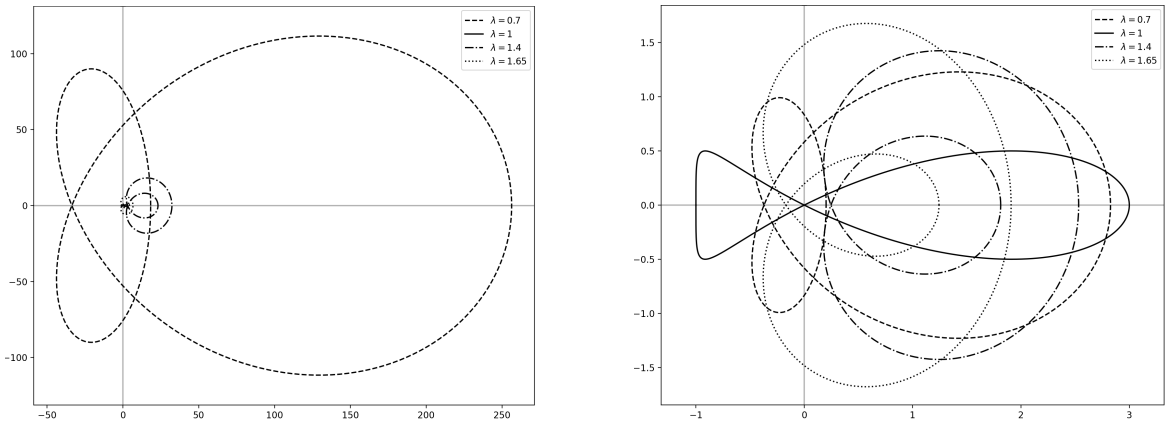


Figure 5.3 – Kreiss-Lopatinskii Determinant Δ when z is on $\mathbb{S}$ for scheme (5.25) for $\lambda \in \{0.7, 1, 1.4, 1.65\}$ (left) and the rescaled one $a_{-2}^2\Delta$ (right).

On Figure 5.3, the curve $\Delta(\mathbb{S})$ is represented successively for different values of the CFL parameter λ. The goal is to compute the winding number of 0, concerned with Corollary 5.15 in order to tackle stability thanks to the Theorem 5.3 (Kreiss). A premultiplication of the quantity Δ by a_{-2}^2 may reduce the order of magnitude of the curves, without changing the winding number. The left and right figures correspond to the case with or without rescaling.

By Theorem 5.15, we have $\#\text{zeros}_\Delta = r - \text{Ind}_{\Delta(\mathbb{S})}(0)$ but after dividing Δ by z^r:

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

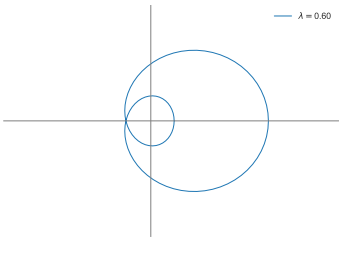
we obtain $\#\text{zeros}_\Delta = -\text{Ind}_{\mathring{\Delta}(\mathbb{S})}(0)$, because $\text{Ind}_{\mathring{\Delta}(\mathbb{S})}(0) = \text{Ind}_{\Delta(\mathbb{S})}(0) - r$, see Figure 5.4.

A particular situation occurs for $\lambda = 1$, since $a_{-2} = 0$ and assumption (H0) fails if we consider $r = 2$. In that case, the equation (5.10) reads $\widetilde{U}_{j-1}(z) = z\widetilde{U}_j(z)$ which is the Beam-Warming scheme for $\lambda = 1$ after $\mathcal{Z}$-transform. Finally, in that case, we find $\det C(z) = (\frac{1}{2}+z(-1+z(\frac{1}{2}-z)))$ that we must multiply by $\frac{1}{\beta^2} = \frac{1}{z^2}$ to find the Kreiss-Lopatinskii determinant (because $m = 3$, $r = 1$ and $\beta = \frac{z-a_0}{a_{-1}} = z$).

All these computations can be done for different boundary conditions and after drawing the curves, the winding number can be computed, as explained in Section 5.4.1, to tackle stability and that the purpose of the following subsection.

5.4.6 Numerical illustration

Figure 5.4 may help to tackle the stability of the scheme (5.25) as we said in Section 5.2.4, indeed, as we said in Section 5.4.1, one can compute the winding number using a numerical procedure [GZDM12] and draw the winding number with respect to λ, as seen in Figure 5.5 for the case S2ILW3. It simplifies the observation of the number of zeros of the Kreiss-Lopatinskii determinant. Hence, the numerical experiments indicate that the scheme (5.25) is strongly stable



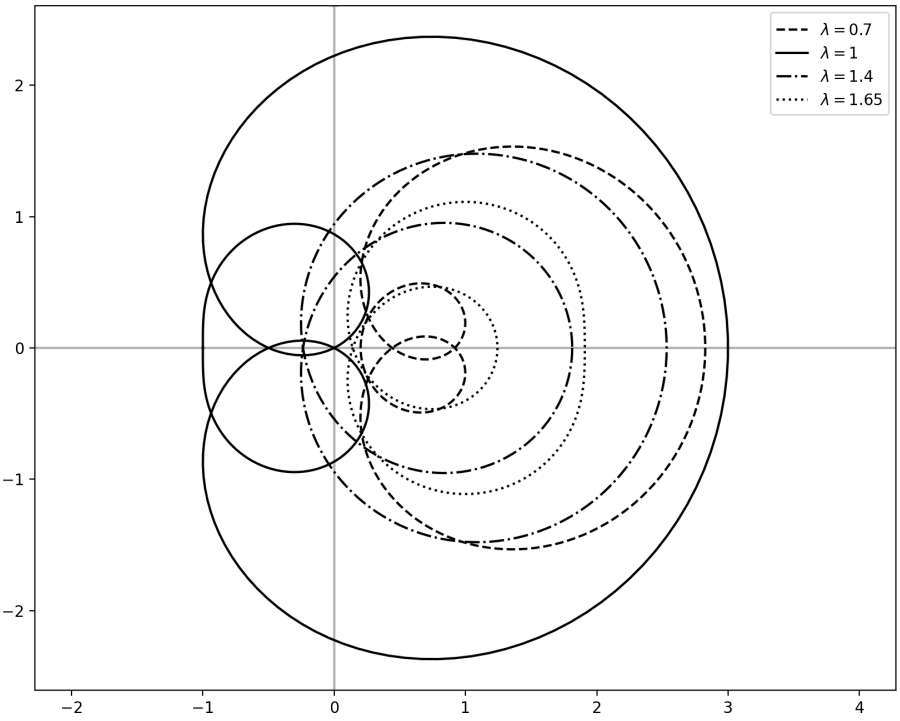


Figure 5.4 – Rescaled Kreiss-Lopatinskii determinant $\frac{a^2_{-2}\Delta}{z^2}$ for z in $\mathbb{S}$.

for $\lambda \in]0, 1[$ but also for $\lambda \in]1.52, 1, 78[$ approximately, but is unstable outside these domains.

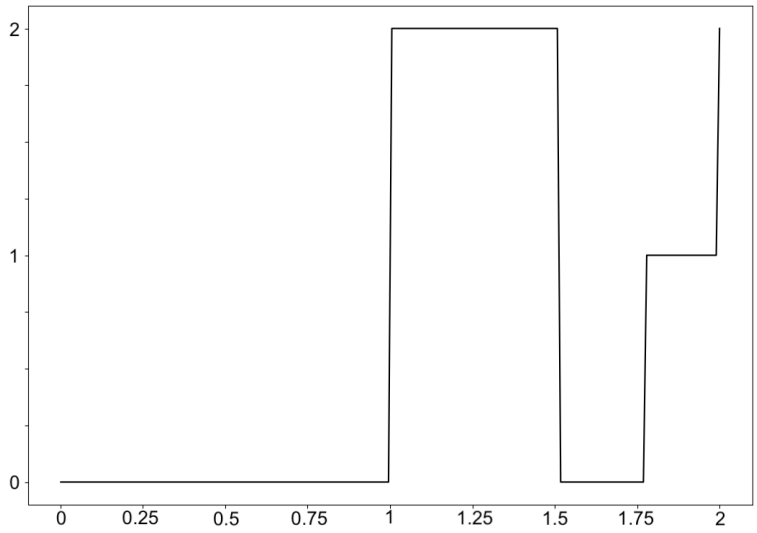
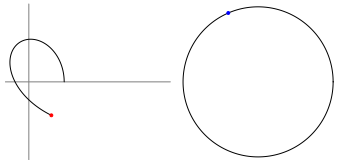


Figure 5.5 – Number of zeros of Kreiss-Lopatinskii determinant with respect to λ for Beam-Warming scheme (5.25) with S2ILW3 boundary condition.

Moreover, instead of taking the Y-axis to represent the number of zeros of the Kreiss-Lopatinskii determinant and having a step function, one can draw areas and compute it for other simplified inverse Lax-Wendroff boundary conditions (defined by the equation (5.24)) as done in Figure 5.6. Note that the stability domain contains a full interval of the form $]0, \lambda_\star[$



(except for the S1ILW4 scheme), but also another disjoint interval included in $]1, 2[$ (except for the S2ILW4 scheme). This property may be used to increase the speed of the computations.

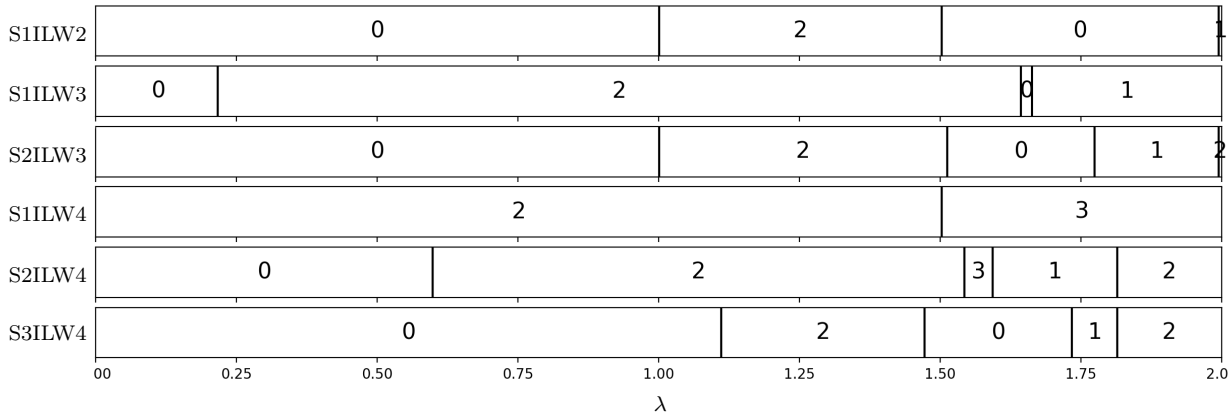


Figure 5.6 – Number of zeros of Kreiss-Lopatinskii determinant for Beam-Warming scheme with different simplified inverse Lax-Wendroff boundary with respect to λ.

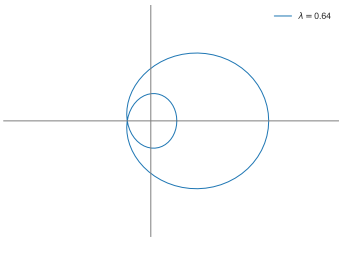
All the figures can be easily computed in Python with the common NumPy [HMvdW+20] library. The algorithm is really quick (less than one minute of computation achieved on a standard laptop). Moreover, our procedure provides sharp results, directly available on ℓ^2. In particular, contrary to numerical investigations of stability which are based on the computation of the spectral radius, no arbitrary truncation of (quasi-)Toeplitz matrices is needed.

5.4.7 Misalignment between boundaries and grid points

Motivated for example by solving multidimensional problems discretized on a cartesian grids, or of one-dimensional problems with moving boundaries as well, a usual idea consists in extrapolating the physical boundary condition to the first boundary points. This idea may be combined with the inverse Lax-Wendroff procedure in order to improve the accuracy at the boundary, see [DDJ18], [VS15] and [LSZ16]. As an archetype for such a situation, we consider hereafter a simple misalignment between the left physical boundary and the first numerical grid point. The advection equation (5.1) is set on the space domain $[x_\sigma, 1]$:

$$\begin{cases} \partial_t u + a\partial_x u = 0, & t \geqslant 0, x \in [x_\sigma, 1], \\ u(t, x_\sigma) = g(t), & t \geqslant 0, \\ u(0, x) = f(x), & x \in [x_\sigma, 1]. \end{cases} \tag{5.27}$$

The space discretization $j\Delta x$ for $j \in \mathbb{Z}$, does not take into account the point x_σ, so that one may write $x_\sigma = (j_\sigma + \sigma)\Delta x$ for some integer $j_\sigma \in \mathbb{Z}$ and the gap (generally nonzero) $\sigma \in [-\frac{1}{2}, \frac{1}{2}[$. The scheme (5.4)-(5.5)-(5.6) is then implemented for $j \geqslant j_\sigma$ only, but with r ghost points at $j_\sigma - 1, \ldots, j_\sigma - r$. For simplicity in the presentation and by translational invariance, we assume from now on that $j_\sigma = 0$. We obtain the discretization represented on Figure 5.7.



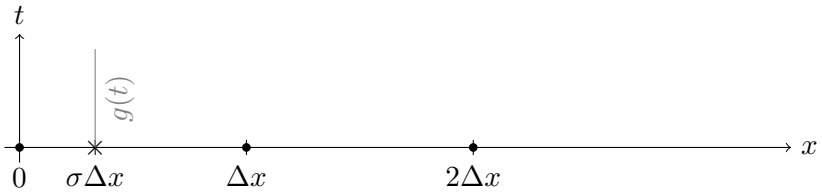


Figure 5.7 – Representation of the mesh.

As explained above, because of the misalignment between the mesh and the boundary position, the simplified inverse Lax-Wendroff procedure (5.24) presented above has to be slightly adapted (see [VS15]). The numerical boundary condition reads

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

We perform the stability analysis of the above scheme, according to the values of both the CFL parameter λ and the gap parameter σ. For example, with the Beam-Warming scheme (5.25) supplemented with the numerical boundary condition S2ILW3 at the point x_σ, the procedure based on the Kreiss-Lopatinskii determinant counts the number of zeros of the Kreiss-Lopatinskii determinant. The corresponding results are represented on Figure 5.8. Of course, on the line $\sigma = 0$, we recover the results obtained on Figure 5.5.

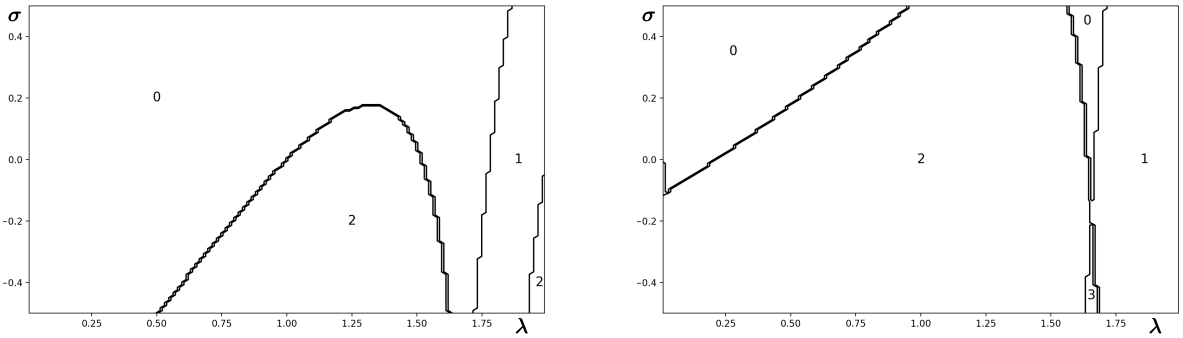
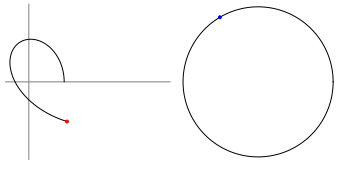


Figure 5.8 – Stability of the Beam-Warming (5.25) with S2ILW3 boundary condition (left) and with S1ILW3 boundary condition (right).

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



parameter σ and where the first reference grid cell is again $x_\sigma = 0$ ($j_\sigma = 0$).

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

In the simulations, the velocity is $a = 1$, the boundary condition is $g(t,y) = e^{-200(t-0.25)^2}$ and the initial datum is $f \equiv 0$. The numerical solution is computed at time $T = 0.3$ using the Beam-Warming scheme with S2ILW3, and with $N = 1000$ grid points in the (truncated) x-direction. The Figure 5.9 represents the amplitude of the numerical solution with respect to the space variable x and to the gap $\sigma = y$, the discrete solution being truncated beyond the value 1 so that unstable boundary oscillations appear as white areas. The two black lines represents the computational domain of Figure 5.8 to confront the left image of Figure 5.8 and the images of Figure 5.9.We observe a good agreement between the corresponding stable/unstable values of σ in Figure 5.8 and Figure 5.9.

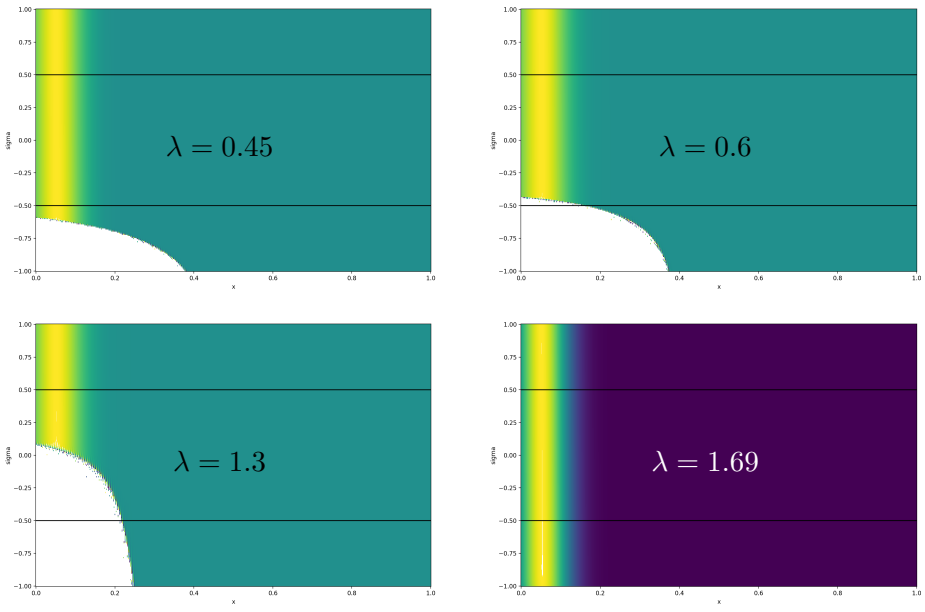
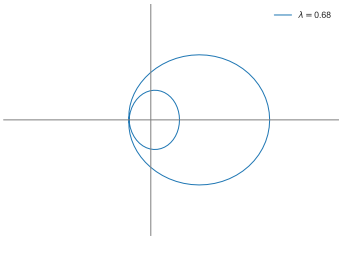


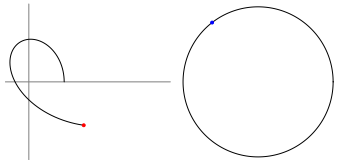
Figure 5.9 – Numerical simulation of Beam-Warming scheme with S2ILW3 for CFL number $\lambda \in \{0.45, 0.6, 1.3, 1.69\}$.

5.5 Future directions

The main drawback of the present theoretical and numerical results is the restriction to the class of totally upwind schemes. This assumption enables a specific simple analysis of the Kreiss-Lopatinskii determinant, using the explicit formula (5.20), and a numerical strategy to conclude



to the existence of eigenvalue or generalized eigenvalue. In this way, it answers the stability issue. This is only an initial effort on the method of designing efficient and automatic numerical tools for stability analysis based on the Kreiss-Lopatinskii determinant. A first extension of the present work is the extension to the case of one-time step explicit schemes without the totally upwind assumption that limits the application of our approch to second-order schemes, see Iserles [IS83]. Such an extension is natural but not straightforward because of the loss of Lemma 5.21: the intrinsic Kreiss-Lopatinskii determinant cannot be reduced easily into a formulation involving square matrices. Another challenging issue is the treatment of multistep schemes and multistep boundary conditions as well. In this direction, explicit schemes may be the most practicable case because many theoretical tools remain available (Hersh, Kreiss...). The difference is the dependence on z in the coefficients of the characteristic equation (5.11). Indeed, each coefficient is a polynomial in z of degree s where s is the number of time steps. Hence, an explicit formula of a Kreiss-Lopatinskii is more difficult to compute. In another direction, for implicit schemes or for more general boundary conditions, such as absorbing boundary conditions [EM77] and [Ehr10] or transparent boundary conditions [AES03] and [Cou19], it seems to be even more challenging to have a such easy-to-use theory.



5.6 Compléments

5.6.1 Preuve des résultats d'holomorphie

Dans la démonstration du Corollaire 5.15, on utilise le principe de l'argument énoncé et démontré en Théorème A.3 (page 200) avec les notations $a = 0$, $r = 1$, $\Omega = B(0,1) = \mathbb{D}$, $\Gamma = \mathbb{S}$ et $f = \Delta$.

La démonstration du Lemme 5.4 (Hersh) utilise le théorème de Rouché que l'on rappelle en Théorème A.5 (page 202). Une version plus générale en est donnée en Lemme 3.4 et en Lemme 6.3 (page 147) et il est démontré page 74.

Le Lemme 5.26 est démontré dans l'Annexe A au Lemme A.4 (page 201).

On donne ici une preuve du Lemme 5.25 de l'article que l'on traduit ici de la manière suivante.

Lemme 5.29. *Soient deux polynômes P et Q avec $\deg P > \deg Q$. Si la fonction $z \mapsto \frac{P(z)}{Q(z)}$ est holomorphe sur $\mathcal{U}$ alors $z \mapsto \frac{P(1/z)}{Q(1/z)}$ est méromorphe sur $\mathbb{D}$ avec un seul pôle en 0 d'ordre $\deg P - \deg Q$.*

Démonstration. Soient $P(X) = \sum_{i=0}^{n} a_i X^i$ et $Q(X) = \sum_{i=0}^{m} b_i X^i$ avec $n > m$. Comme la fonction $z \mapsto \frac{P(z)}{Q(z)}$ est holomorphe sur $\mathcal{U}$, alors la fonction $z \mapsto \frac{P(1/z)}{Q(1/z)}$ est holomorphe sur $\mathbb{D}^*$. Le seul pôle potentiel de la fonction $z \mapsto \frac{P(1/z)}{Q(1/z)}$ est en 0.

On a

$$\frac{P(1/z)}{Q(1/z)} = \frac{\frac{1}{z^n}\sum_{i=0}^{n} a_i z^{n-i}}{\frac{1}{z^m}\sum_{i=0}^{m} b_i z^{m-i}} = \frac{1}{z^{n-m}} \frac{\sum_{i=0}^{n} a_i z^{n-i}}{\sum_{i=0}^{m} b_i z^{m-i}}.$$

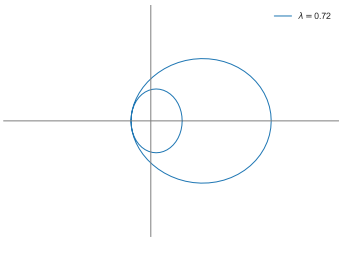
Donc $z^{n-m}\frac{P(1/z)}{Q(1/z)} \xrightarrow[z\to 0]{} \frac{a_n}{b_m} \neq 0$, car les deux coefficients dominants de P et Q sont non nuls par hypothèse. Ainsi, la fonction $z \mapsto \frac{P(z)}{Q(z)}$ est méromorphe sur $\mathbb{D}$ avec un unique pôle en 0 d'ordre $n - m$. □

5.6.2 Réflexion sur les coefficients du schéma

On a besoin de localiser le coefficient a_0 afin que la division par $(a_0 - z)$ dans l'expression (5.15) ne soit pas problématique. Pour cela, on va utiliser l'hypothèse de consistance qui est supposée dans le Théorème 5.13. Pour commencer, on rappelle la propriété suivante des schémas consistants.

Proposition 5.30. *Si le schéma (1.2a) est consistant avec (1.1) d'ordre au moins 1, alors*

$$\sum_{k=-r}^{p} a_k = 1 \quad \text{et} \quad \sum_{k=-r}^{p} k a_k = -\lambda.$$



Démonstration. Soit u une solution régulière de (1.1). On l'injecte dans le schéma :

$$\begin{aligned}
&u(n\Delta t+\Delta t, j\Delta x) - \sum_{k=-r}^{p} a_k u(n\Delta t, j\Delta x + k\Delta x)\\
&= u(n\Delta t, j\Delta x) + \Delta t \partial_t u(n\Delta t, j\Delta x) + O(\Delta t^2)\\
&\qquad - \sum_{k=-r}^{p} a_k \left(u(n\Delta t, j\Delta x) + k\Delta x \partial_x u(n\Delta t, j\Delta x) + o(\Delta x)\right)\\
&= u(n\Delta t, j\Delta x) - a\Delta t \partial_x u(n\Delta t, j\Delta x) + O(\Delta t^2)\\
&\qquad - \sum_{k=-r}^{p} a_k \Big(u(n\Delta t, j\Delta x) + k\Delta x \partial_x u(n\Delta t, j\Delta x) + O(\Delta x^2)\Big)
\end{aligned}$$

Comme le schéma est au moins d'ordre 1, l'erreur de consistance est $O(\Delta x^2 + \Delta t^2)$. On peut donc identifier les termes constants et les termes devant $\partial_x u(n\Delta t, j\Delta x)$. On a alors

$$1 = \sum_{k=-r}^{p} a_k \quad \text{et} \quad -a\Delta t = \sum_{k=-r}^{p} a_k k\Delta x.$$

On conclut par définition de λ. □

Remarque 5.31. En montant en ordre de consistance, on a les identités suivantes :

$$\sum_{k=-r}^{p} a_k = 1, \quad \sum_{k=-r}^{p} k a_k = -\lambda, \quad \sum_{k=-r}^{p} k^2 a_k = \lambda^2, \quad \sum_{k=-r}^{p} k^3 a_k = -\lambda^3, \quad \sum_{k=-r}^{p} k^4 a_k = \lambda^4, \text{ etc.}$$

Le lemme suivant donne une démonstration plus détaillée du Lemme 5.27.

Lemme 5.32. *Si un schéma de la forme (1.2a) est Cauchy-stable et consistant au moins à l'ordre 1, on a $|a_0| < 1$.*

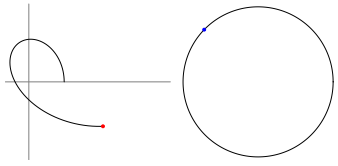
Démonstration. **Étape 1 :** *Montrons que $|a_0| \leqslant 1$.*
On sait que

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

Par inégalité triangulaire et par Cauchy-stabilité, on en déduit

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

Étape 2 : *Montrons que le cas d'égalité amène à une contradiction.*
Pour avoir $|a_0| = 1$, il faut avoir le cas d'égalité dans l'inégalité triangulaire complexe. Ainsi, il faut que $\sum_{k=-r}^{p} a_k e^{ik\xi} = \alpha g(\xi)$ où $\alpha \in \mathbb{C}$ et g est une fonction à valeurs réelles positives



(Prop 7.4, [BP12, p.124]). En appliquant en $\xi = 0$, on trouve

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

Cette égalité permet de montrer que le coefficient α est en fait réel. Cela veut dire que, pour tout $\xi \in \mathbb{R}$, $\sum_{k=-r}^{p} a_k e^{ik\xi}$ est réel. Ainsi, on a $\sum_{k=-r}^{p} a_k e^{ik\xi} = \sum_{k=-r}^{p} a_k e^{-ik\xi}$ (égalité avec son conjugué). En faisant un changement d'indice, on a $\sum_{k=-r}^{p} a_k e^{ik\xi} - \sum_{k=-p}^{r} a_{-k} e^{ik\xi} = 0$ et par unicité des coefficients de Fourier, on trouve $p = r$ et $a_k = a_{-k}$ pour tout $k \in [\![1 : r]\!]$. Cela permet d'obtenir

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

ce qui mène à la contradiction puisque $\lambda \neq 0$ par définition. Ainsi, le cas $|a_0| = 1$ est impossible, ce qui conclut la preuve de ce théorème. □

5.6.3 Généralisation du Lemme 5.21

Pour tout $z \in \overline{\mathcal{U}}$, on considère l'espace vectoriel $\mathcal{E}^s(z)$ des solutions $(\widetilde{U}_j(z))_{j \geqslant -r}$ qui sont dans $\ell^2 \overset{\text{def}}{=} \ell^2([\![-r : -1]\!] \cup \mathbb{N})$ de l'équation

$$z\widetilde{U}_j(z) = \sum_{k=-r}^{p} a_k \widetilde{U}_{j+k}(z), \tag{5.28}$$

pour tout $j \geqslant 0$ avec $a_{-r} \neq 0$ et $a_p \neq 0$.

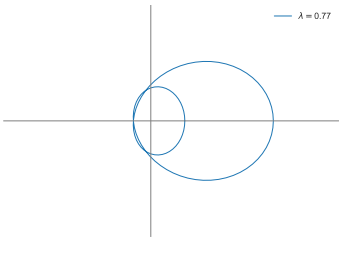
Ici, on considère un schéma avec potentiellement $p \neq 0$. On peut alors généraliser le Lemme 5.21 de la manière suivante.

Lemme 5.33. *Soit $B \in \mathcal{M}_{r,N}(\mathbb{C})$ une matrice vérifiant $N \geqslant r + p$. Il existe une unique matrice $C(z) \in \mathcal{M}_{r,r+p}(\mathbb{C})$ telle que*

$$B \sim_{\mathcal{E}^s(z)} \left(\begin{array}{ccc|c} 0 & \cdots & 0 & \\ \vdots & & \vdots & C(z) \\ 0 & \cdots & 0 & \end{array}\right) \begin{matrix} \\ \\ \updownarrow r \end{matrix}$$
$$\underbrace{\qquad\qquad}_{N-(r+p)} \quad \underbrace{\qquad}_{r+p}$$

De plus, les coefficients de $C(z)$ sont des polynômes en z.

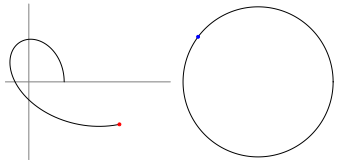
Démonstration. C'est la même preuve que celle du Lemme 5.21 mais sans l'hypothèse de récurrence (e) (qui n'a plus de sens car la matrice n'est plus carrée). Dans le corps de la récurrence, la matrice $B^{(j+1)}$ est définie par $B^{(j+1)} \overset{\text{def}}{=} B^{(j)} - \widetilde{B^{(j)}}$ avec



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

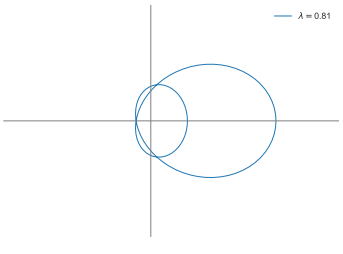
□

Dans le cas où $p \neq 0$, le Lemme 5.33 ne permet pas d'obtenir une matrice carrée. Cela a motivé l'étude du cas particulier $p = 0$, car le fait d'avoir une matrice carrée permet d'utiliser la propriété de multiplicativité du déterminant de Kreiss–Lopatinskii (voir Remarque 5.22). Le problème vient du fait qu'on utilise une suite récurrente linéaire (5.28) d'ordre $p+r$ pour réduire la matrice B initiale. Cependant, comme précisé en Remarque 5.22, les matrices résultant de ce lemme sont toujours multipliées par des matrices $K_{-r,j}(z)$ contenant uniquement les r racines de l'équation caractéristique (3.4) appartenant au disque unité. Concrètement, les p racines de (5.28) se trouvant dans $\mathcal{U}$ ne sont pas utiles. On voudrait donc séparer la suite récurrente linéaire d'ordre $p+r$ en deux suites récurrentes linéaire : l'une d'ordre r et l'autre d'ordre p. On se servirait uniquement de celle d'ordre r pour réduire la matrice B initiale. C'est la stratégie mise en place dans le Lemme 6.16 (page 153).



Chapitre 6

Stability of one-step explicit schemes



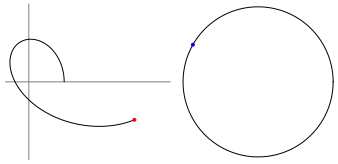
Ce chapitre est dédié à l'article [BLBS23b] : *Stability of finite difference schemes for the hyperbolic initial boundary value problem by winding number computations* écrit avec Benjamin Boutin et Nicolas Seguin et soumis pour publication.

Résumé.

Dans cet article, on présente une stratégie numérique permettant de conclure sur la stabilité forte d'un schéma explicite à un pas avec condition de bord numérique. Toute l'étude est faite sur le modèle jouet de l'équation de transport en dimension 1. La stabilité forte est étudiée sous le prisme de la théorie GKS et du théorème de Kreiss. On introduit un nouvel outil : le déterminant intrinsèque de Kreiss–Lopatinskii qui possède de meilleures propriétés de régularité que le déterminant de Kreiss–Lopatinskii classique. En utilisant des résultats standards d'analyse complexe et d'algèbre linéaire, on peut relier le caractère stable du schéma au calcul de l'indice complexe d'une courbe, ce qui est numériquement robuste et peu coûteux. Cette étude est illustrée par le schéma O3 et le schéma de Lax-Wendroff à l'ordre 5 muni de bords définis par une procédure de reconstruction expliquée dans [DDJ18].

La Section 6.6 (page 167) donne des preuves plus complètes des résultats de l'article. Notamment, on mentionne les preuves du principe de l'argument qui sert dans la démonstration du Corollaire 6.14 et du théorème de Rouché utile pour la démonstration du Lemme 6.3 (Hersh) et du Lemme 6.22. Afin de préciser la preuve de la Proposition 6.18 et la preuve de la Proposition 6.19, on justifie la régularité de l'intégrale à paramètre dans le Lemme 6.23. De plus, on ajoute à la preuve du Lemme 6.17 la justification rigoureuse du lien entre la matrice de Hermite $H_{0,j}(z)$ et la matrice $K_{0,j}(z)$ à la Section 6.6.2. Pour conclure la section, on donne une preuve algébrique de la Proposition 6.20 dans le cas où la matrice $K_{0,j}(z)$ est une matrice de Vandermonde classique, autrement dit, dans le cas où les racines $(\kappa_j(z))_{j=1}^r$ de l'équation caractéristique (3.4) appartenant au disque unité sont simples. De plus, le Lemme 6.22 est démontré intégralement dans la Section 7.1.1 (page 177) du Chapitre 7 sur l'implémentation numérique des schémas.

Pour conclure le chapitre, la Section 6.7 (page 172) donne une extension des résultats de l'article dans le cas de schémas multipas et est illustrée par le schéma leap-frog.



STABILITY OF FINITE DIFFERENCE SCHEMES FOR THE HYPERBOLIC INITIAL BOUNDARY VALUE PROBLEM BY WINDING NUMBER COMPUTATIONS

Abstract.

In this paper, we present a numerical strategy to check the strong stability (or GKS-stability) of one-step explicit finite difference schemes for the one-dimensional advection equation with an inflow boundary condition. The strong stability is studied using the Kreiss–Lopatinskii theory. We introduce a new tool, the intrinsic Kreiss–Lopatinskii determinant, which possesses the same regularity as the vector bundle of discrete stable solutions. By applying standard results of complex analysis to this determinant, we are able to relate the strong stability of numerical schemes to the computation of a winding number, which is robust and cheap. The study is illustrated with the O3 scheme and the fifth-order Lax-Wendroff (LW5) scheme together with a reconstruction procedure at the boundary.

6.1 Introduction

6.1.1 Motivations and assumptions

The purpose of this work is to establish an efficient numerical strategy to determine whether a given finite difference method on the half line is stable or not. We work on an approximation of the rightgoing linear transport equation set on the positive real axis:

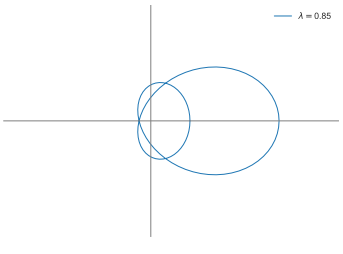
$$\begin{cases} \partial_t u + a\partial_x u = 0, & t \geqslant 0, x \geqslant 0, \\ u(t,0) = g(t), & t \geqslant 0, \\ u(0,x) = f(x), & x \geqslant 0, \end{cases} \tag{6.1}$$

where $u(t,x) \in \mathbb{R}$ is the unknown, f an initial datum at time $t = 0$, g is a prescribed physical boundary datum at the point $x = 0$ which corresponds to the inflow boundary because the velocity a is assumed to be positive $a > 0$.

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

$$U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{j+k}^n, \tag{6.2}$$

with integers $r, p \geqslant 1$ and a_p, a_{-r} non zero. The case where $p = 0$ or $r = 0$ will be discussed in Section 6.1.2. Here, the unknown of the scheme U_j^n is expected to approximate the quantity $u(n\Delta t, j\Delta x)$. The time step $\Delta t > 0$ and the space step $\Delta x > 0$ are usually chosen with respect to some CFL condition $\lambda = a\Delta t/\Delta x \leqslant \lambda_{\mathsf{CFL}}$ discussed later on.



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

As a central idea in numerical analysis, the Lax equivalence theorem [LR56] asserts that a consistent scheme is convergent if and only if it is stable. Therefore, all along the paper only consistent numerical schemes are considered and the discussion concentrates only on their stability issues. The Cauchy-stability for the space-periodic problem is handled with the Fourier symbolic analysis, the so-called Von-Neumann stability analysis (see [CFL28] and [CN47]) and makes use of the symbol γ. The *symbol* associated with the scheme (6.2) is defined, for $\xi \in \mathbb{R}$, by

$$\gamma(\xi) = \sum_{k=-r}^{p} a_k e^{ik\xi}. \tag{6.3}$$

Assumption (H1). The scheme (6.2) is *Cauchy-stable*, meaning that the symbol γ satisfies $|\gamma(\xi)| \leqslant 1$ for all $\xi \in \mathbb{R}$.

When dealing with discrete schemes set over the full line $j \in \mathbb{Z}$, the algebraic characterization of the Cauchy-stability follows classically from the Fourier analysis but in the scalar case, it reduces to a geometric property concerning the *symbol curve* Γ which is a closed complex parametrized curve defined by

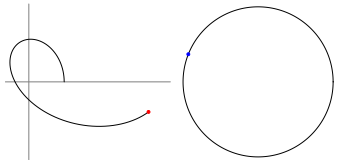
$$\Gamma = \{\theta \in [0, 2\pi] \mapsto \gamma(\theta)\}.$$

This curve enables a geometric interpretation of the Cauchy-stability assumption (H1) reformulated equivalently using the inclusion $\Gamma \subset \overline{\mathbb{D}}$. The stability condition (H1) can be easily illustrated graphically in the complex plane. In some sense, our goal is to extend this kind of graphical study when including the numerical boundary conditions.

For solving the Initial Boundary Value Problem (IBVP) (6.1) with the discrete scheme (6.2), r additional ghost points are needed to take into account the left boundary condition and to fully define the discrete approximation. We assume that the values at these ghost points are obtained from a linear combination of the first values of the solution close to the boundary and at the same time step. More clearly, the considered numerical schemes reads

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{k=-r}^{p} a_k U_{k+j}^n, & j \in \mathbb{N},\ n \in \mathbb{N}, \qquad (6.4) \\ U_j^n = \displaystyle\sum_{k=0}^{m-1} b_{j,k} U_k^n + g_j^n, & j \in [\![-r : -1]\!],\ n \in \mathbb{N}, \qquad (6.5) \\ U_j^0 = f_j, & j \in \mathbb{N}, \qquad (6.6) \end{cases}$$

where the integer m satisfies $p + r \leqslant m$, $(f_j)_j$ are approximations of the initial condition f and $(g_j^n)_{n,j}$ are numerical data related to the boundary datum g and possibly its derivatives (see for



instance the example in Section 6.4.3). The assumption $p + r \leqslant m$ is not restrictive since some of the coefficients $b_{j,k}$ are possibly zero.

In order to define the stability on $\ell^2(\mathbb{N})$ and for the sake of convenience in the Kreiss-Lopatinskii determinant formulation (see Definition 6.10), the explicit use of the r ghost points U_j^n, for $j \in [\![-r : -1]\!]$, can be avoided by substituting the r boundary condition (6.5) into the recurrence formula (6.4) for $j \in [\![0 : r-1]\!]$. After straightforward calculations, the boundary part reads also under the form

$$\mathfrak{U}_r^{n+1} = \mathcal{B}\mathfrak{U}_m^n + \mathcal{G}^n \tag{6.7}$$

where we denote

$$\mathfrak{U}_r^{n+1} = \begin{pmatrix} U_0^{n+1} \\ \vdots \\ U_{r-1}^{n+1} \end{pmatrix}, \ \mathfrak{U}_m^n = \begin{pmatrix} U_0^n \\ \vdots \\ U_{m-1}^n \end{pmatrix}, \ \mathcal{G}^n = \begin{pmatrix} a_{-r} & \cdots & a_{-1} \\ & \ddots & \vdots \\ 0 & & a_{-r} \end{pmatrix} \begin{pmatrix} g_{-r}^n \\ \vdots \\ g_{-1}^n \end{pmatrix} \in \mathcal{M}_r(\mathbb{C}).$$

Here, the matrix $\mathcal{B} \in \mathcal{M}_{r,m}(\mathbb{C})$ encodes the boundary treatment in another way. It corresponds to the boundary part of the quasi-Toeplitz matrix form of the scheme used by Beam and Warming [BW93]. In the detail, the explicit relationship between $\mathcal{B}$ and B is as follows:

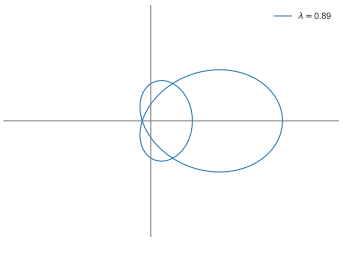
$$\mathcal{B} = \begin{pmatrix} a_{-r} & \cdots & a_{-1} \\ & \ddots & \vdots \\ 0 & & a_{-r} \end{pmatrix} B + \begin{pmatrix} a_0 & \cdots & a_p & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & & \ddots & \ddots & & & \vdots \\ a_{-r+1} & \cdots & a_0 & \cdots & a_p & 0 & \cdots & 0 \end{pmatrix} \in \mathcal{M}_{r,m}(\mathbb{C}) \tag{6.8}$$

with the notation

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

The relation (6.8) is inversible since the coefficient a_{-r} is supposed to be non zero. For example, for the very naive scheme $U_j^{n+1} = \frac{U_{j-1}^n + U_{j+1}^n}{2}$ and the boundary condition $U_{-1}^n = \frac{U_0^n + U_1^n}{2}$, we obtain $B = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ and $\mathcal{B} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \end{pmatrix}$.

This class of boundary conditions, (6.5) or (6.7), encompasses the Dirichlet and Neumann extrapolation procedures, see, for example, the work of Goldberg [Gol77]. This class also takes into account the more general simplified inverse Lax-Wendroff procedure analyzed by Vilar and Shu [VS15] in the framework of central compact schemes, and Li, Shu and Zhang for the advection equation [LSZ16] and for diffusion equations [LSZ17]. We will focus on the so-called reconstruction technique for the boundary condition, which enables to deal with a boundary which is not superposed with a grid point (presented by Dakin, Després and Jaouen [DDJ18] and also in Section 6.4.3) in our numerical examples. Other treatments at the boundary exist, as for example absorbing boundary conditions [EM77] and [Ehr10], or transparent boundary conditions [AES03] and [Cou19], however, in general, they do not enter the present framework.



For finite difference schemes applied to discrete IBVP's, the stability study is a principal issue and is the subject of different approaches. For example, Beam and Warming [BW93] study the spectral properties of the Toeplitz or quasi-Toeplitz representation of the scheme. In the same spirit, the computation of the spectral radius of the truncated (i.e. finite dimensional) quasi-Toeplitz matrix may provide significant information for the power boundedness of the method. This is the method used for example by Dakin, Després and Jaouen [DDJ18]. This strategy is sometimes called *eigenvalue spectrum visualization method*, especially by Li, Shu and Zhang [LLS22, LSZ16, LSZ17]. In Section 6.4, we will compare this latter approach with our own strategy presented hereafter for the O3 scheme in the case of reconstruction boundary conditions. Our strategy is based on the so-called GKS-stability theory introduced by Gustafsson, Kreiss and Sundström [GKS72] which handles the discrete IBVP (6.4)-(6.5)-(6.6) with a zero initial data. The reader can refer to the work by Wu [Wu95] and Coulombel [Cou13] for more recent developments on semigroup estimates in order to deduce the stability of the discrete IBVP (6.4)-(6.5)-(6.6) with non zero initial data from the GKS-stability. The notion of GKS-stability (or also called strong stability) for the boundary problem makes use of the following discrete norms:

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

The so-called strong stability, or GKS-stability, is defined by:

Definition 6.1 (Strong stability)**.** The scheme (6.4)-(6.5)-(6.6) is strongly stable if, for $(f_j) = 0$, there exist $C > 0$ and α_0, such that for all $\alpha > \alpha_0$, for all boundary data (g_j^n), for all $\Delta x > 0$, for all $n \in \mathbb{N}$, the solution satisfies

$$\sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2 + \left(\frac{\alpha - \alpha_0}{\alpha \Delta t + 1}\right) \|e^{-\alpha n \Delta t} U\|_{\Delta x,\Delta t}^2 \leqslant C \sum_{j=-r}^{-1} \|e^{-\alpha n \Delta t} g_j\|_{\Delta t}^2. \tag{6.9}$$

We warn the reader that $\|e^{-\alpha n \Delta t} U_j\|_{\Delta t}^2$ is here an abuse of notation to describe $\sum_{n=0}^{+\infty} \Delta t e^{-2\alpha n \Delta t} |U_j^n|^2$ and similarly for $\|e^{-\alpha n \Delta t} U\|_{\Delta x,\Delta t}^2$. This stability definition admits a similar but continuous form for the solutions to continuous hyperbolic PDE's [BGS06]. Namely, it provides some a priori estimates that are useful for a general analysis of such problems.

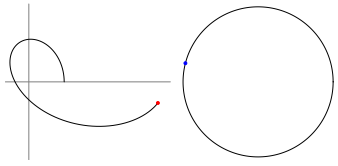
The following Kreiss theorem [Kre68] expresses a necessary and sufficient condition for the strong stability. We provide hereafter a condensed formulation of this theorem, obtained from [GKS72, Thm 5.1] combined with [GKO13, Lem 13.1.4] or with [Gus08, Def 2.23].

Theorem 6.2 (Kreiss)**.** *The following statements are equivalent:*

(i) The scheme (6.4)-(6.5)-(6.6) is strongly stable in the sense of Definition 6.1.

(ii) The Uniform Kreiss-Lopatinskii Condition is satisfied.

The Uniform Kreiss-Lopatinskii Condition corresponds to the absence of zeros for the so-



called Kreiss-Lopatinskii determinant Δ_{KL} that we present here by the informal definition:

$$\Delta_{\mathrm{KL}}(z) = \det(\mathfrak{B}e_1(z), \dots, \mathfrak{B}e_r(z)) \tag{6.10}$$

where $(e_1(z), \dots, e_r(z))$ is a explicit basis of the linear space of the $\ell^2(\mathbb{N})$-stable solutions of the $\mathcal{Z}$-transform of the interior equation (6.4) and $\mathfrak{B}$ is an encoding of the $\mathcal{Z}$-transform of the boundary equation (6.5). For a proper definition of this determinant, the reader can look at Definition 6.10 or the book by Gustafsson, Kreiss and Oliger [GKO13]. Before going on, let us provide some comments to a particular case we already studied.

6.1.2 The case of totally upwind schemes and summary of [BLBS23a]

The present article is a non trivial extension of our previous work [BLBS23a] that deals with the restricted case of totally upwind schemes. Totally upwind schemes are schemes of the form (6.2) with $p = 0$ if $a > 0$ or $r = 0$ if $a < 0$. Without loss of generality, we restrict here the discussion to the case $p = 0$ since flipping the indices may turn a case to the other. In this section, we summarize the result of [BLBS23a] and introduce the novelty of the present work. The first step of the analysis conducted in [BLBS23a] is based on the introduction of the intrinsic Kreiss-Lopatinskii determinant:

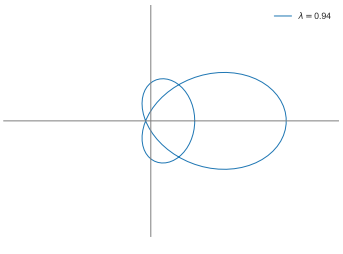
$$\Delta(z) = \frac{\det(\mathfrak{B}e_1(z), \dots, \mathfrak{B}e_r(z))}{\det(e_1(z), \dots, e_r(z))} \tag{6.11}$$

using the same informal notation as in (6.10). Under appropriate assumptions, an explicit formula for the intrinsic Kreiss-Lopatinskii determinant is obtained:

$$\forall |z| \geqslant 1, \quad \Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a_{-r}}{a_0 - z}\right)^{m-r} \tag{6.12}$$

where $\det C(z)$ is an computable polynomial in z depending only on the coefficients $(a_j)_{j=-r}^{0}$ and on $\mathfrak{B}$. Thanks to this result, we prove that Δ is holomorphic on $\overline{\mathcal{U}}$. Note that this property may be wrong as long as the standard Kreiss-Lopatinskii determinant is concerned.

Applying the residue theorem to Δ, we developed a numerical strategy to count the number of zeros of the Kreiss-Lopatinskii determinant in $\mathcal{U}$. By Theorem 6.2 (Kreiss), we conclude that if $\mathrm{Ind}_{\Delta(\mathbb{S})}(0) < r$ then the scheme is not stable where $\mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ is the notation for the winding number of 0 with respect to the Kreiss-Lopatinskii curve $\Delta(\mathbb{S})$. This result allows us to establish an efficient and practical method (see Method 19 of [BLBS23a] or Method 6.15 of the present paper) to study the stability of a scheme with boundary. It provides sharp results for the solution $(U_j^n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$ to the problem set on the half line $\mathbb{N}$. In particular, contrary to numerical investigations of stability which are based on the computation of the spectral radius, no arbitrary truncation of (quasi-)Toeplitz matrices is needed. In return, a problem set on a bounded space domain needs, for a whole convergence study, superposition techniques for



truncated data, as used in [Cou19] and [BNS+21]. This feature restricts mainly the study to explicit scheme.

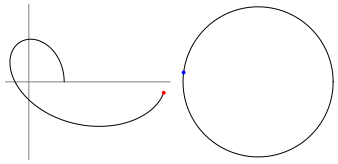
In the present article, we extend to $p \geqslant 1$ the connection between the winding number of 0 and the stability of the scheme. Indeed, even if we do not have an explicit formula as (6.12), the holomorphic property of the intrinsic Kreiss-Lopatinskii determinant is sufficient to have the same efficient method to study the stability. This approach, using the winding number, is robust since instead of finding zeros of an algebraic curve, it only requires the computation of a winding number, which is an integer, to count the number of zeros. One can mention the work of Thuné [Thu86] who develops a numerical method to check the GKS-stability. He looks for the precise location of the zeros of the Kreiss-Lopatinskii determinant approximating the roots of some parameterized characteristic polynomial equations which is significantly different with our work. One can also cite the work of Tadmor and Goldberg [GT78, GT81] which gives a sufficient condition to have the stability of a scheme with boundaries. One can see this condition as a weaker version of the Kreiss-Lopatinskii determinant but it has the advantage of decorrelating the study of the interior equation and the boundary equations. In our paper, the framework is less restrictive since the scheme is not necessary dissipative and the boundary conditions is more general.

6.1.3 Outline of the paper

After constructing the intrinsic Kreiss-Lopatinskii determinant in Section 6.2 and Section 6.3, we see that, in such a general case, the lack of an explicit formula for $\Delta(z)$ does not preclude holomorphic properties (see Theorem 6.12). From there, we obtain the following stability criterion: if $\mathrm{Ind}_{\Delta(\mathbb{S})}(0) < r$ then the scheme is not stable (see Corollary 6.14). We prove these results by the use of Hermite interpolation and residue theorem. To compute the Kreiss-Lopatinskii determinant numerically, in Section 6.4, we use a an easy-to-use formulation of it which is, in some sense, close to the explicit formulation (6.12). Moreover, Section 6.4 gathers the numerical procedure to draw the Kreiss-Lopatinskii curve, several examples and numerical experiments for illustrating the efficiency of the proposed strategy.

6.2 Kreiss-Lopatinskii determinants

In this section, we introduce the Kreiss-Lopatinskii determinant, a usual tool to check the Uniform Kreiss-Lopatinskii Condition. Then we define the intrinsic Kreiss-Lopatinskii determinant, namely a reshaping of the previous one, which is more convenient in practice and has better properties than the classical Kreiss-Lopatinskii determinant: holomorphicity, continuity, independence on the basis...



6.2.1 Stable subspace $\mathcal{E}^s(z)$ and matrix representation

First, we study the solutions to the interior equation:

$$U_j^{n+1} = \sum_{k=-r}^{p} a_k U_{k+j}^n, \ j \in \mathbb{N}, \ n \in \mathbb{N}. \tag{6.13}$$

To study this equation, the $\mathcal{Z}$-transform (see [GW99, Lesson 40]) is applied. This transformation is defined for $(x_n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$ such that $x_0 = 0$ and $z \in \mathcal{U}$ by $\widetilde{x}(z) = \sum_{n\geqslant 0} z^{-n} x_n$. The previous equation then reads

$$z\widetilde{U}_j(z) = \sum_{k=-r}^{p} a_k \widetilde{U}_{j+k}(z), \ j \in \mathbb{N}, \ z \in \mathcal{U}. \tag{6.14}$$

To solve the linear recurrence equation (6.14), let us introduce the following characteristic equation where z plays the role of a parameter and κ is the indeterminate:

$$z\kappa^r = \sum_{k=-r}^{p} a_k \kappa^{r+k}. \tag{6.15}$$

This equation is nothing but the discrete dispersion relation of the finite difference scheme (6.13), with frequency parameter κ in space and z in time. It is formally obtained by looking for solutions to the interior equation (6.13) having the form $U_j^n = z^n \kappa^j$.

In the spirit of a classic result by Hersh [Her63], the following lemma provides a property of separation for the roots with respect to the unit circle.

Lemma 6.3 (Hersh)**.** *Assume (H1). For z in the unbounded connected component of $\mathbb{C} \setminus \Gamma$,*

1. *there is no root of the characteristic equation (6.15) on $\mathbb{S}$,*
2. *there are r roots (with multiplicity) of the characteristic equation (6.15) in $\mathbb{D}$ and p roots (with multiplicity) of the characteristic equation (6.15) in $\mathcal{U}$.*

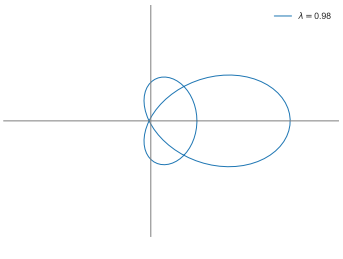
Remark 6.4. Under the Cauchy-stability assumption (H1), the inclusion $\Gamma \subset \overline{\mathbb{D}}$ is known. From there, it follows that the unbounded connected component of $\mathbb{C} \setminus \Gamma$ contains the whole set $\mathcal{U}$ so that a weaker form of the lemma is available for considering $z \in \mathcal{U}$ only. If in addition, the considered scheme is also *dissipative*, meaning that its symbol γ satisfies

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

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

Proof of Lemma 6.3.

1. Assume there exists a root κ of (6.15) on the unit circle, then one can find $\theta \in \mathbb{R}$ such



that $\kappa = e^{i\theta}$. So we have

$$z = \sum_{j=-r}^{p} a_j \kappa^j = \sum_{j=-r}^{p} a_j e^{ij\theta} = \gamma(\theta).$$

This is a contradiction because $z \in \Gamma$ and by assumption $z \in \mathbb{C} \setminus \Gamma$. It concludes the proof.

2. We denote $\mathcal{C}$ the unbounded connected component of $\mathbb{C} \setminus \Gamma$. The polynomial (6.15) has $p+r$ roots (with multiplicity). It is sufficient to count how many roots there are inside the unit disk to deduce the number of roots outside. By continuity of the roots with respect to coefficients and because there is no root on the unit circle for $z \in \mathcal{C}$, we know that there is a constant number of roots inside the unit disk for all $z \in \mathcal{C}$. By Rouché's theorem, one can study the zeros of $f_z(\kappa) = \kappa^r - \frac{1}{z}(a_{-r} + a_{-r+1}\kappa + \cdots + a_p\kappa^{p+r})$ and $g_z(\kappa) = \kappa^r - \frac{1}{z}a_{-r}$ in $\mathbb{D}$ for z sufficiently large to have the result.

□

Lemma 6.3 (Hersh) above is illustrated in Figure 6.1. The first two lines correspond to the Lemma 6.3 (Hersh) and the third one describes the possible configuration for $z \in \Gamma \cap \mathbb{S}$, typically not meeting the assumptions. This case will be the object of a subsequent discussion.

For $|z| > 1$, by Lemma 6.3 (Hersh), the linear subspace of solutions to (6.14) living in $\ell^2(\mathbb{N})$ is generated by the following r vectors:

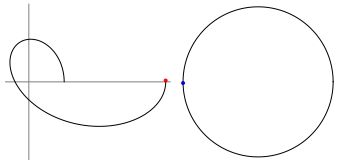
$$\begin{pmatrix} 1 \\ \kappa_\ell \\ \kappa_\ell^2 \\ \kappa_\ell^3 \\ \kappa_\ell^4 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ \kappa_\ell \\ 2\kappa_\ell^2 \\ 3\kappa_\ell^3 \\ 4\kappa_\ell^4 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ \kappa_\ell \\ 2^2\kappa_\ell^2 \\ 3^2\kappa_\ell^3 \\ 4^2\kappa_\ell^4 \\ \vdots \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \kappa_\ell \\ 2^{\beta_\ell - 1}\kappa_\ell^2 \\ 3^{\beta_\ell - 1}\kappa_\ell^3 \\ 4^{\beta_\ell - 1}\kappa_\ell^4 \\ \vdots \end{pmatrix}, \quad \ell = 1, \ldots, M \tag{6.16}$$

where $\kappa_1, \ldots, \kappa_M$ of multiplicity $\beta_1, \ldots, \beta_M$ are the solutions to (6.15) living in $\mathbb{D}$, with $\beta_1 + \cdots + \beta_M = r$ (we omit the z-dependence of $\kappa(z)$ for the sake of readability).

Notation. We denote $\mathcal{E}^s(z)$ the linear subspace of solutions to (6.14) living in $\ell^2(\mathbb{N})$ and $K_{i,j}(z) \in \mathcal{M}_{j-i+1,r}(\mathbb{C})$ the matrix where we put in columns the extraction of all the lines between i and j (included) of the r vectors of (6.16), where $0 \leqslant i \leqslant j$.

Remark 6.5. For $r = 2$, if the solutions to (6.15) are $\kappa_1(z) \neq \kappa_2(z)$, then there are exactly two roots with multiplicity 1. The solutions to (6.14) can be written $\widetilde{U}_j(z) = \alpha_1\kappa_1(z)^j + \alpha_2\kappa_2(z)^j$, and we have

$$K_{0,2}(z) = \begin{pmatrix} 1 & 1 \\ \kappa_1(z) & \kappa_2(z) \\ \kappa_1(z)^2 & \kappa_2(z)^2 \end{pmatrix}.$$



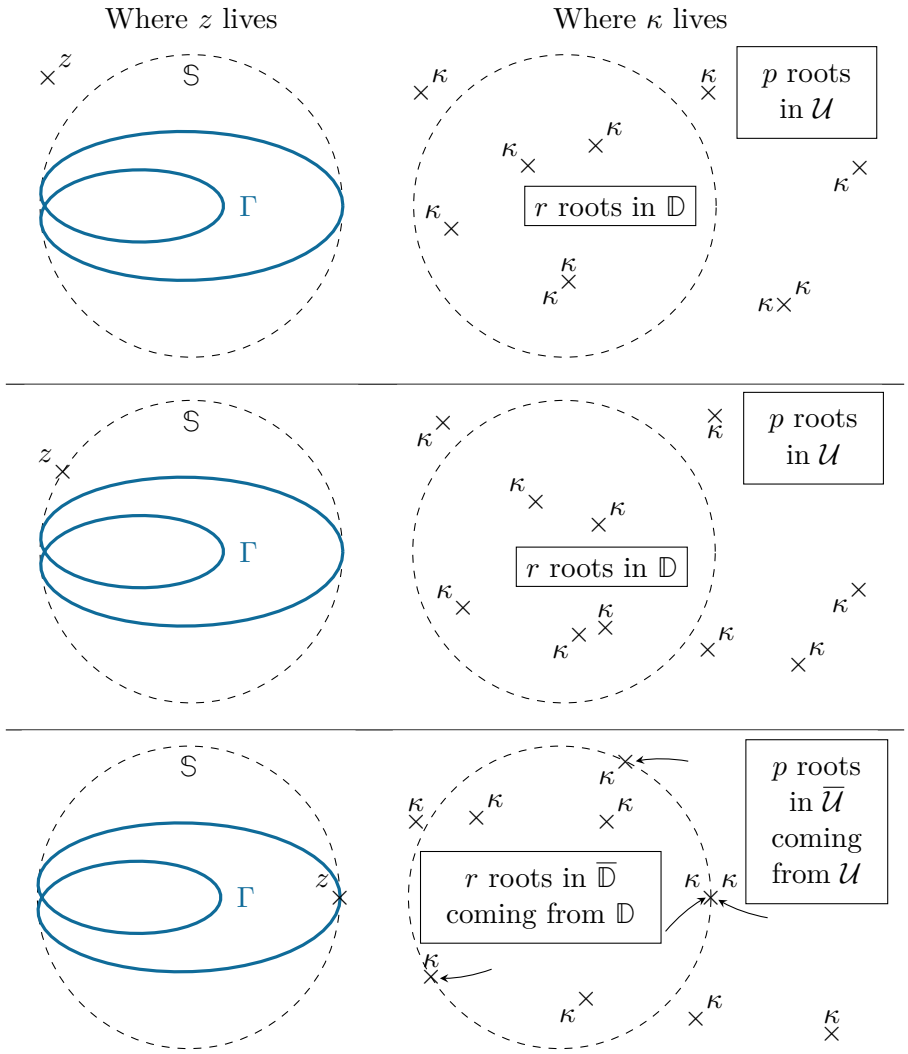


Figure 6.1 – Illustration of Lemma 6.3: case $|z| > 1$ (first line), case $|z| = 1$ and $z \notin \Gamma$ (second line) and case $z \in \Gamma$ where Lemma 6.3 does not hold (third line).

Remark 6.6. Still for $r = 2$, if the solution to (6.15) now is $\kappa(z)$ with multiplicity 2, then the solutions to (6.14) can be written $\widetilde{U}_j(z) = (\alpha_1 + \alpha_2 j)\kappa(z)^j$, and we have

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

We raise awareness of the dependence in z and of the continuity issues because the map $z \mapsto K_{i,j}(z)$ is not continuous whereas the set of roots of (6.15) is a continuous mapping with respect to z. Indeed, the root curves $(\kappa_j(z))_j$ can intersect, when a multiple root occurs. For example, for $r = 2$, if there is $(z_n)_{n\in\mathbb{N}} \subset \mathcal{U}$ with $\kappa_1(z_n) \neq \kappa_2(z_n)$ which converge to $z_\infty \in \mathcal{U}$ such that $\kappa_1(z_\infty) = \kappa_2(z_\infty)$ a double root, then we have

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

but

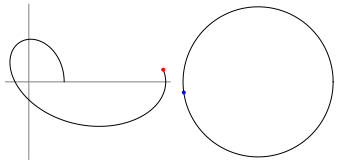
$$K_{0,3}(z_n) = \begin{pmatrix} 1 & 1 \\ \kappa_1(z_n) & \kappa_2(z_n) \\ \kappa_1^2(z_n) & \kappa_2^2(z_n) \\ \kappa_1^3(z_n) & \kappa_2^3(z_n) \end{pmatrix} \not\xrightarrow[n\to\infty]{} K_{0,3}(z_\infty) = \begin{pmatrix} 1 & 0 \\ \kappa_1(z_\infty) & \kappa_1(z_\infty) \\ \kappa_1^2(z_\infty) &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

Consequently, the considered basis (6.16) of $\mathcal{E}^s(z)$ does not generally define a continuous mapping with respect to z.

In spite of the difficulty enlightened above, it turns out that $\mathcal{E}^s(z)$ is a continuous and even holomorphic vector bundle over $\mathcal{U}$ as it is discussed in [Cou13, Thm 4.3]. It is also proved that this vector bundle $\mathcal{E}^s(z)$ can even be continuously extended over $\overline{\mathcal{U}}$, thus considering $z \in \mathbb{S}$ as well (see also [MZ04] for a similar property for the hyperbolic-parabolic PDE case). The main point therein is that for some $z_0 \in \mathbb{S}$, there may exists one (or several) root $\kappa_0(z_0)$ of (6.15) on $\mathbb{S}$. At such points z_0 the Lemma 6.3 (Hersh) does not hold anymore. This situation is depicted on the third line of Figure 6.1. For z on $\mathbb{S}$, the space $\mathcal{E}^s(z)$ still is of dimension r and we extend the notation $K_{i,j}(z)$. We can summarize the above discussion in the following theorem.

Theorem 6.7 ([Cou13])**.** *Under assumption (H1), the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over $\mathcal{U}$ and can be extended in a unique way to a continuous vector bundle over $\overline{\mathcal{U}}$.*

Remark 6.8. For the extension, the first difficulty is to select the roots of (6.15) coming from the inside, indeed, if there is a root on $\mathbb{S}$, it can be coming from the inside of $\mathbb{D}$, the outside or both (in case of multiplicity). In Section 6.4.2, we will explain the numerical strategy to select the good ones. The second difficulty is to prove the continuity of $\mathcal{E}^s(z)$ after the extension, it follows from the existence of a K-symmetrizer and is obtained e.g. in [Cou13, Thm 4.3]. As previously observed, $K_{i,j}(z)$ is generally not continuous with respect to z.



6.2.2 Intrinsic Kreiss-Lopatinskii determinant

In this section, we define properly formulas (6.10) and (6.11). Let us consider the $\mathcal{Z}$-transformed version of the boundary condition (6.7), that is

$$z\begin{pmatrix}\widetilde{U}_0(z)\\ \vdots \\ \widetilde{U}_{r-1}(z)\end{pmatrix} - \mathcal{B}\begin{pmatrix}\widetilde{U}_0(z)\\ \vdots \\ \widetilde{U}_{m-1}(z)\end{pmatrix} = \begin{pmatrix}a_{-r} & \cdots & a_{-1}\\ & \ddots & \vdots \\ 0 & & a_{-r}\end{pmatrix}\begin{pmatrix}\widetilde{g}_{-r}(z)\\ \vdots \\ \widetilde{g}_{-1}(z)\end{pmatrix}. \tag{6.17}$$

Injecting the solution $(\widetilde{U}_j(z))_{j\in\mathbb{N}} \in \mathcal{E}^s(z)$ to (6.14) into (6.17), we obtain a system of r equations with r scalar unknowns: they are the coefficients of $(\widetilde{U}_j(z))_{j\in\mathbb{N}}$ written in the basis (6.16) of $\mathcal{E}^s(z)$.

Remark 6.9. For $r = 2$ and a given value of z (we skip for convenience the dependence in z hereafter), if $\kappa_1 \neq \kappa_2$ so that the solution to (6.17) has the form $\alpha_1\kappa_1^j + \alpha_2\kappa_2^j$, then that solution is constrained by the system (6.17). The matricial form of that system reads

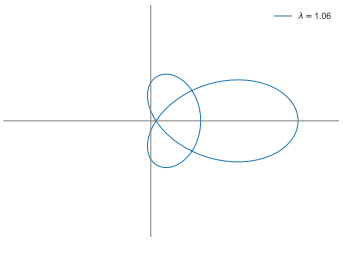
$$\left(z\begin{pmatrix}1 & 1\\ \kappa_1 & \kappa_2\end{pmatrix} - \mathcal{B}\begin{pmatrix}1 & 1\\ \kappa_1 & \kappa_2\\ \kappa_1^2 & \kappa_2^2\\ \vdots & \vdots \\ \kappa_1^{m-1} & \kappa_2^{m-1}\end{pmatrix}\right)\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} = \begin{pmatrix}a_{-1}\widetilde{g}_{-1} + a_{-2}\widetilde{g}_{-2}\\ a_{-2}\widetilde{g}_{-1}\end{pmatrix}.$$

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

Definition 6.10 (Kreiss-Lopatinskii determinant)**.** The *Kreiss-Lopatinskii determinant* is the complex-valued function defined for $|z| \geqslant 1$ by:

$$\Delta_{\mathrm{KL}}(z) = \det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z)).$$

Despite the fact that the space $\mathcal{E}^s(z)$ is a holomorphic vector bundle over $\mathcal{U}$ and continuous over $\overline{\mathcal{U}}$ (Theorem 6.7), this determinant Δ_{KL} is not holomorphic on $\mathcal{U}$. To retrieve those properties, we define the *intrinsic Kreiss-Lopatinskii determinant* Δ that we can motivate by the following informal discussion. The above Kreiss-Lopatinskii determinant is actually not well defined until we order in some way the roots $(\kappa_j(z))_{j=1,\dots,r}$ of (6.15). There are two points to emphasize. The first one is related to crossing roots and already discussed after Remark 6.6. The second one is that, outside crossing cases, being given any choice for the ordering of the roots (and thus of the vectors of the basis (6.16) for the vector bundle), there is in general no chance to obtain a holomorphicity property for the components of the matrix $K_{0,m-1}(z)$ over $\mathcal{U}$. For example, even the roots of $X^2 - z$ are not holomorphic w.r.t $z \in \mathcal{U}$ because of the logarithm determination. On the other side, any symmetric functions of the roots $(\kappa_j(z))_{j=1,\dots,r}$ however are holomorphic because they can be obtained directly in terms of the coefficients of the poly-



nomial (6.15). So except for crossing roots, the same holds for the quantity $\Delta_{\mathrm{KL}}(z)$ since the matrices B and $\mathcal{B}$ are constants and the determinant itself is a symmetric function.

A very natural way to reach the holomorphic property and go beyond the last difficulties consists in dividing Δ_{KL} by the quantity $\det K_{0,r-1}(z)$. Hence, the same permutation or combination of the vectors of the basis (6.16) is involved in both computations. This *intrinsic Kreiss-Lopatinskii determinant* has already been introduced and studied in a particular case in [BLBS23a].

Definition 6.11 (Intrinsic Kreiss-Lopatinskii determinant)**.** The *intrinsic Kreiss-Lopatinskii determinant* is the complex-valued function defined for $|z| \geqslant 1$ by:

$$\Delta(z) = \frac{\Delta_{\mathrm{KL}}(z)}{\det K_{0,r-1}(z)}. \tag{6.18}$$

Let us note that the intrinsic Kreiss-Lopatinskii determinant can be rewritten

$$\Delta(z) = \frac{\det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z))}{\det K_{0,r-1}(z)} = z^r \det\left(I_r - \frac{\mathcal{B}K_{0,m-1}(z)K_{0,r-1}(z)^{-1}}{z}\right). \tag{6.19}$$

To conclude with these definitions, let us state a little more about the *Uniform Kreiss-Lopatinksii Condition*. With the above notations and additionally to the invertibility of $zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z)$, it corresponds to the existence of a constant $C > 0$ such that for any $z \in \overline{\mathcal{U}}$, any $\widetilde{U} \in \mathcal{E}^s(z)$ solution to (6.17) satisfies the uniform estimate $\|\widetilde{U}\| \leqslant C\|\widetilde{g}\|$. From the Parseval identity for the $\mathcal{Z}$-transform, this inequality gives directly the first necessary half-part of the strong stability estimate (6.9). We refer the reader to [GKO13] for a more detailed presentation.

6.2.3 Main results

Theorem 6.12 is our main theoretical result. It states that the intrinsic Kreiss-Lopatinskii determinant has the same regularity properties as $\mathcal{E}^s(z)$, see Theorem 6.7.

Theorem 6.12 (Smoothness of the intrinsic Kreiss-Lopatinskii determinant)**.** *Assume (H1). The intrinsic Kreiss-Lopatinskii determinant Δ is holomorphic on $\mathcal{U}$ and continuous on $\overline{\mathcal{U}}$.*

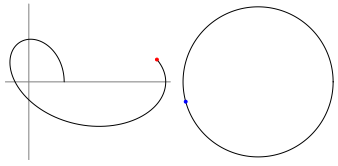
By equation (6.18), the function Δ shares the same zeros with the Kreiss-Lopatinskii determinant Δ_{KL}, so that it can be used as an alternative in the Uniform Kreiss-Lopatinskii Condition, see Theorem 6.2 (Kreiss). Another property, important for the forthcoming applications, lies in the next Corollary 6.14 and involves the following important geometrical object:

Definition 6.13. The *Kreiss-Lopatinskii curve* $\Delta(\mathbb{S})$ is the closed complex parameterized curve

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

Using the residue theorem [1] thanks to Theorem 6.12, we obtain the following result.

1. All the complex analysis results can be found in [Lan99].



Corollary 6.14 (Number of zeros of the intrinsic Kreiss-Lopatinskii determinant)**.** *Assume (H1). If* $0 \notin \Delta(\mathbb{S})$ *then the equation* $\Delta(z) = 0$ *has exactly* $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ *zeros in* $\mathcal{U}$*.*

Here above and in all the paper, $\mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ denotes the winding number of the origin with respect to the closed oriented curve $\Delta(\mathbb{S})$ (see [Lan99] for a definition of the winding number). This corollary helps us to establish an efficient and practical method to study the stability of a given IBVP through Theorem 6.2 (Kreiss). In particular, the low computational cost of the following procedure is very appealing for the study of parameterised IBVP's, see Section 6.4.

Method 6.15 (Uniform Kreiss-Lopatinskii Condition check)**.** *There are two different cases:*

- *if* $0 \in \Delta(\mathbb{S})$*, then there exists* $z_0 \in \mathbb{S}$ *such that* $\Delta(z_0) = 0$*.*
- *if* $0 \notin \Delta(\mathbb{S})$*,* Δ *does not vanish on* $\mathbb{S}$ *and it has* $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ *zeros in* $\mathcal{U}$ *by Corollary 6.14. It follows that if* $\mathrm{Ind}_{\Delta(\mathbb{S})}(0) = r$ *then the scheme is stable. Otherwise the scheme is unstable.*

In summary, by Theorem 6.2 (Kreiss) and since Uniform Kreiss-Lopatinskii Condition is fulfilled if and only if the Kreiss-Lopatinskii determinant has no zero in $\overline{\mathcal{U}}$, Method 6.15 can be used to conclude that the scheme is stable or not. Some illustrations for the O3 scheme and the fifth-order Lax-Wendroff scheme follow in Section 6.4.

6.3 Proof of Theorem 6.12 and Corollary 6.14

6.3.1 Constant-recursive sequence of order r

For each $z \in \mathcal{U}$, we denote P_z the polynomial linked to the characteristic equation (6.15), i.e.

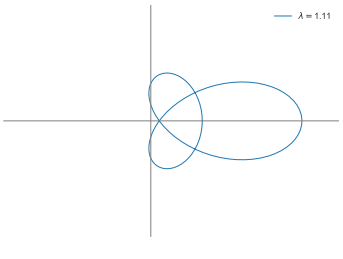
$$P_z(\kappa) = a_p\kappa^{r+p} + \cdots + a_1\kappa^{r+1} + (a_0 - z)\kappa^r + a_{-1}\kappa^{r-1} + \cdots + a_{-r+1}\kappa + a_{-r}. \tag{6.20}$$

By Lemma 6.3 (Hersh), the polynomial $P_z(\kappa)$ can be factorized into two polynomials: one with the r roots in $\mathbb{D}$, denoted $R_z(\kappa)$ and one with the p roots in $\mathcal{U}$, denoted $Q_z(\kappa)$. We know that the coefficients of P_z are holomorphic in z. We already said that the basis (6.16) is not holomorphic because the roots κ are not. In the next result, we prove that the symmetric functions of the r roots κ living in $\mathbb{D}$ are indeed holomorphic in $\mathcal{U}$, in other words, the coefficients of R_z are holomorphic in $\mathcal{U}$.

Lemma 6.16. *For all* $z \in \mathcal{U}$*, the polynomial* $R_z(X) = \prod_{j=1}^{r}(X - \kappa_j(z))$ *has holomorphic coefficients in* $\mathcal{U}$*, where* $(\kappa_j(z))_{j=1}^{r}$ *are the* r *roots (with multiplicity) in* $\mathbb{D}$ *of* (6.15)*.*

Proof. We use the Dunford-Taylor formula with C_z the companion matrix of the polynomial (6.15):

$$\Pi(z) = \frac{1}{2\pi}\int_{\mathbb{S}}(\zeta I_{r+p} - C_z)^{-1}d\zeta$$



It is the projection along $\mathbb{E}_s(z) = \ker \prod_{j=1}^{r}(C_z - \kappa_j(z))$ onto $\mathbb{E}_u(z) = \ker \prod_{j=r+1}^{r+p}(C_z - \kappa_j(z))$ where $(\kappa_j(z))_{j=r+1}^{r+p}$ are the roots of (6.15) in $\mathcal{U}$, because $(\kappa_j(z))_{j=1}^{r}$ are surrounded by $\mathbb{S}$ and $(\kappa_j(z))_{j=r+1}^{r+p}$ are not. The projector $\Pi(z)$ is holomorphic on $\mathcal{U}$ since it is a holomorphic parameter integral. We have $C_z \circ \Pi(z)_{|\mathbb{E}_u(z)} = 0$ and $C_z \circ \Pi(z)_{|\mathbb{E}_s(z)} = C_z$, then the characteristic polynomial of $C_z \circ \Pi(z)$ is $X^p R_z(X)$ because $\mathbb{C}^{r+p} = \mathbb{E}_s(z) \oplus \mathbb{E}_u(z)$. The function $z \mapsto C_z \circ \Pi(z)$ is holomorphic on $\mathcal{U}$, then the coefficients of its characteristic polynomial are too. It concludes the proof. □

6.3.2 Hermite interpolation

To prove the holomorphic properties of Δ, by (6.19), it is sufficient to study the function $z \mapsto K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$. To simplify this study, for $z \in \overline{\mathcal{U}}$, we introduce on the linear map

$$\varphi_z : Q \in \mathbb{C}_{m-1}[X] \mapsto \varphi_z(Q) \in \mathbb{C}_{r-1}[X] \tag{6.21}$$

where $\varphi_z(Q)$ is the Hermite interpolation polynomial of degree less than $r-1$ defined by the value $(Q(\kappa_1(z)), \dots, Q^{(\beta_1-1)}(\kappa_1(z)), Q(\kappa_2(z)), \dots, Q^{(\beta_2-1)}(\kappa_2(z)), \dots, Q^{(\beta_M-1)}(\kappa_M(z)))$, where the κ's are the same as in (6.16). The link between the matrix $K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$ and φ_z is given by:

Lemma 6.17. *For all $z \in \overline{\mathcal{U}}$, the transpose of the matrix $K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$ is the representation in the canonical basis of φ_z defined in (6.21).*

Proof. The Hermite interpolation make appear the following matrix

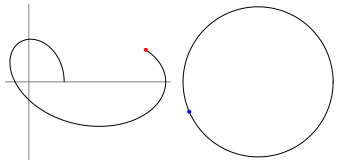
$$H_{0,j}(z) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 1 & 0 & \cdots & \cdots & 1 & 0 & \cdots \\ \kappa_1 & 1 & 0 & \cdots & \kappa_2 & 1 & \cdots & \cdots & \kappa_M & 1 & \cdots \\ \kappa_1^2 & 2\kappa_1 & 2 & \cdots & \kappa_2^2 & 2\kappa_2 & \cdots & \cdots & \kappa_M^2 & 2\kappa_M & \cdots \\ \kappa_1^3 & 3\kappa_1^2 & 6\kappa_1 & \cdots & \kappa_2^3 & 3\kappa_2^2 & \cdots & \cdots & \kappa_M^3 & 3\kappa_M^2 & \cdots \\ \kappa_1^4 & 4\kappa_1^3 & 12\kappa_1^2 & \cdots & \kappa_2^4 & 4\kappa_2^3 & \cdots & \cdots & \kappa_M^4 & 4\kappa_M^3 & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots & \cdots \\ \underbrace{\kappa_1^j}_{\beta_1 \text{ columns linked to } \kappa_1} & \underbrace{j\kappa_1^{j-1}}_{} & \underbrace{j(j-1)\kappa_1^{j-2}}_{} & \cdots & \underbrace{\kappa_2^j}_{\beta_2 \text{ columns linked to } \kappa_2} & \underbrace{j\kappa_2^{j-1}}_{} & \cdots & \cdots & \underbrace{\kappa_M^j}_{\beta_M \text{ columns linked to } \kappa_M} & \underbrace{j\kappa_M^{j-1}}_{} & \cdots \end{pmatrix}.$$

The representation of $z \mapsto \varphi_z$ in the canonical basis is $(H_{0,m-1}(z)H_{0,r-1}^{-1}(z))^\top$. Besides, there exists an invertible matrix $M(z) \in \mathcal{M}_r(\mathbb{C})$ such that $K_{0,j}(z) = H_{0,j}(z)M(z)$. Therefore, we have $K_{0,m-1}(z)K_{0,r-1}^{-1}(z) = H_{0,m-1}(z)M(z)M(z)^{-1}H_{0,r-1}^{-1}(z) = H_{0,m-1}(z)H_{0,r-1}^{-1}(z)$. The result follows. □

Proposition 6.18. *The function $z \mapsto \varphi_z$ is holomorphic on $\mathcal{U}$.*

Proof. For all $k \in [\![0 : m-1]\!]$, we want every coefficient of the polynomial $\varphi_z(X^k)$ to be holomorphic on $\mathcal{U}$. Writing $\varphi_z(X^k)(x) = \sum_{j=0}^{r-1} \alpha_{j,k}(z)x^j$, we know that

$$\forall j \in [\![0 : r-1]\!], \quad j!\alpha_{j,k}(z) = \partial_x^j \varphi_z(X^k)(x)_{|x=0}. \tag{6.22}$$



By the error of Hermite interpolation (see [Her77]), we have

$$\varphi_z(X^k)(x) - x^k = \frac{1}{2i\pi}\int_{\mathbb{S}} \frac{\zeta^k R_z(x)}{(x-\zeta)R_z(\zeta)} d\zeta \tag{6.23}$$

where $R_z(X)$ is defined in Lemma 6.16. Differentiating equation (6.23) (with the Leibniz product rule), one obtains

$$j!\alpha_{j,k}(z) = k!\delta_k^j + \sum_{s=0}^{j} \binom{j}{s} R_z^{(j-s)}(0) \frac{1}{2i\pi}\int_{\mathbb{S}} \frac{-s!\ \zeta^{k-s-1}}{R_z(\zeta)} d\zeta. \tag{6.24}$$

By Lemma 6.16 and the holomorphicity of parameter-dependent integrals, the function $z \mapsto \alpha_{j,k}(z)$ is holomorphic on $\mathcal{U}$ for all $j \in [\![0 : r-1]\!]$ and $k \in [\![0 : m-1]\!]$. The proof is now complete. □

Proposition 6.19. *The function $z \mapsto \varphi_z$ is continuous on $\overline{\mathcal{U}}$.*

Proof. Because Lemma 6.3 (Hersh) does not hold anymore for $z \in \mathbb{S}$, the roots $\kappa(z)$ of characteristic equation (6.15) can be on the unit circle $\mathbb{S}$. To prove the continuity of $z \mapsto \alpha_{j,k}(z)$, we use equation (6.24) but replacing $\mathbb{S}$ by $\mathbb{S}_\varepsilon \overset{\text{def}}{=} \{z \in \mathbb{C}, |z| = 1+\varepsilon\}$ for $\varepsilon > 0$. Using the continuity of parameter-dependent integrals and the continuity of the roots $\kappa(z)$ of characteristic equation (6.15), we obtain the continuity of the coefficients of R_z and thus the function $z \mapsto \alpha_{j,k}(z)$ is continuous on $\overline{\mathcal{U}}$ for all $j \in [\![0 : r-1]\!]$ and $k \in [\![0 : m-1]\!]$. The proof is now complete. □

Proposition 6.20. *The function $z \mapsto \varphi_z$ is bounded on $\overline{\mathcal{U}}$.*

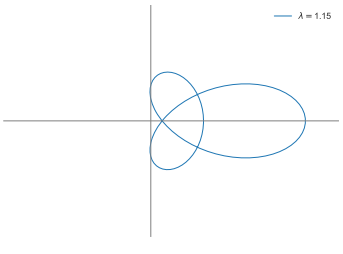
Proof. Equation (6.24) can give a bound of every components of $K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$. Indeed, using Rouché's theorem, as in the proof of Lemma 6.3 (Hersh), we can see that for $|z| > R$ for a certain R, all the roots $\kappa(z)$ of the characteristic equation (6.15) satisfy $|\kappa(z)| < \frac{1}{2}$. Then, for $|z| > R$, one can have

$$\left|\frac{1}{2i\pi}\int_{\mathbb{S}} \frac{-s!\ \zeta^{k-s-1}}{R_z(\zeta)} d\zeta\right| = \left|\frac{s!}{2i\pi}\int_0^{2\pi} \frac{e^{i\theta(k-s-1)}}{\prod_{j=1}^r (e^{i\theta} - \kappa_j(z))} d\theta\right| \leqslant \frac{s!}{2\pi}\int_0^{2\pi} \frac{1}{|1-\frac{1}{2}|^r} d\theta \leqslant s!2^r.$$

By Gauss-Lucas theorem, the roots of all the derivatives of R_z are in $\mathbb{D}$ for all $|z| > 1$, it follows that $R_z^{(j-s)}(0)$ is bounded independently of z. Then for $|z| > R$, the quantity $K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$ is bounded. Moreover, by Proposition 6.19, the quantity $K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$ is bounded on the compact set $\{1 \leqslant |z| \leqslant R\}$. The proof is now complete. □

6.3.3 Conclusion

Proof of Theorem 6.12. By Lemma 6.17, the continuity and holomorphicity properties of $z \mapsto \varphi_z$ provided in Propositions 6.18 and 6.19 are shared by the function $z \mapsto K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$. The expression of the intrinsic determinant (6.19) concludes the proof. □



The next proof is close to the proof of the Corollary 15 of [BLBS23a]. We reproduce it here for completeness.

Proof of Corollary 6.14. Let us define the following function

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

By Theorem 6.12, the function $\widetilde{\Delta}$ is meromorphic on $\mathbb{D}$ with one only pole in 0 and is continuous on $\overline{\mathbb{D}} \setminus \{0\}$. By Proposition 6.20, the function $z \mapsto K_{0,m-1}(1/z)K_{0,r-1}(1/z)^{-1}$ is bounded on $\overline{\mathbb{D}}$, it follows that 0 is a pole of order r of the function $\widetilde{\Delta}$. The residue theorem applied on $\widetilde{\Delta}$ with the path $\mathbb{S}$ gives the following equality:

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

It follows that

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

□

6.4 Numerical results

The results presented below can be reproduced using the code available in the GitHub repository [LB23], it can be used as a Python library, following the link : `https://doi.org/10.5281/zenodo.7773742`.

6.4.1 New formulation of Δ

In [BLBS23a], an explicit formula of the Kreiss-Lopatinskii determinant is given. Unfortunately to reduce the boundary matrix $\mathcal{B}$, the characteristic equation of degree $r+p$ is used to find a final matrix of size $r \times (r+p)$ and, since in the present analysis $p \neq 0$, the matrix is not square. To skirt that problem we will use the polynomial R_z defined in Lemma 6.16 instead of using the complete characteristic equation (6.20). It reads also

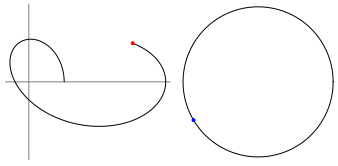
$$R_z(X) = \prod_{j=1}^{r}(X - \kappa_j(z)) = X^r + \sigma_{r-1}(z)X^{r-1} + \cdots + \sigma_1(z)X + \sigma_0(z)$$

where $(\sigma_j(z))_j$ are the symmetric functions of $(\kappa_j(z))_j$.

Because $\widetilde{U}_j(z)$ is in $\mathcal{E}^s(z)$ and can be expressed in the basis (6.16), we have, for all $j \in \mathbb{N}$,

$$\widetilde{U}_{j+r}(z) + \sigma_{r-1}(z)\widetilde{U}_{j+r-1}(z) + \cdots + \sigma_1(z)\widetilde{U}_{j+1}(z) + \sigma_0(z)\widetilde{U}_j(z) = 0. \tag{6.25}$$

Notation. We note, for all $j \in \mathbb{N}$, $\widetilde{\mathfrak{U}}_j(z)$ the vector $(\widetilde{U}_0(z) \cdots \widetilde{U}_j(z))^\top$ of size $j+1$.



Proposition 6.21. *Let $\mathcal{B} \in \mathcal{M}_{r,m}(\mathbb{C})$. There exists a function $\widetilde{\mathcal{B}} : \mathbb{C}^r \to \mathcal{M}_{r,r}(\mathbb{C})$ constructible such that, for all $z \in \overline{\mathcal{U}}$, we have*

$$\mathcal{B}\widetilde{\mathfrak{U}}_{m-1}(z) = \widetilde{\mathcal{B}}(\sigma_0(z), \dots, \sigma_{r-1}(z))\widetilde{\mathfrak{U}}_{r-1}(z) \tag{6.26}$$

where $(\widetilde{U}_j(z))_j$ satisfies (6.25) for all $j \in \mathbb{N}$.

By "constructible function", we mean here that we establish a computable algorithm to get the matrix $\widetilde{\mathcal{B}}(\sigma_0(z), \dots, \sigma_{r-1}(z))$. This algorithm, based on a Gaussian elimination, is fully described in the following proof.

Proof. For $z \in \overline{\mathcal{U}}$ and $\varsigma_0 = \sigma_0(z), \dots, \varsigma_{r-1} = \sigma_{r-1}(z)$. By a descending induction on j between $m-1$ to $r-1$, we construct a matrix $\mathcal{B}_j(\varsigma_0, \dots, \varsigma_{r-1}) \in \mathcal{M}_{r,j+1}(\mathbb{C})$ such that

$$\mathcal{B}\widetilde{\mathfrak{U}}_{m-1}(z) = \mathcal{B}_j(\varsigma_0, \dots, \varsigma_{r-1})\widetilde{\mathfrak{U}}_j(z).$$

Initialization: if $j = m-1$ then one can take $\mathcal{B}$ for the matrix $\mathcal{B}_{m-1}(\varsigma_0, \dots, \varsigma_{r-1})$.
Induction: we assume the induction hypotheses for some $j \in [\![m-1 : r]\!]$ and we want to prove the result for $j-1$. By equation (6.25), we have $\widetilde{\mathfrak{U}}_j(z) = \mathcal{P}_j\widetilde{\mathfrak{U}}_{j-1}(z)$ where

$$\mathcal{P}_j = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & & 1 \\ (0) & -\varsigma_0 & \cdots & -\varsigma_{r-1} \end{pmatrix} \in \mathcal{M}_{j+1,j}(\mathbb{C}).$$

We define $\mathcal{B}_{j-1}(\varsigma_0, \dots, \varsigma_{r-1}) = \mathcal{B}_j(\varsigma_0, \dots, \varsigma_{r-1})\mathcal{P}_j \in \mathcal{M}_{r,j}(\mathbb{C})$ then we have

$$\mathcal{B}_{j-1}\widetilde{\mathfrak{U}}_{j-1}(z) = \mathcal{B}_j\mathcal{P}_j\widetilde{\mathfrak{U}}_{j-1}(z) = \mathcal{B}_j\widetilde{\mathfrak{U}}_j(z) = \mathcal{B}\widetilde{\mathfrak{U}}_{m-1}(z).$$

Conclusion: we define $\widetilde{\mathcal{B}}$ by $\mathcal{B}_{r-1}$.
The function $\widetilde{\mathcal{B}}$ is easily computable because $(\mathcal{P}_j)_j$ are just matrices of Gaussian elimination. □

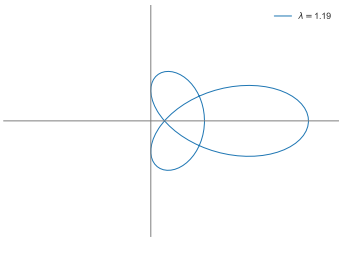
By (6.26), the intrinsic Kreiss-Lopatinskii determinant can be written

$$\Delta(z) = \frac{\det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z))}{\det K_{0,r-1}(z)} = \frac{\det(zK_{0,r-1}(z) - \widetilde{\mathcal{B}}K_{0,r-1}(z))}{\det K_{0,r-1}(z)} = \det(zI_r - \widetilde{\mathcal{B}}). \tag{6.27}$$

The matrix $\widetilde{\mathcal{B}}$ from Proposition 6.21 depends on coefficients $(\sigma_j(z))_j$. By (6.26), we have

$$\widetilde{\mathcal{B}}(\sigma_0(z), \dots, \sigma_{r-1}(z)) = \mathcal{B}K_{0,m-1}(z)K_{0,r-1}^{-1}(z).$$

Using any computer algebra system, we can compute the matrix $\widetilde{\mathcal{B}}(\varsigma_0, \dots, \varsigma_{r-1})$ from the ma-



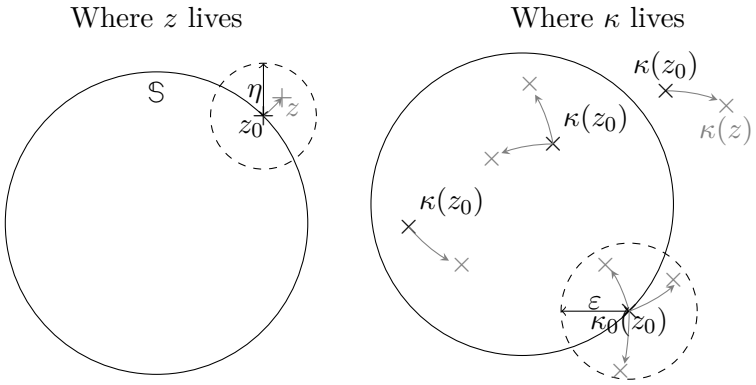


Figure 6.2 – Illustration of Lemma 6.22

trix $\mathcal{B}$ and then we compute the coefficients $(\sigma_j(z))_j$ and replace ς_j by $\sigma_j(z)$ for all $j \in [\![0 : r-1]\!]$. It provides that the computation of $\widetilde{\mathcal{B}}(\varsigma_0, \dots, \varsigma_{r-1})$ can be done only once and then apply for different z (and so different $(\sigma_j(z))_j$).

With (6.27), we find again the holomorphic property of Δ by Lemma 6.16 which states that $z \mapsto \sigma_j(z)$ is holomorphic on $\mathcal{U}$ for all $j \in [\![0 : r-1]\!]$.

6.4.2 Computation of $\Delta(\mathbb{S})$

Let us fix a $z_0 \in \mathbb{S}$. To compute $(\sigma_j(z_0))_j$, we need the r roots $(\kappa_j(z_0))_j$ that come from the inside of the unit disk, see Remark 6.8. By the continuity of the roots of polynomial P_{z_0} defined in (6.20) with respect to the parameter z_0, for each $\kappa_0(z_0)$ of multiplicity β on the unit circle, for a sufficiently small $\varepsilon > 0$, there exists $\eta > 0$ such that for all $z \in B(z_0, \eta)$, the polynomial P_z has exactly β roots with multiplicity in $B(\kappa_0(z_0), \varepsilon)$. The explicit value of η is given in the following statement.

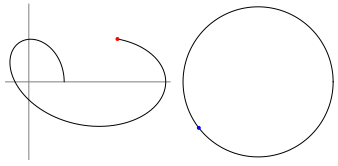
Lemma 6.22. *Let z_0 be on the unit circle. Let $\kappa_0(z_0) \in \mathbb{S}$ be a root of multiplicity β of the polynomial P_{z_0} defined in (6.20). Let $\varepsilon > 0$ be such that $\kappa_0(z_0)$ is the only root of P_{z_0} in $\overline{B(\kappa_0(z_0), \varepsilon)}$ and set*

$$\eta = (1+\varepsilon)^{-r} \min_{\kappa \in \partial B(\kappa_0(z_0), \varepsilon)} |P_{z_0}(\kappa)|.$$

Then for all $z \in B(z_0, \eta)$, the polynomial P_z has exactly β roots with multiplicity in $B(\kappa_0(z_0), \varepsilon)$.

The proof of Lemma 6.22 is a consequence of Rouché's theorem, comparing the number of zeros between P_{z_0} and P_z for z close to z_0 in $B(\kappa_0(z_0), \varepsilon)$, the details of the proof are not given here, but we refer to [Lan99] for a proof of Rouché's theorem. This lemma is illustrated in Figure 6.2 where $\kappa_0(z_0)$ is of multiplicity 3. The black points are related to z_0 and $(\kappa_j(z_0))_j$, and the gray point are related to z and $(\kappa_j(z))_j$.

From the numerical point of view, for a multiple root $\kappa_0(z_0)$, one can take the smallest distance between two roots of P_{z_0} as ε, take $z = (1 + \frac{\eta}{2})z_0 \in \mathcal{U}$. The value of η is obtained discretizing the circle of radius ε centered in $\kappa_0(z_0)$. By Lemma 6.3 (Hersh), there is no roots of P_z



on the unit circle, then one can count the roots in $B(\kappa_0(z_0),\varepsilon)\cap\mathbb{D}$ and $B(\kappa_0(z_0),\varepsilon)\cap\mathcal{U}$ to know the number of roots linked to $\kappa_0(z)$ that come from the inside and the outside of the unit disk. After selecting the roots $(\kappa_j(z))_j$ that come from the inside of the unit disk, one may compute their symmetric functions $(\sigma_j(z))_j$. By replacing the formal variables $(\varsigma_j)_j$ of $\widetilde{\mathcal{B}}(\varsigma_0,\dots,\varsigma_{r-1})$ with $(\sigma_j(z))_j$, one may compute $\Delta(z)$ with expression (6.27). Instead of computing $(\kappa_j(z))_j$ for each z on the unit circle independently, one may use the continuity of $(\kappa_j(z))_j$ with respect to z in order to describe the movement of the roots $(\kappa_j(z))_j$ for $z\in\mathbb{S}$. After drawing the Kreiss-Lopatinskii curve, the winding number has to be computed in order to use Method 6.15. To do so, we use the geometric algorithm proposed by García Zapata and Díaz Martín in [GZDM12] and [GZDM14].

6.4.3 Boundary condition: reconstruction procedure

To define the boundary condition, we use the reconstruction procedure explained in [DDJ18]. The framework is the advection equation with a misalignment between the space boundary and the discrete grid points

$$\begin{cases} \partial_t u + a\partial_x u = 0, & t\geqslant 0, x\in[x_\sigma,1], \\ u(t,x_\sigma) = g(t), & t\geqslant 0, \\ u(0,x) = f(x), & x\in[x_\sigma,1]. \end{cases} \tag{6.28}$$

Without loss of generality, we can assume that $x_\sigma=\sigma\Delta x$ with $\sigma\in[-\frac{1}{2},\frac{1}{2}[$ (as it is explained in [BLBS23a]). Let us introduce x_j for $j\Delta x$ and t^n for $n\Delta t$ when $j\geqslant -r$ and $n\geqslant 0$. Let $n\in\mathbb{N}$ be a fixed time. The solution u of (6.28) (assumed here to be smooth enough) satisfies

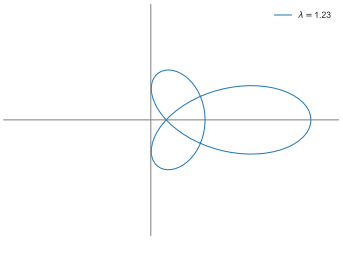
$$\frac{1}{\Delta x}\int_{x_j-\frac{\Delta x}{2}}^{x_j+\frac{\Delta x}{2}} u(t^n,y)dy = \frac{1}{\Delta x}\int_{x_j-\frac{\Delta x}{2}}^{x_j+\frac{\Delta x}{2}} \sum_{k=0}^{d-1}\partial_x^k u(t^n,x_\sigma)\frac{(y-x_\sigma)^k}{k!}dy + O(\Delta x^d) \tag{6.29}$$

using a Taylor expansion of order d. Now, let us take a solution $(U_j^n)_{j\geqslant 0}$ of a scheme of the form (6.2) approximating u. Using (6.29), we want to define the r ghost points $(U_j^n)_{-r\leqslant j\leqslant -1}$. The approximation of equation (6.29) reads, for $j\geqslant -r$,

$$U_j^n \approx \sum_{k=0}^{d-1}\partial_x^k u(t^n,x_\sigma)\left(\frac{(j+\frac{1}{2}-\sigma)^{k+1}}{(k+1)!}-\frac{(j-\frac{1}{2}-\sigma)^{k+1}}{(k+1)!}\right). \tag{6.30}$$

On the one hand, we use the PDE to convert space derivatives into time derivatives until an index $k_d<d$. The index k_d allows us to know only the first derivatives of the boundary datum g and use extrapolation for the rest. For the advection equation, we have, for all $k\leqslant k_d$,

$$\partial_x^k u(t^n,x_\sigma) = (-a)^{-k}\partial_t^k u(t^n,x_\sigma) = (-a)^{-k}g^{(k)}(t^n).$$



Equation (6.30) becomes, for $j \geqslant -r$,

$$U_j^n \approx \sum_{k=0}^{k_d} (-a)^{-k} g^{(k)}(t^n) \left(\frac{(j+\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} - \frac{(j-\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} \right) + \sum_{k=k_d+1}^{d-1} \partial_x^k u(t^n, x_\sigma) \left(\frac{(j+\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} - \frac{(j-\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} \right). \quad (6.31)$$

On the other hand, we need to define $(\partial_x^k u(t^n, x_\sigma))_{k=k_d+1}^{d-1}$, but using (6.31) for $j \in [\![0 : d-k_d-2]\!]$, we can deduce the unknowns $(U_j^n)_{-r \leqslant j \leqslant -1}$. Writing $\mathfrak{U}_- = (U_{-r}^n, \dots, U_{-1}^n)^\top$, $\mathfrak{U}_+ = (U_0^n, \dots, U_{d-k_d-2}^n)^\top$ and $\Theta^n = (\partial_x^{k_d+1} u(t^n, x_\sigma), \dots, \partial_x^{d-1} u(t^n, x_\sigma))^\top$, we have a condensed formulation of (6.31):

$$\begin{cases} \mathfrak{U}_- &= \mathcal{S}_-^n + \mathcal{Y}_- \Theta^n, \\ \mathfrak{U}_+ &= \mathcal{S}_+^n + \mathcal{Y}_+ \Theta^n, \end{cases} \quad (6.32)$$

where $\mathcal{S}_-^n \in \mathbb{R}^r$, $\mathcal{S}_+^n \in \mathbb{R}^{d-k_d-1}$, $\mathcal{Y}_- \in \mathcal{M}_{r,d-k_d-1}(\mathbb{R})$ and $\mathcal{Y}_+ \in \mathcal{M}_{d-k_d-1}(\mathbb{R})$ with

$$\begin{cases} (\mathcal{S}_-^n)_i = \displaystyle\sum_{k=0}^{k_d} (-a)^{-k} g^{(k)}(t^n) \left(\frac{(-i+\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} - \frac{(-i-\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} \right) & \text{for } i \in [\![1 : r]\!] \\ (\mathcal{S}_+^n)_i = \displaystyle\sum_{k=0}^{k_d} (-a)^{-k} g^{(k)}(t^n) \left(\frac{(i-1+\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} - \frac{(i-1-\frac{1}{2}-\sigma)^{k+1}}{(k+1)!} \right) & \text{for } i \in [\![1 : d-k_d-1]\!] \\ (\mathcal{Y}_-)_{i,j} = \left(\dfrac{\left(-(r-i+1)+\frac{1}{2}-\sigma\right)^{j+k_d+1}}{(j+k_d+1)!} - \dfrac{\left(-(r-i+1)-\frac{1}{2}-\sigma\right)^{j+k_d+1}}{(j+k_d+1)!} \right) & \text{for } i \in [\![1 : r]\!], \\ & \quad j \in [\![1 : d-k_d-1]\!] \\ (\mathcal{Y}_+)_{i,j} = \left(\dfrac{(i-1+\frac{1}{2}-\sigma)^{j+k_d+1}}{(j+k_d+1)!} - \dfrac{(i-1-\frac{1}{2}-\sigma)^{j+k_d+1}}{(j+k_d+1)!} \right) & \text{for } i,j \in [\![1 : d-k_d-1]\!] \end{cases}$$

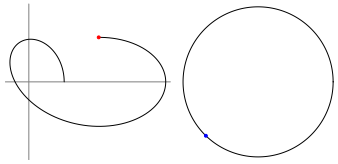
Eliminating the space derivatives Θ^n in (6.32) gives the following boundary condition:

$$\mathfrak{U}_- = \mathcal{Y}_- \mathcal{Y}_+^{-1} \mathfrak{U}_+ + \mathcal{S}_-^n - \mathcal{Y}_- \mathcal{Y}_+^{-1} \mathcal{S}_+^n. \quad (6.33)$$

Equation (6.33) is exactly the boundary equations (6.5) of the scheme which define the r ghost points of the scheme. To write the boundary condition as equation (6.7) with expression (6.8), we identify

$$B \overset{def}{=} \mathcal{Y}_- \mathcal{Y}_+^{-1} \text{ and } \begin{pmatrix} g_{-r}^n \\ \vdots \\ g_{-1}^n \end{pmatrix} \overset{def}{=} \mathcal{S}_-^n - \mathcal{Y}_- \mathcal{Y}_+^{-1} \mathcal{S}_+^n.$$

As in [DDJ18], $\mathcal{R}^{d,k_d}$ denotes the reconstruction procedure where d is the order of consistency of the method and k_d the index when we change from time derivatives to extrapolation. For



example, the reconstruction procedure $\mathcal{R}^{3,0}$ for $r = 2$ and $\sigma = 0.4$ leads to

$$\mathcal{Y}_- = \begin{pmatrix} -\frac{12}{5} & \frac{1753}{600} \\ -\frac{7}{5} & \frac{613}{600} \end{pmatrix}, \mathcal{Y}_+ = \begin{pmatrix} -\frac{2}{5} & \frac{73}{600} \\ \frac{3}{5} & \frac{133}{600} \end{pmatrix} \text{ and } B = \begin{pmatrix} \frac{1371}{97} & \frac{526}{97} \\ \frac{554}{97} & \frac{143}{97} \end{pmatrix}$$

6.4.4 Example of O3 scheme

As it is done in [DDJ18], we want to find the stability area for the O3 scheme defined, for $j \in \mathbb{N}$ and $n \in \mathbb{N}$, by

$$U_j^{n+1} = \left(\frac{\lambda^3}{6} - \frac{\lambda}{6}\right) U_{j-2}^n + \left(\lambda + \frac{\lambda^2}{2} - \frac{\lambda^3}{2}\right) U_{j-1}^n + \left(1 - \frac{\lambda}{2} - \lambda^2 + \frac{\lambda^3}{2}\right) U_j^n + \left(\frac{\lambda^2}{2} - \frac{\lambda^3}{6} - \frac{\lambda}{3}\right) U_{j+1}^n \tag{6.34}$$

The O3 scheme is a scheme with $r = 2$ and $p = 1$ and is Cauchy-stable for $\lambda \in]0, 1]$. The reconstruction $\mathcal{R}^{3,0}$ for the O3 scheme and $\sigma = 0.4$ leads to

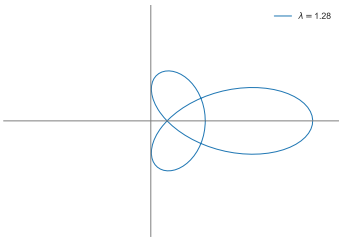
$$B = \begin{pmatrix} \frac{1371}{97} & \frac{526}{97} & 0 \\ \frac{554}{97} & \frac{143}{97} & 0 \end{pmatrix} \text{ and } \mathcal{B} = \begin{pmatrix} \frac{180\lambda^2}{97} + \frac{277\lambda}{97} + 1 & \frac{120\lambda^2}{97} + \frac{23\lambda}{97} & 0 \\ \frac{263\lambda^3}{582} + \frac{\lambda^2}{2} + \frac{14\lambda}{291} & \frac{217\lambda^3}{291} - \lambda^2 - \frac{217\lambda}{291} + 1 & -\frac{\lambda^3}{6} + \frac{\lambda^2}{2} - \frac{\lambda}{3} \end{pmatrix}.$$

Using the reformulation (6.27) of the Kreiss-Lopatinskii determinant, we have

$$\begin{aligned} \widetilde{\mathcal{B}}(\varsigma_0, \varsigma_1) &= \mathcal{B} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -\varsigma_0 & -\varsigma_1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{180\lambda^2}{97} + \frac{277\lambda}{97} + 1 & \frac{120\lambda^2}{97} + \frac{23\lambda}{97} \\ \frac{(263+97\varsigma_0)\lambda^3}{582} + \frac{(1-\varsigma_0)\lambda^2}{2} + \frac{(14+97\varsigma_0)\lambda}{291} & \frac{(434+97\varsigma_1)\lambda^3}{582} - \frac{(2+\varsigma_1)\lambda^2}{2} - \frac{(217-97\varsigma_1)\lambda}{291} + 1 \end{pmatrix}. \end{aligned}$$

For example, for $\sigma = 0.4$, Figure 6.3 shows that the O3 scheme with $\mathcal{R}^{3,0}$ boundary is stable for $\lambda = 0.4$ (because $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0) = 0$) and is unstable for $\lambda = 0.9$ (because $r - \mathrm{Ind}_{\Delta(\mathbb{S})}(0) = 1$).

We can draw the same figure as the Figure 4 of [DDJ18] but instead of using a computation of the spectral radius of the truncated quasi-Toeplitz matrix, we use our strategy of counting the number of instability modes, see Figure 6.4 which is much more reliable (since it is parameter free) and efficient. In Figure 6.4, every area stamped with 0 is a domain where the O3 scheme is stable. The odd pattern for very small λ (approximatively between 0 and 0.01) of Figure 6.4 may be due to difficulties for computing the winding number. Indeed, for very small values of λ, the Kreiss-Lopatinskii determinant is really close to the origin and even with a refinement (see the next example for more details on this procedure), the computation of the winding number may become inaccurate, which is not a problem in practice since it would correspond to very small, then unusable, time steps.



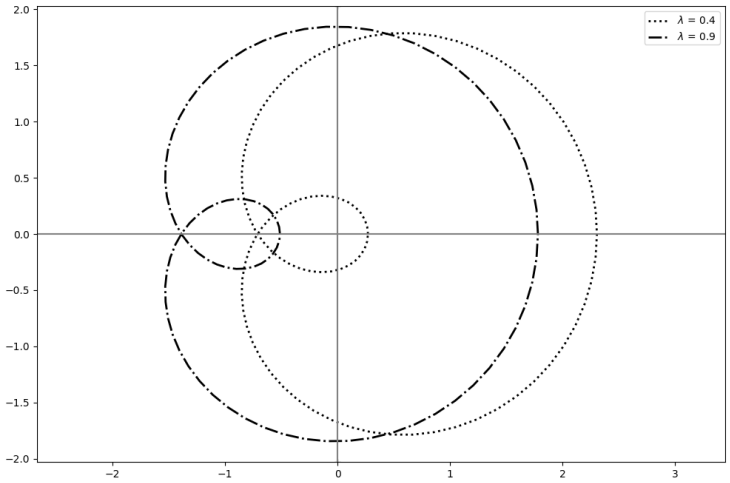


Figure 6.3 – Curve $\Delta(\mathbb{S})$ for O3 scheme for $\sigma = 0.4$, for $\lambda \in \{0.4, 0.9\}$ with reconstruction boundary $\mathcal{R}^{3,0}$.

6.4.5 Example of Lax-Wendroff 5

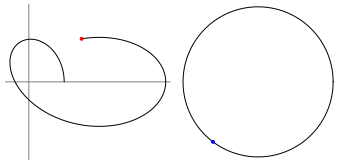
The fifth-order Lax-Wendroff scheme, called LW5, which has been proposed in [LM06], can be written, for all $n \in \mathbb{N}$, for all $j \in \mathbb{N}$, as

$$U_j^{n+1} = \frac{\lambda(\lambda-2)(\lambda-1)(\lambda+1)(\lambda+2)}{120} U_{j-3}^n - \frac{\lambda(\lambda-1)(\lambda-3)(\lambda+1)(\lambda+2)}{24} U_{j-2}^n + \frac{\lambda(\lambda-2)(\lambda-3)(\lambda+1)(\lambda+2)}{12} U_{j-1}^n + \left(1 - \frac{\lambda(\lambda^4-3\lambda^3-5\lambda^2+15\lambda+4)}{12}\right) U_j^n + \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)(\lambda+2)}{24} U_{j+1}^n - \frac{\lambda(\lambda-1)(\lambda-2)(\lambda-3)(\lambda+1)}{120} U_{j+2}^n. \tag{6.35}$$

This scheme LW5 is Cauchy-stable for $\lambda \in]0, 1]$.

Figure 6.5 illustrates the computation of the number of instabilities for LW5 for different reconstruction boundaries where $\sigma = 0.4$ with respect to $\lambda \in]0, 1]$. As in the previous example, it may happen that the Kreiss-Lopatinskii curve is too close to the origin, the winding number of the origin cannot be computed correctly. Following the geometric algorithm proposed by García Zapata in [GZDM14], a refinement of the discretization then improves the effective computation of the winding number. Figure 6.6 represents such refinement with close-up close to the origin. However, even with this strategy, for very small values of λ, we cannot refine more than the machine precision, that is why there is still some odd pattern for very small λ in Figure 6.4, Figure 6.5 and Figure 6.7. As we already discussed, such very small time step are however not used in practice. The stability area with respect to both parameters λ and σ are drawn in Figure 6.7, considering again successively various reconstruction boundary conditions.

All the figures can be easily computed in Python with the common NumPy [HMvdW+20] library and the SymPy [MSP+17] library for the computer algebra system. The algorithm is really efficient. For each subfigure of Figure 6.4, the 1600 runs takes less than a couple of minutes of computation achieved on a standard laptop. Moreover, our procedure provides sharp results. In



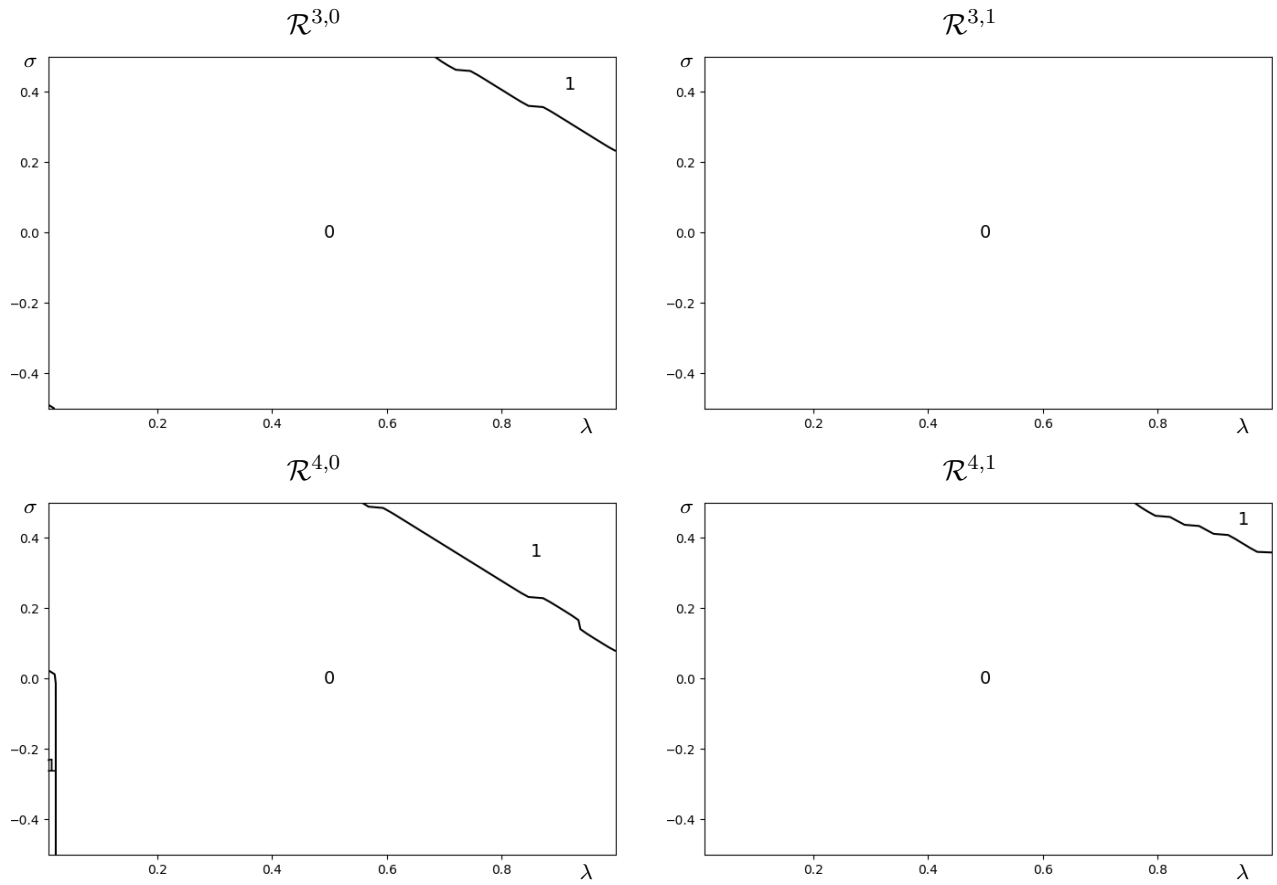


Figure 6.4 – Number of zeros of the Kreiss-Lopatinskii determinant of O3 scheme with different reconstruction boundaries for $\lambda \in]0, 1]$ and $\sigma \in]-0.5, 0.5[$.

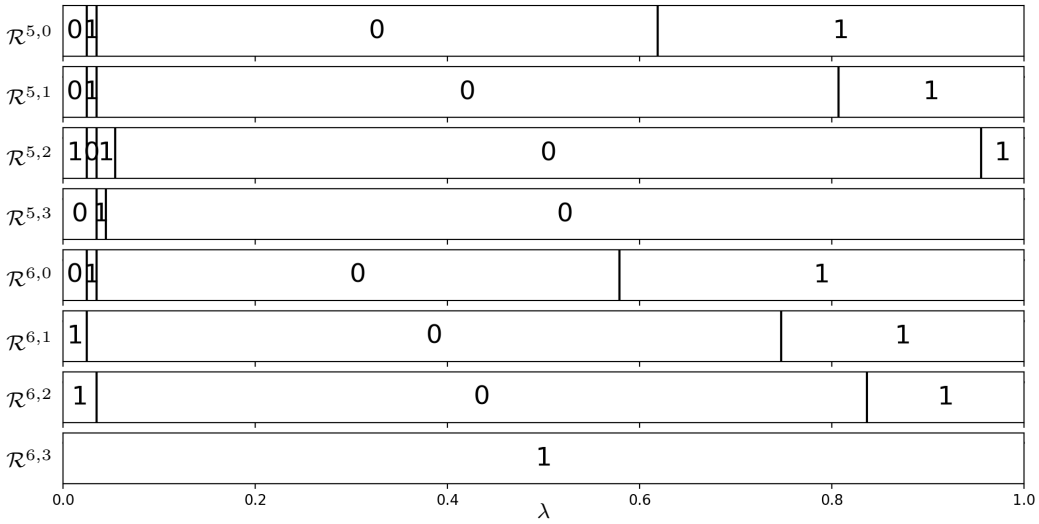
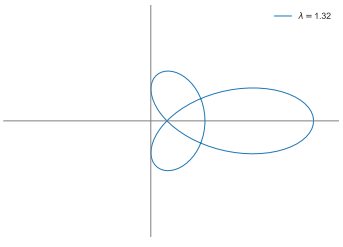


Figure 6.5 – Number of zeros of the Kreiss-Lopatinskii determinant of LW5 scheme with different reconstruction boundaries for $\lambda \in]0, 1]$ and $\sigma = 0.4$.



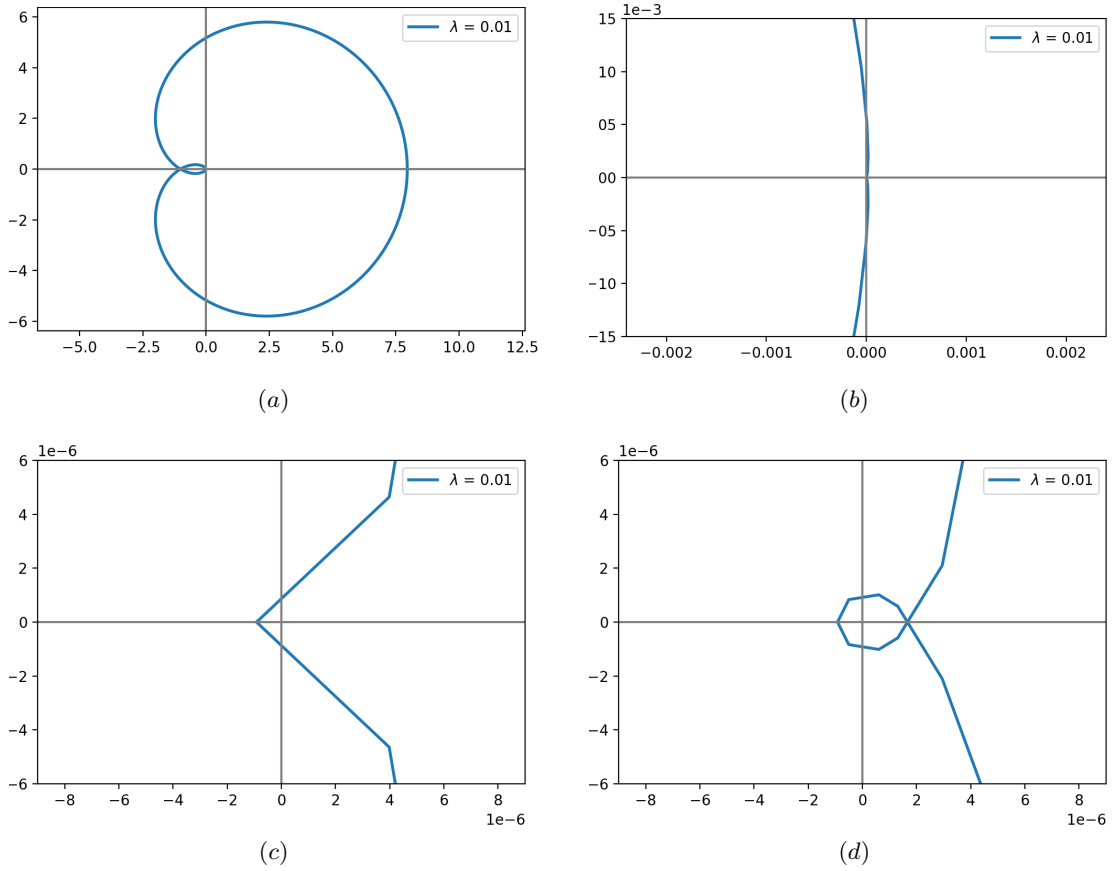
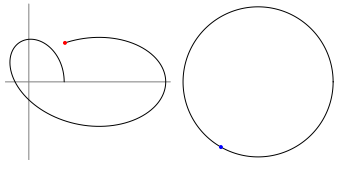


Figure 6.6 – (a) representation of $\Delta(\mathbb{S})$ for LW5 with the boundary condition $\mathcal{R}^{6,1}$, for $\lambda = 0.01$ and $\sigma = 0$, zoom (b) without refinement (c) and with refinement (d).



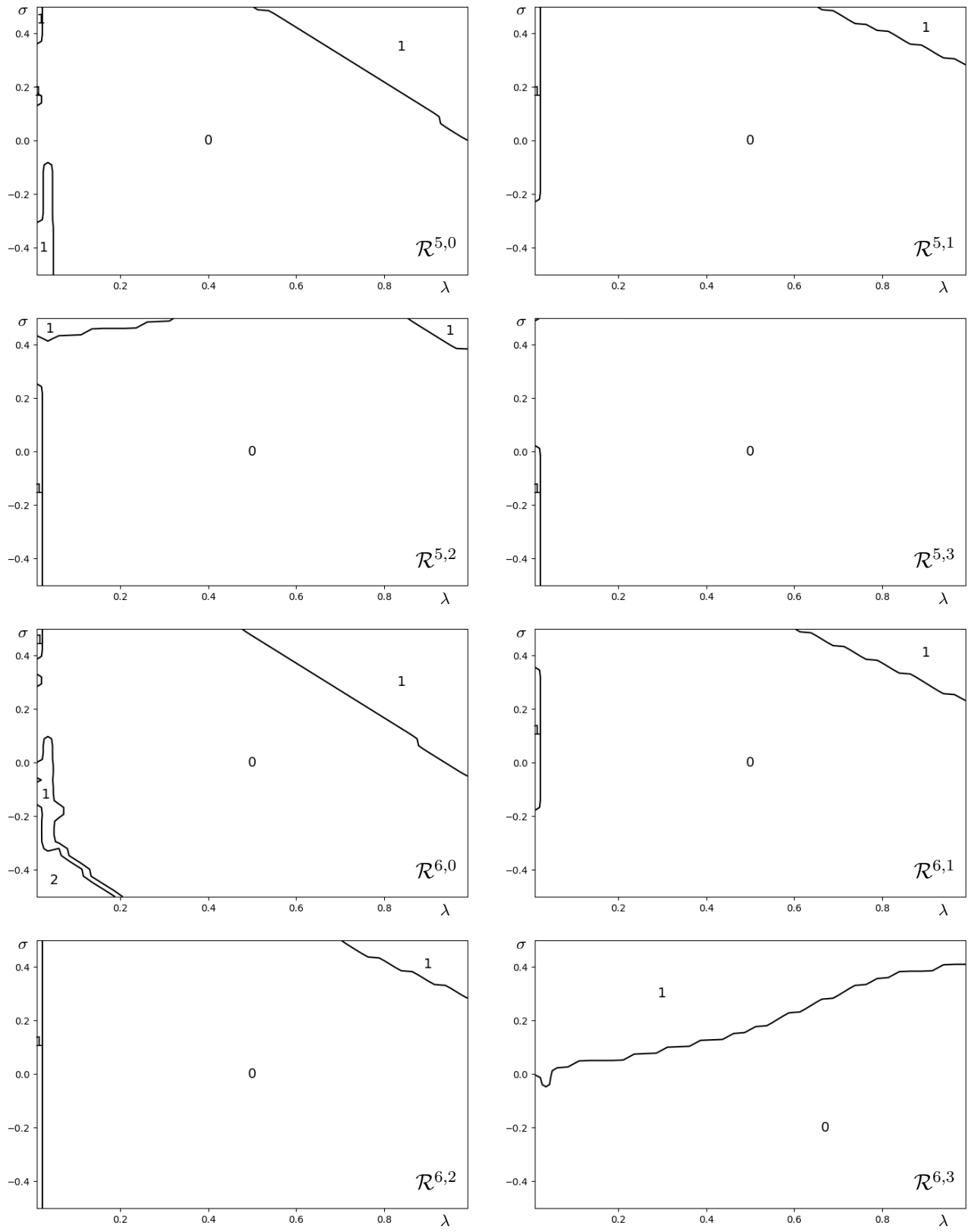
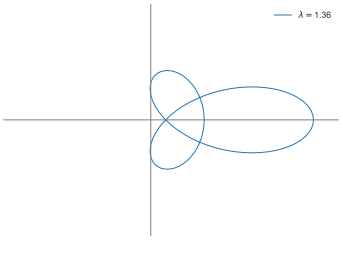


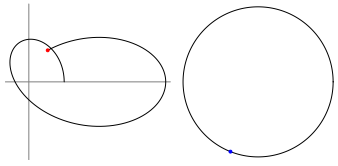
Figure 6.7 – Number of zeros of the Kreiss-Lopatinskii determinant of LW5 scheme with different reconstruction boundaries for $\lambda \in]0, 1]$ and $\sigma \in]-0.5, 0.5[$.



particular, contrary to numerical investigations of stability which are based on the computation of the spectral radius, no arbitrary truncation of (quasi-)Toeplitz matrices is needed.

6.5 Future directions

The present theoretical and numerical results is restricted to the class of one-step schemes. Actually, the multistep case presents similarities with the one-time step case thanks to the work of Coulombel [Cou13]. Hence, Theorem 6.12 still holds and Corollary 6.14 has to be adapted: the number of zeros in $\mathcal{U}$ of the equation $\Delta(z) = 0$ is now $r(s+1) - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$ where $s+1$ is the number of time-steps of the scheme, this work is initiated in Section 6.7. In another direction, for implicit schemes or for more general boundary conditions, such as absorbing boundary conditions [EM77] and [Ehr10] or transparent boundary conditions [AES03] and [Cou19], it seems to be more challenging to have a such easy-to-use theory. The work of Benoit [Ben22] may be a source of inspiration to work on a bounded space domain. In that case, we have to modify the stability estimates (6.9) to deal with the two boundaries and introduce two Kreiss-Lopatinskii determinant, one for each side.



6.6 Compléments

6.6.1 Preuve des résultats d'holomorphie

Dans la démonstration du Corollaire 6.14, on utilise le principe de l'argument énoncé et démontré en Théorème A.3 (page 200) avec les notations $a = 0$, $r = 1$, $\Omega = B(0,1) = \mathbb{D}$, $\Gamma = \mathbb{S}$ et $f = \Delta$.

La démonstration du Lemme 6.3 (Hersh) utilise le théorème de Rouché que l'on rappelle en Théorème A.5 (page 202) et il est démontré à la page 74.

Dans la preuve de la Proposition 6.18, on utilise l'holomorphie d'une intégrale à paramètre, et dans celle de la Proposition 6.19, on utilise la continuité d'une intégrale à paramètre, on justifie cela dans le résultat suivant.

Défini dans le Lemme 6.16, on rappelle que $R_z(X) = \prod_{j=1}^{r}(X - \kappa_j(z))$.

Lemme 6.23. *Soit $\varepsilon > 0$. On pose $\mathbb{S}_\varepsilon \overset{def}{=} \{z \in \mathbb{C}, |z| = 1+\varepsilon\}$. Pour tout $k \in \mathbb{Z}$, l'application $z \mapsto \int_{\mathbb{S}_\varepsilon} \frac{\zeta^k}{R_z(\zeta)} d\zeta$ est holomorphe sur $\mathcal{U}$ et continue sur $\overline{\mathcal{U}}$.*

Démonstration. On a

$$\int_{\mathbb{S}_\varepsilon} \frac{\zeta^k}{R_z(\zeta)} d\zeta = i \int_0^{2\pi} \frac{(1+\varepsilon)e^{i(k+1)\theta}}{R_z((1+\varepsilon)e^{i\theta})} d\theta.$$

On utilise le théorème d'holomorphie sous l'intégrale pour la fonction

$$f : (\theta, z) \in [0, 2\pi] \times \overline{\mathcal{U}} \mapsto \frac{(1+\varepsilon)e^{i(k+1)\theta}}{R_z((1+\varepsilon)e^{i\theta})} \in \mathbb{C}.$$

Comme pour tout $z \in \overline{\mathcal{U}}$, on a $|\kappa_j(z)| \leqslant 1$ où $j \in [\![1 : r]\!]$ et que $|(1+\varepsilon)e^{i\theta}| > 1$, la fonction $z \mapsto f(\theta, z)$ est bien continue sur $\overline{\mathcal{U}}$ pour tout $\theta \in [0, 2\pi]$ et holomorphe sur $\mathcal{U}$ grâce au Lemme 6.16. De plus, on a

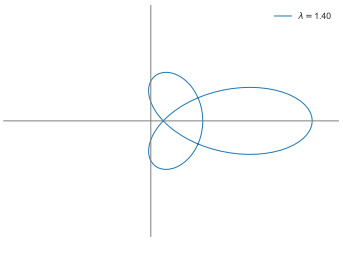
$$|f(\theta, z)| = \frac{(1+\varepsilon)}{\left|\prod_{j=1}^{r}((1+\varepsilon)e^{i\theta} - \kappa_j(z))\right|} \leqslant \frac{(1+\varepsilon)}{\varepsilon^r}$$

qui est une fonction intégrable sur $[0, 2\pi]$ et indépendante de z. Ce qui conclut la preuve. □

6.6.2 Lien entre $H_{0,j}(z)$ et $K_{0,j}(z)$

La matrice de Hermite $H_{0,j}(z)$ définie dans la preuve de la Proposition 6.17 est construite par blocs, pour chaque racine $\kappa_i(z)$ de multiplicité β_i, on construit un bloc avec β_i colonnes et $j+1$ lignes, en commençant par la colonne ${}^t(1\ \kappa_i(z)\ \kappa_i(z)^2\ \kappa_i(z)^3 \cdots \kappa_i(z)^j)$, puis en dérivant successivement les monômes de la colonne précédente.

L'intérêt de la matrice de Hermite est que si on prend un vecteur $Q = {}^t(q_0\ q_1 \cdots q_{m-1})$ qui représente le polynôme $\sum_{k=0}^{m-1} q_k X^k$ dans la base canonique, alors le produit ${}^t H_{0,m-1} Q$ donne le



vecteur d'évaluation en les $\kappa_1, \dots, \kappa_M$ avec les dérivées quand il y a de la multiplicité, autrement dit :

$$ {}^tH_{0,m-1}Q = {}^t(Q(\kappa_1)\ Q'(\kappa_1)\cdots Q^{(\beta_1-1)}(\kappa_1)\ Q(\kappa_2)\cdots Q^{(\beta_2-1)}(\kappa_2)\cdots Q^{(\beta_M-1)}(\kappa_M)) \in \mathbb{C}^r. $$

De plus, pour obtenir les coefficients du polynôme interpolateur de degré $r-1$, il suffit de multiplier par la matrice ${}^tH^{-1}_{0,r-1}$. Ainsi, l'application φ_z (définie en (6.21)) est représentée par la matrice ${}^tH^{-1}_{0,r-1}\,{}^tH_{0,m-1} \in \mathcal{M}_{r,m}(\mathbb{C})$ dans les bases canoniques de $\mathbb{C}_{m-1}[X]$ et $\mathbb{C}_{r-1}[X]$.

On veut maintenant expliciter le lien entre la matrice $H_{0,j}(z)$ et $K_{0,j}(z)$.

Supposons dans un premier temps que les matrices $H_{0,j}(z)$ et $K_{0,j}(z)$ ne dépendent que d'une seule racine $\kappa(z)$ de multiplicité β, ce qui est représenté par les deux matrices de dimension $(j+1)\times\beta$ suivantes :

$$ H_{0,j}(z) = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ \kappa & 1 & 0 & 0 & \cdots \\ \kappa^2 & 2\kappa & 2 & 0 & \cdots \\ \kappa^3 & 3\kappa^2 & 6\kappa & 6 & \cdots \\ \kappa^4 & 4\kappa^3 & 12\kappa^2 & 24\kappa & \cdots \\ \kappa^5 & 5\kappa^4 & 20\kappa^3 & 60\kappa^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ \kappa^j & j\kappa^{j-1} & j(j-1)\kappa^{j-2} & & \cdots \end{pmatrix} \text{ et } K_{0,j}(z) = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ \kappa & \kappa & \kappa & \kappa & \cdots \\ \kappa^2 & 2\kappa^2 & 4\kappa^2 & 8\kappa^2 & \cdots \\ \kappa^3 & 3\kappa^3 & 9\kappa^3 & 27\kappa^3 & \cdots \\ \kappa^4 & 4\kappa^4 & 16\kappa^4 & 64\kappa^4 & \cdots \\ \kappa^5 & 5\kappa^5 & 25\kappa^5 & 125\kappa^5 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ \kappa^j & j\kappa^j & j^2\kappa^j & j^3\kappa^j & \cdots \end{pmatrix} $$

On note $(h_{i,\ell})_{\substack{0\leqslant i\leqslant j\\ 0\leqslant\ell\leqslant\beta-1}}$ les coefficients de la matrice $H_{0,j}(z)$ et $(k_{i,\ell})_{\substack{0\leqslant i\leqslant j\\ 0\leqslant\ell\leqslant\beta-1}}$ les coefficients de la matrice $K_{0,j}(z)$. On a alors, pour tout $0\leqslant\ell\leqslant\beta-1$ et $0\leqslant i\leqslant j$,

$$ h_{i,\ell} = i(i-1)\dots(i-\ell+1)\kappa^{i-\ell} \quad \text{et} \quad k_{i,\ell} = i^\ell\kappa^i. $$

On peut remarquer que l'on a $h_{i,\ell+1} = \frac{\partial}{\partial\kappa}h_{i,\ell}$ et $k_{i,\ell+1} = \kappa\frac{\partial}{\partial\kappa}k_{i,\ell}$.

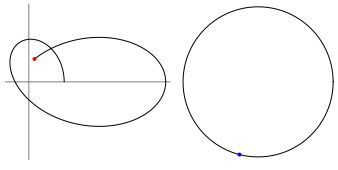
On va faire des opérations sur les colonnes afin d'exprimer les colonnes de $K_{0,j}(z)$ en fonction des colonnes de $H_{0,j}(z)$. On utilise pour cela les coefficients définis dans l'Annexe E (page 217).

On peut montrer par récurrence sur n que, pour tout $n\in[\![1:\beta-1]\!]$, on a

$$ \forall i\in[\![0:j]\!], \quad k_{i,n} = \sum_{\ell=1}^{n}\begin{bmatrix} n \\ \ell \end{bmatrix}\kappa^\ell h_{i,\ell}. \tag{6.36} $$

La preuve de ce résultat est très similaire à la preuve de la Proposition E.2 (page 218).

Soit $P_{\kappa,\beta}$ la matrice de changement de base telle que $H_{0,j}(z)P_{\kappa,\beta} = K_{0,j}(z)$. Grâce à la for-



mule (6.36), on a

$$P_{\kappa,\beta} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \binom{1}{1}\kappa & \binom{2}{1}\kappa & \cdots & \binom{\beta-1}{1}\kappa \\ \vdots & \ddots & \binom{2}{2}\kappa^2 & & \binom{\beta-1}{2}\kappa^2 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \binom{\beta-1}{\beta-1}\kappa^{\beta-1} \end{pmatrix} \in \mathcal{M}_\beta(\mathbb{C}).$$

Il faut remarquer que la matrice $P_{\kappa,\beta}$ ne dépend pas de j (puisque $P_{\kappa,\beta}$ agit sur les colonnes).

Pour le cas général avec des racines $\kappa_1(z), \dots, \kappa_M(z)$ de multiplicités respectives $\beta_1, \dots, \beta_M$ (avec $\beta_1 + \cdots + \beta_M = r$), comme on ne fait que des opérations par colonnes, il suffit de poser la matrice par blocs suivante :

$$P = \begin{pmatrix} P_{\kappa_1(z),\beta_1} & & 0 \\ & \ddots & \\ 0 & & P_{\kappa_M(z),\beta_M} \end{pmatrix} \in \mathcal{M}_r(C).$$

afin d'obtenir $H_{0,j}(z)P = K_{0,j}(z)$.
En particulier, on a

$$K_{0,m-1}(z)K_{0,r-1}^{-1}(z) = H_{0,m-1}(z)P(H_{0,r-1}(z)P)^{-1} = H_{0,m-1}(z)H_{0,r-1}^{-1}(z). \tag{6.37}$$

Cela justifie la preuve du Lemme 6.17.

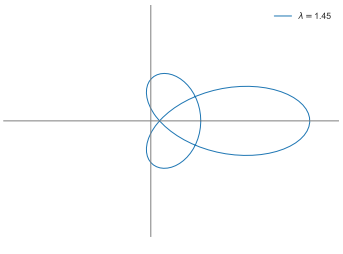
6.6.3 Preuve algébrique de la Proposition 6.20

On présente dans cette section une preuve alternative de la Proposition 6.20. Au lieu d'utiliser l'interpolation de Hermite qui permet de traiter les multiplicités aussi facilement que si les racines $(\kappa_j(z))_{j=1}^r$ étaient distinctes, on utilise une preuve algébrique en calculant explicitement les coefficients de la matrice $K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$. Cette preuve alternative se place dans le cas où toutes les racines $(\kappa_j(z))_{j=1}^r$ sont distinctes, les matrices $K_{0,m-1}(z)$ et $K_{0,r-1}(z)$ sont alors des matrices de Vandermonde classiques dont une expression explicite de l'inverse est connue : pour tout $i, j \in [\![1 : r]\!]$,

$$(K_{0,r-1}(z)^{-1})_{i,j} = (-1)^{j+i} \frac{\sum_{\substack{1 \leqslant i_1 < \cdots < i_{r-j} \leqslant r \\ i_1, \dots, i_{r-j} \neq i}} \kappa_{i_1}(z) \dots \kappa_{i_{r-j}}(z) \prod_{\substack{k>\ell \\ k,\ell \neq s}}^r (\kappa_k(z) - \kappa_\ell(z))}{\prod_{v>w}^r (\kappa_v(z) - \kappa_w(z))}.$$

Proposition 6.24. *Supposons les racines $(\kappa_j(z))_{j=1}^r$ distinctes, alors $\|K_{0,m-1}(z)K_{0,r-1}(z)^{-1}\|$ est borné indépendamment de $z \in \mathcal{U}$.*

Dans toute la démonstration suivante, on s'affranchit de la dépendance en z des racines



$(\kappa_j(z))_{j=1}^r$ afin de rendre la preuve plus lisible.

Démonstration. En utilisant la formule de la comatrice, pour tout $i \in [\![1:m]\!]$ et $j \in [\![1:r]\!]$, on peut effectuer le calcul suivant :

$$\begin{aligned}(K_{0,m-1}(z)K_{0,r-1}(z)^{-1})_{i,j} &= \sum_{s=1}^{r}(K_{0,m-1}(z))_{i,s}(K_{0,r-1}(z)^{-1})_{s,j} \\ &= \sum_{s=1}^{r}\kappa_s^{i-1}(-1)^{j+s}\frac{\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\dots,i_{r-j}\neq s}}\kappa_{i_1}\dots\kappa_{i_{r-j}}\prod_{\substack{k>\ell\\ k,\ell\neq s}}^{r}(\kappa_k-\kappa_\ell)}{\prod_{v>w}^{r}(\kappa_v-\kappa_w)}.\end{aligned}$$

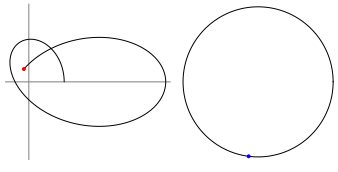
On veut montrer que le coefficient en (i,j) est un polynôme en $\kappa_1,\dots,\kappa_r$. Pour cela on se place dans l'anneau $\mathbb{Q}[X_1,\dots,X_r]$. Tous les polynômes $(X_v - X_w)$ sont premiers entre eux (pour tout $v > w$). Donc il suffit de vérifier que la somme

$$Q(X_1,\dots,X_r) = \sum_{s=1}^{r}X_s^{i-1}(-1)^{j+s}\left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\dots,i_{r-j}\neq s}}^{r}X_{i_1}\dots X_{i_{r-j}}\right)\prod_{\substack{k>\ell\\ k,\ell\neq s}}^{r}(X_k-X_\ell)$$

est divisible par $X_v - X_w$ pour tout $v > w$ pour que Q soit divisible par $\prod_{v>w}(X_v - X_w)$. Pour cela, il faut et il suffit que $Q(X_1,\dots,X_v,\dots,X_v,\dots,X_r) = 0$ où X_v est placé en v et en w.

Dans les termes de la somme sur s dans l'expression de Q, seuls les termes $s = v$ et $s = w$ nous intéressent car pour tout $s \neq v, w$ le produit pour $k > \ell$ et $k, \ell \neq s$ contient $(X_v - X_v)$ et donc annule les termes pour $s \neq v, w$.
D'où

$$\begin{aligned}&Q(X_1,\dots,X_v,\dots,X_v,\dots,X_r)\\ =&\left(X_w^{i-1}(-1)^{j+w}\left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\dots,i_{r-j}\neq w}}^{r}X_{i_1}\dots X_{i_{r-j}}\right)\prod_{\substack{k>\ell\\ k,\ell\neq w}}^{r}(X_k-X_\ell)\right)_{|X_w=X_v}\\ &+\left(X_v^{i-1}(-1)^{j+v}\left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\dots,i_{r-j}\neq v}}^{r}X_{i_1}\dots X_{i_{r-j}}\right)\prod_{\substack{k>\ell\\ k,\ell\neq v}}^{r}(X_k-X_\ell)\right)_{|X_w=X_v}\\ =&X_v^{i-1}(-1)^{j+w}\left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\dots,i_{r-j}\neq w}}^{r}X_{i_1}\dots X_{i_{r-j}}\right)\prod_{\substack{k>\ell\\ k,\ell\neq w}}^{r}(X_k-X_\ell)\\ &+X_v^{i-1}(-1)^{j+v}\left(\left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\dots,i_{r-j}\neq v}}^{r}X_{i_1}\dots X_{i_{r-j}}\right)\prod_{\substack{k>\ell\\ k,\ell\neq v}}^{r}(X_k-X_\ell)\right)_{|X_w=X_v}.\end{aligned}$$



Lemme 6.25. *On a*

$$(-1)^w \prod_{\substack{k>\ell\\ k,\ell\neq w}}^{r} (X_k - X_\ell) = -(-1)^v \left(\prod_{\substack{k>\ell\\ k,\ell\neq v}}^{r} (X_k - X_\ell) \right)_{|X_w=X_v}. \tag{6.38}$$

Démonstration. Il suffit de faire le calcul suivant :

$$\begin{aligned}
&(-1)^v \left(\prod_{\substack{k>\ell\\ k,\ell\neq v}}^{r} (X_k - X_\ell) \right)_{|X_w=X_v} \\
=&(-1)^v \prod_{\substack{k>\ell\\ k,\ell\neq v,w}}^{r} (X_k - X_\ell) \left(\prod_{\ell=1}^{w-1} (X_w - X_\ell) \prod_{k=w+1}^{v-1} (X_k - X_w) \prod_{v+1}^{r} (X_k - X_w) \right)_{|X_w=X_v} \\
=&(-1)^v \prod_{\substack{k>\ell\\ k,\ell\neq v,w}}^{r} (X_k - X_\ell) \prod_{\ell=1}^{w-1} (X_v - X_\ell) \prod_{k=w+1}^{v-1} (X_k - X_v) \prod_{v+1}^{r} (X_k - X_v) \\
=&(-1)^v \prod_{\substack{k>\ell\\ k,\ell\neq v,w}}^{r} (X_k - X_\ell) \prod_{\ell=1}^{w-1} (X_v - X_\ell)(-1)^{v-w-1} \prod_{\ell=w+1}^{v-1} (X_v - X_\ell) \prod_{v+1}^{r} (X_k - X_v) \\
=&(-1)^{w+1} \prod_{\substack{k>\ell\\ k,\ell\neq w}}^{r} (X_k - X_\ell).
\end{aligned}$$

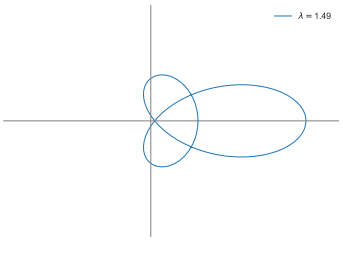
□

Lemme 6.26. *On a*

$$\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\ldots,i_{r-j}\neq w}}^{r} X_{i_1}\ldots X_{i_{r-j}} = \left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\ldots,i_{r-j}\neq v}}^{r} X_{i_1}\ldots X_{i_{r-j}} \right)_{|X_w=X_v}. \tag{6.39}$$

Démonstration. Il suffit de faire le calcul suivant :

$$\begin{aligned}
&\left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\ldots,i_{r-j}\neq v}}^{r} X_{i_1}\ldots X_{i_{r-j}} \right)_{|X_w=X_v} \\
&\qquad = \sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\ldots,i_{r-j}\neq v,w}}^{r} X_{i_1}\ldots X_{i_{r-j}} + \left(\sum_{\substack{1\leqslant i_1<\cdots<i_{r-j}\leqslant r\\ i_1,\ldots,i_{r-j}\neq v\\ \exists k, i_k=w}}^{r} X_{i_1}\ldots X_{i_{r-j}} \right)_{|X_w=X_v}.
\end{aligned}$$



L'application $\sigma : (i_1, \dots, i_{r-j}) \mapsto (i_1, \dots, w, \dots, i_{k-1}, i_{k+1}, \dots, i_{r-j})$ où k est l'indice tel que $i_k = v$ et w est inséré au bon endroit (ordre strict) est une bijection entre les ensembles

$$\{(i_1, \dots, i_{r-j}) | 1 \leqslant i_1 < \dots < i_{r-j} \leqslant r, i_1, \dots, i_{r-j} \neq w, \exists k \ i_k = v\}$$

$$\text{et} \quad \{(i_1, \dots, i_{r-j}) | 1 \leqslant i_1 < \dots < i_{r-j} \leqslant r, i_1, \dots, i_{r-j} \neq v, \exists k \ i_k = w\}.$$

Ainsi, on trouve

$$\begin{aligned}\left(\sum_{\substack{1 \leqslant i_1 < \dots < i_{r-j} \leqslant r \\ i_1, \dots, i_{r-j} \neq v}}^{r} X_{i_1} \dots X_{i_{r-j}} \right)_{|X_w = X_v} &= \sum_{\substack{1 \leqslant i_1 < \dots < i_{r-j} \leqslant r \\ i_1, \dots, i_{r-j} \neq v, w}}^{r} X_{i_1} \dots X_{i_{r-j}} + \sum_{\substack{1 \leqslant i_1 < \dots < i_{r-j} \leqslant r \\ i_1, \dots, i_{r-j} \neq w \\ \exists k, i_k = v}}^{r} X_{i_1} \dots X_{i_{r-j}} \\ &= \sum_{\substack{1 \leqslant i_1 < \dots < i_{r-j} \leqslant r \\ i_1, \dots, i_{r-j} \neq w}}^{r} X_{i_1} \dots X_{i_{r-j}}.\end{aligned}$$

□

Pour conclure la preuve de la Proposition 6.24, on utilise les équations (6.38) et (6.39), on trouve donc bien $Q(X_1, \dots, X_v, \dots, X_v, \dots, X_r) = 0$ pour X_v placé en v et en w.

Comme, pour tout $k \in [\![1 : r]\!]$ et pour tout $|z| > 1$, on a $|\kappa_k(z)| < 1$, cela prouve le résultat. □

6.7 Extensions aux schémas multi-pas

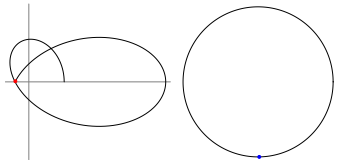
On se fixe des entiers r, p, m et s. L'entier s correspond au nombre de pas en temps.

6.7.1 Résultats théoriques

On définit un schéma multipas de la manière suivante :

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} \sum_{\ell=-r}^{p} a_{\ell,\sigma} U_{j+\ell}^{n-\sigma}, & n \geqslant s, \ j \geqslant r, \quad (6.40a) \\ U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} \sum_{\ell=0}^{m-1} b_{j,\ell}^{(\sigma)} U_\ell^{n-\sigma} + g_j^{n+1}, & n \geqslant s, \ j \in [\![0 : r-1]\!], \quad (6.40b) \\ U_j^n = f_j^n, & n \leqslant s, \ j \geqslant 0. \quad (6.40c) \end{cases}$$

Le nombre de pas en temps peut différer entre l'équation intérieure (6.40a) et la condition de bord (6.40b), mais il suffit de prendre le nombre de pas le plus élevé et de mettre à zéros les coefficients des indices rajoutés. On peut visualiser ce schéma (6.40) en Figure D.1 (page 209).



En passant à la transformée en $\mathcal{Z}$, l'équation (6.40a) devient :

$$z^{s+1}\widetilde{U}_j(z) = \sum_{\ell=-r}^{p} \left(\sum_{\sigma=0}^{s} a_{\ell,s-\sigma} z^{\sigma} \right) \widetilde{U}_{j+\ell}(z). \tag{6.41}$$

L'équation caractéristique de (6.41) est alors

$$c_p(z)X^{p+r} + \cdots + c_0(z)X^r + \cdots + c_{-r}(z) = 0 \tag{6.42}$$

avec $c_\ell(z) = \sum_{k=0}^{s} a_{\ell,s-k} z^k$ pour tout $\ell \neq 0$ et $c_0(z) = \sum_{k=0}^{s} a_{0,s-k} z^k - z^{s+1}$. Classiquement (voir [GKS72] et [Cou13]) et afin que (6.42) soit une suite récurrente linéaire d'ordre exactement $p+r$, on effectue l'hypothèse suivante.

Hypothèse 6.27 (Non dégénérescence)**.** *On suppose que $c_p(z)$ et $c_{-r}(z)$ sont non nuls pour tout $z \in \overline{\mathcal{U}}$.*

De plus, la Cauchy-stabilité du schéma (6.40a) se traduit par l'hypothèse suivante.

Hypothèse 6.28 (Cauchy-stabilité)**.** *Le schéma* (6.40a) *est* Cauchy-stable *s'il existe une constante $C > 0$ telle que*

$$\forall n \in \mathbb{N},\ \forall \xi \in \mathbb{R}, \quad |\mathscr{A}(\xi)^n| \leqslant C$$

où

$$\mathscr{A}(\xi) = \begin{pmatrix} \sum_{\ell=-r}^{p} a_{\ell,0} e^{i\ell\xi} & \cdots & \cdots & \sum_{\ell=-r}^{p} a_{\ell,s} e^{i\ell\xi} \\ 1 & 0 & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix} \in \mathcal{M}_{s+1}(\mathbb{C}).$$

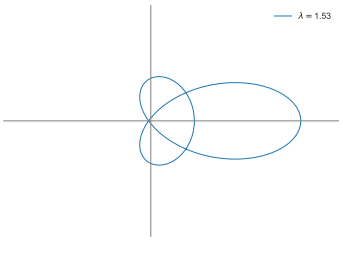
Sous l'Hypothèse 6.28 (Cauchy-stabilité), le Lemme 3.4 (Hersh) est toujours valable dans le cas multipas (voir [Cou13, Lem.3.7]). Pour $z \in \overline{\mathcal{U}}$, on sélectionne donc les r racines $(\kappa_j(z))$ venant du disque unité. On utilise à nouveau la notation $K_{0,j}(z)$ comme explicitée autour de l'équation (6.16).

Pour définir le déterminant de Kreiss–Lopatinskii, il faut injecter les solutions $(\widetilde{U}_j(z))_j$ dans la version transformée en $\mathcal{Z}$ des équations du bord (6.40b), ce qui donne

$$z^{s+1}\widetilde{\mathfrak{U}}_{r-1}(z) = (z^s B_0 + \cdots + zB_{s-1} + B_s)\widetilde{\mathfrak{U}}_{m-1}(z)$$

où

$$\forall j \geqslant 0, \quad \widetilde{\mathfrak{U}}_j(z) = \begin{pmatrix} \widetilde{U}_0(z) \\ \vdots \\ \widetilde{U}_j(z) \end{pmatrix} \quad \text{et} \quad \forall \sigma \in [\![0:s]\!], \quad B_\sigma = \begin{pmatrix} b_{0,0}^{(\sigma)} & \cdots & b_{0,m-1}^{(\sigma)} \\ \vdots & & \vdots \\ b_{r-1,0}^{(\sigma)} & \cdots & b_{r-1,m-1}^{(\sigma)} \end{pmatrix}.$$



Le déterminant intrinsèque de Kreiss–Lopatinskii se définit alors de la manière suivante :

$$\Delta(z) = \frac{\det\left(z^{s+1}K_{0,r-1}(z) - \left(\sum_{k=0}^{s} B_k z^{s-k}\right) K_{0,m-1}(z)\right)}{\det K_{0,r-1}(z)}.$$

De la même façon qu'à l'équation (6.19), on peut réécrire $\Delta(z)$ de la manière suivante :

$$\Delta(z) = z^{r(s+1)} \det\left(I_r - \sum_{k=0}^{s} B_k \frac{1}{z^{k+1}} K_{0,m-1}(z) K_{0,r-1}^{-1}(z)\right). \tag{6.43}$$

Grâce à l'Hypothèse 6.27, pour tout $\ell \in [\![-r : p-1]\!]$, les fonctions $z \mapsto \frac{c_\ell(z)}{c_p(z)}$ sont holomorphes sur $\mathcal{U}$, ainsi le Lemme 6.16 s'applique. La fonction $z \mapsto K_{0,m-1}(z)K_{0,r-1}^{-1}(z)$ est donc holomorphe sur $\mathcal{U}$ et continue sur $\overline{\mathcal{U}}$ par les mêmes arguments que dans le Théorème 6.12.

Grâce à la formulation (6.43), comme dans la preuve du Corollaire 6.14, le pôle en 0 de la fonction $z \mapsto \Delta(1/z)$ est d'ordre $r(s+1)$. Ainsi le Corollaire 6.14 se traduit dans le cas multipas de la manière suivante.

Corollary 6.29 (Nombre de zéros du déterminant intrinsèque de Kreiss–Lopatinskii)**.** *Sous les Hypothèses 6.27 et 6.28, si* $0 \notin \Delta(\mathbb{S})$*, alors l'équation* $\Delta(z) = 0$ *a exactement*

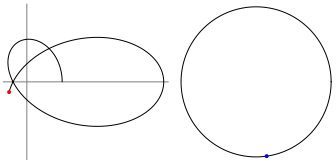
$$r(s+1) - \mathrm{Ind}_{\Delta(\mathbb{S})}(0)$$

solutions dans $\mathcal{U}$*.*

6.7.2 Résultats numériques

En pratique, on peut utiliser la même technique que dans le Lemme 6.21. On part de la matrice $\mathcal{B} = \sum_{k=0}^{s} B_k z^{s-k}$ qui est de taille $r \times m$, on effectue les opérations du Lemme 6.21 afin d'arriver à la matrice $\widetilde{\mathcal{B}}(\sigma_0, \dots, \sigma_{r-1})$. Ensuite, pour chaque $z \in \overline{\mathcal{U}}$, on calcule les coefficients $c_\ell(z)$ pour $\ell \in [\![-r : p]\!]$, on trouve numériquement les solutions de (6.42), on sélectionne les r racines qui viennent de l'intérieur du disque unité avec le Lemme 6.22 (dont la preuve doit être adaptée par rapport à la preuve donnée en Section 7.1.1, page 177), on calcule les r fonctions symétriques que l'on injecte dans la formulation $\widetilde{B}(\sigma_0(z), \dots, \sigma_{r-1}(z))$. On a alors une expression numérique du déterminant intrinsèque de Kreiss–Lopatinskii, comme à l'équation (6.27). En calculant l'indice complexe (voir Section 7.2 pour plus de détails), on peut trouver s'il y a ou non des instabilités.

Pour faire le lien avec l'article [BLBS23a], si le schéma multipas est totalement décentré (*i.e.* $p = 0$), les coefficients $\sigma_0(z), \dots, \sigma_{r-1}(z)$ sont alors très facile à calculer car ils correspondent aux coefficients de l'équation caractéristique (6.42). On peut alors donner une expression du déterminant de Kreiss–Lopatinskii sans avoir à calculer les racines de l'équation caractéristique, comme dans le cas à un pas du Chapitre 5 dédié à l'article [BLBS23a].



On trace maintenant les courbes du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma leap-frog avec les conditions de bord de l'article [Tre84] que l'on rappelle ci-dessous :

$$\alpha : \quad U_0^{n+1} = U_1^n \tag{α}$$
$$\beta : \quad U_0^{n+1} = U_1^{n-2} \tag{β}$$
$$\gamma : \quad U_0^{n+1} = \frac{1}{2}(U_0^n + U_2^n) \tag{γ}$$
$$\delta : \quad U_0^{n+1} = U_1^{n+1} \tag{δ}$$

Pour plus de détails sur le schéma leap-frog et sur ces conditions de bord, on peut se référer à l'Annexe D (page 209).

La Figure 6.8 est à mettre en parallèle avec les pseudospectres tracés dans la Section 2.4.4 (page 66).

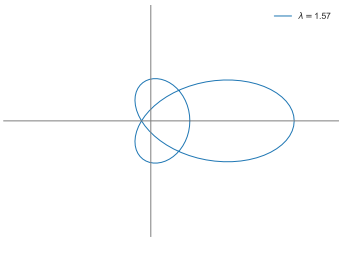
Les conditions de bord (β), (γ) et (δ) sont des cas instables car elles engendrent des valeurs propres généralisées (voir Section D.2, page 213). On voit qu'en effet la courbe $\Delta(\mathbb{S})$ du déterminant intrinsèque de Kreiss–Lopatinskii passe par l'origine. Par exemple, pour le cas (δ) où la valeur propre généralisée est en $z = -1$, le déterminant de Kreiss–Lopatinskii s'annule justement pour cette valeur $z = -1$.

La condition de bord (α) est un cas stable. Contrairement à ce que laisse penser la Figure 6.8, la courbe fait deux tours autour de l'origine car quand z parcourt $\mathbb{S}$, la courbe $\Delta(\mathbb{S})$ parcourt deux fois la courbe de la Figure 6.8. On peut voir que la courbe du déterminant intrinsèque de Kreiss–Lopatinskii ne passe pas par l'origine, donc il n'y a pas de zéros de Δ sur $\mathbb{S}$. Ensuite, on peut appliquer le Corollaire 6.29 : le nombre de zéros de Δ sur $\mathcal{U}$ est

$$r(s+1) - \mathrm{Ind}_{\Delta(\mathbb{S})}(0) = 1 \times (1+1) - 2 = 0.$$

Cela confirme que le schéma dans le cas (α) est stable.

Le folioscope de gauche entre les pages 199 et 230 montre que la courbe du déterminant intrinsèque de Kreiss–Lopatinskii pour le cas (α) fait deux tours autour de l'origine.



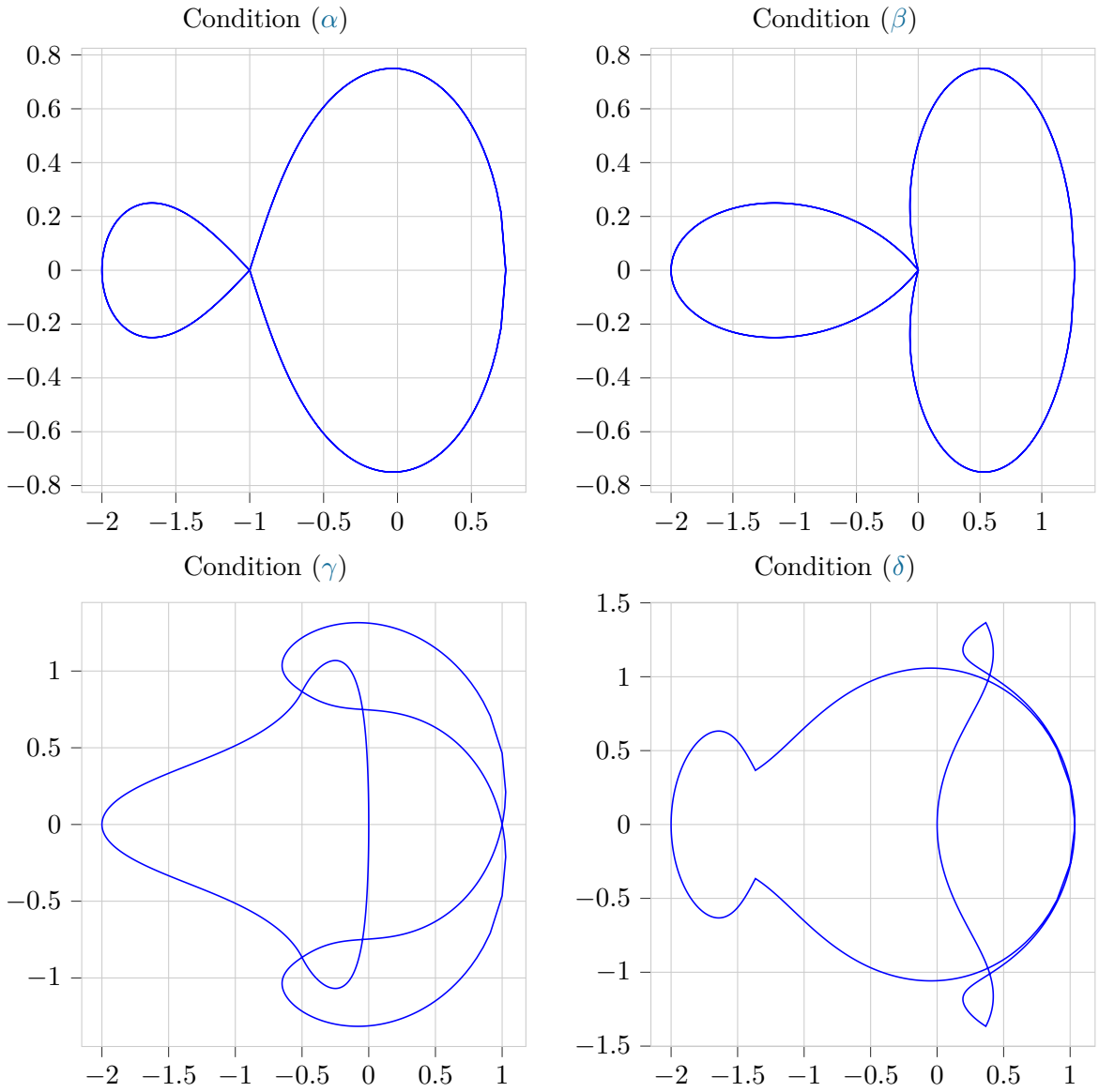
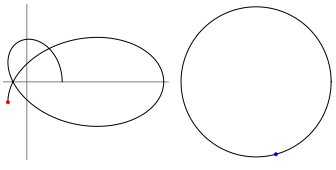


FIGURE 6.8 – Courbe du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma leap-frog avec différentes conditions de bord.



Chapitre 7

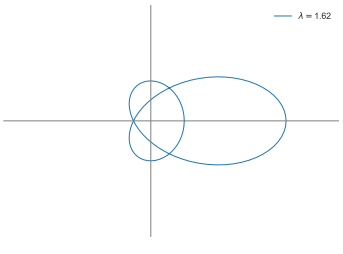
Aspects d'implémentation numérique

Dans ce chapitre, on va voir les détails de l'implémentation numérique de la stratégie (Méthode 5.19, page 116, et Méthode 6.15, page 153) adoptée dans les Chapitres 5 et 6. On commence par décrire comment calculer le déterminant intrinsèque de Kreiss–Lopatinskii, notamment en expliquant comment sélectionner les racines $(\kappa_j)_{j=1}^r$ de (3.4) issues du disque unité ouvert, qu'on appelle *racines stables* dans tout ce chapitre. Ensuite, on donne la stratégie numérique utilisée pour calculer l'indice complexe, étape cruciale pour utiliser les Corollaire 5.15 (page 114) et Corollaire 6.14 (page 153) et conclure sur la stabilité des schémas. Enfin, on décrit l'implémentation en Python des schémas et des conditions de bord. Le but, à moyen terme, est de mettre le code en libre-accès sur le dépôt GitHub : `https://github.com/PLeBarbenchon/boundaryscheme`.

7.1 Calcul du déterminant intrinsèque de Kreiss–Lopatinskii

7.1.1 Sélection des racines stables

Cette section décrit l'utilisation du Lemme 6.22 (qui permet de compter la multiplicité des racines se trouvant sur le cercle unité) et en donne une preuve rigoureuse.



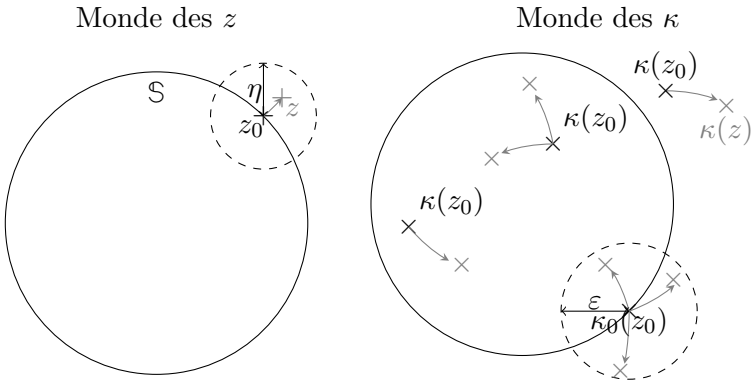


FIGURE 7.1 – Illustration de la Proposition 7.1.

Pour tracer le déterminant intrinsèque de Kreiss–Lopatinskii en $z_0 \in \mathbb{S}$, il faut être capable de construire les fonctions symétriques des racines stables $\kappa(z_0)$.

Quand $|z_0| > 1$, on sait, par le Lemme 3.4 (Hersh), que les r racines $\kappa(z_0)$ qui nous intéressent sont à l'intérieur du disque unité ouvert $\mathbb{D}$, il suffit d'ordonner les racines $\kappa(z_0)$ par module croissant et de prendre les r premières.

Quand $|z_0| = 1$, il faut sélectionner les $\kappa(z_0)$ qui viennent de l'intérieur et non de l'extérieur du disque unité. Pour une valeur $\kappa_0(z_0) \in \mathbb{S}$, on va faire varier z légèrement vers l'extérieur du disque unité et on va compter combien il y a de racines $\kappa_0(z)$ à l'intérieur et à l'extérieur, on utilise la continuité des racines par rapport à z. Mais il faut faire attention à ne prendre que des racines $\kappa_0(z)$ issue de $\kappa_0(z_0)$ et non d'autres racines $\kappa(z_0)$. Pour cela, on va se placer dans une boule autour de $\kappa_0(z_0)$ de rayon ε, où ε est la plus petite distance entre les valeurs $\kappa(z_0)$. En ce plaçant sur un compact autour de z_0, l'uniforme continuité nous donne

$$\forall \varepsilon > 0,\ \exists \eta > 0,\ \forall z \in B(z_0, \eta),\ |\kappa(z) - \kappa(z_0)| < \varepsilon.$$

On veut quantifier le module de continuité η autour de z_0 afin d'être sûr que les racines $\kappa_0(z)$ seront bien issues de $\kappa_0(z_0)$. Pour cela, on va utiliser le théorème de Rouché (voir Théorème A.5, page 202) qui permet de compter les racines à l'intérieur d'une courbe fermée.

On rappelle l'énoncé du Lemme 6.22.

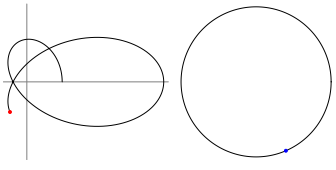
Proposition 7.1. *Soit $\kappa_0(z_0)$ racine de module 1 et de multiplicité μ du polynôme P_{z_0} défini par*

$$a_{-r} + a_{-r+1}X + \cdots + a_{-1}X^{r-1} + (a_0 - z_0)X^r + a_1X^{r+1} + \cdots + a_pX^{p+r}.$$

Soit $\varepsilon > 0$ telle que $\kappa_0(z_0)$ soit la seule racine de P_{z_0} dans $\overline{B(\kappa_0(z_0), \varepsilon)}$ On pose

$$\eta = (1+\varepsilon)^{-r} \min_{\kappa \in \partial B(\kappa_0(z_0), \varepsilon)} |P_{z_0}(\kappa)|.$$

Alors, pour tout $z \in B(z_0, \eta)$, le polynôme P_z possède exactement μ racines comptées avec



multiplicité dans la boule $B(\kappa_0(z_0), \varepsilon)$.

Démonstration. Soit $z \in B(z_0, \eta)$. On utilise le théorème de Rouché pour compter le nombre de zéros de P_z à l'intérieur de $B(\kappa_0(z_0), \varepsilon)$. Pour tout $\kappa \in \partial B(\kappa_0(z_0), \varepsilon)$, on a

$$|P_z(\kappa) - P_{z_0}(\kappa)| < \eta|\kappa|^r \leqslant \eta(1+\varepsilon)^r \leqslant \min_{\kappa \in \partial B(\kappa_0(z_0),\varepsilon)} |P_{z_0}(\kappa)| \leqslant |P_{z_0}(\kappa)|.$$

Il y a, ainsi, autant de racines, comptées avec multiplicité, à l'intérieur de $B(\kappa_0(z_0), \varepsilon)$ pour les polynômes P_z et P_{z_0}. Ce qui conclut la preuve. □

Ainsi si numériquement on tombe sur une valeur de $\kappa_0(z_0)$ sur le cercle unité, il faut chercher les racines $\kappa(z)$ dans $B(\kappa_0(z_0), \varepsilon)$ pour $z = z_0 + \eta\frac{z_0}{|z_0|}$ et pour ε la distance minimale entre deux racines distinctes de P_{z_0}.

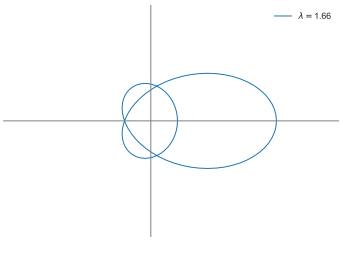
On donne ci-dessous des fragments de code qui permettent de sélectionner les racines stables κ de l'équation caractéristique (3.4) avec la multiplicité adéquate. On utilise le module Polynomial de NumPy [HMvdW+20] pour approcher les racines d'un polynôme.

```
def epsilon(L):
    """L is a list of elements xi and return min |xi - xj| / 2 (when xi != xj)
    """
    assert len(L) > 1
    diff = []
    for i in range (len(L)):
        for j in range (i + 1, len(L)):
            if L[i] != L[j]:
                diff.append(abs(L[i] - L[j]))
    mini = min(diff)
    assert mini > 0
    return mini / 2
```

Code Python 7.1 – Calcul de ε la distance minimale entre deux éléments d'une liste L.

```
def eta_func(eps, kappa0, N, polynom, r):
    """return min |polynom(K)|/(1+eps)^r for K on the circle centered in kappa0
    of radius eps where N is the number of discretization of the circle"""
    theta = np.linspace(0, 2 * pi, N)
    circle = np.cos(theta) + 1j * np.sin(theta)
    kappas = kappa0 + eps * circle
    val_pol = np.abs(polynom(kappas))
    mini = min(val_pol)
    assert mini > 0
    return mini / ((1 + eps)**r)
```

Code Python 7.2 – Calcul de η défini dans la Proposition 7.1.



```python
def count_root(self,eta, eps, z0, kappa):
    z = z0 + eta * z0 / (2*abs(z0))
    NewRoots = self.roots(z)#return the p+r roots of the characteristic equation
    selection = list(filter(lambda k : abs(k-kappa)<eps, NewRoots))
    return len(list(filter(lambda k : abs(k)<1, selection)))

def Kappa(self,z0):
    """ selection of kappas that come from the inside of the unit disk """
    delta = 10**(-10)
    allRoots = self.roots(z0)#return the p+r roots of the characteristic
    equation
    eps, RootsFromInside, stock = epsilon(allRoots), [], []
    for x in allRoots:
        if abs(x)<1-delta:
            RootsFromInside.append(x)
        elif abs(x)<1+delta and x not in stock:
            stock.append(x)
            eta = eta_func(eps,x,1000,self.pol(z0),self.center)
            #self.pol(z0) is the characteristic equation with parameter z0
            n = self.count_root(eta, eps, z0, x)
            for i in range (n):
                RootsFromInside.append(x)
    assert len(RootsFromInside) == self.center
    return RootsFromInside
```

Code Python 7.3 – Sélection des racines stables avec la multiplicité adéquate.

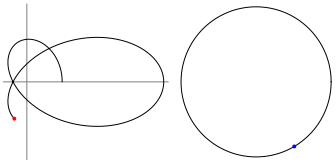
La procédure `self.roots(z)` renvoie toutes les racines de l'équation caractéristique (3.4), elle utilise les fonctions du module Polynomial de NumPy.

7.1.2 Calcul symbolique et algorithme de réduction de la matrice du bord

Maintenant que l'on a sélectionné les racines $(\kappa_j)_{j=1}^r$, il faut réduire la matrice $\mathfrak{B}$ liée au bord pour pouvoir écrire le déterminant intrinsèque de Kreiss–Lopatinskii comme dans l'équation (6.27). Cela correspond au Lemme 5.21 (page 117) et à la Proposition 6.21 (page 157).

La notation $\mathfrak{B}$ correspond à l'encodage de la condition de bord. Comme expliqué en Section 3.3.2, le Chapitre 5 utilise la matrice B et la vision $\det(K_{-r,-1}(z) - BK_{0,m-1}(z))$ du déterminant de Kreiss–Lopatinskii et le Chapitre 6 utilise la matrice $\mathcal{B}$ et la vision $\det(zK_{0,r-1}(z) - \mathcal{B}K_{0,m-1}(z))$ pour le déterminant de Kreiss–Lopatinskii. Dans le code, on utilise cette deuxième vision du déterminant de Kreiss–Lopatinskii.

Pour gagner en efficacité dans le calcul de la courbe du déterminant intrinsèque de Kreiss–Lopatinskii, on peut n'effectuer la réduction de la matrice $\mathfrak{B}$ qu'une seule fois, en utilisant des variables symboliques pour z et pour les fonctions symétriques $\sigma_0, \dots, \sigma_{r-1}$ des racines $(\kappa_j)_j$.



```python
def DKL(self):
    """
    use the boundary part B of a quasi-Toeplitz matrix and return the formula of the intrinsic Kreiss--Lopatinskii determinant with symbolic variables for z0 and the r symmetric functions of the roots kappa (in the vector b)
    """
    z0 = sp.Symbol("z0", imaginary = True)
    B = copy.deepcopy(self.boundary_quasi_toep)
    B = sp.Matrix(B) - z0*sp.eye(self.r,self.m)
    b = sp.ones(1,self.r+1)
    for i in range (self.r+1):
        b[i] = sp.Symbol("b"+str(i), imaginary = True)
    for j in range (self.m-self.r):
        row = sp.zeros(1,self.m)
        for k in range (self.r+1):
            row[self.m-self.r-j+k-1] = b[k]
        B = B - np.dot(B[:,self.m-1-j],row)
    fdetKL = sp.lambdify([z0,b],sp.det(B[:self.r,:self.r]),"numpy")
    return fdetKL
```

Code Python 7.4 – Calcul symbolique du déterminant intrinsèque de Kreiss–Lopatinskii.

Ensuite, pour chaque $z \in \mathbb{S}$, on sélectionne les $(\kappa_j(z))_j$ en utilisant les méthodes décrites à la Section 7.1.1, on calcule les fonctions symétriques $\sigma_0(z), \dots, \sigma_{r-1}(z)$ des $(\kappa_j(z))_j$ et on renvoie le déterminant intrinsèque de Kreiss–Lopatinskii.

```python
DKL_formula = self.DKL()
def scalar_detKL(self,z,DKL_formula):
    Rz = nppol.polyfromroots(self.Kappa(z))
    #polyfromroots(L) returns the canonical coefficients of the polynomial whose roots are the elements of the list L
    return DKL_formula(z,Rz)
```

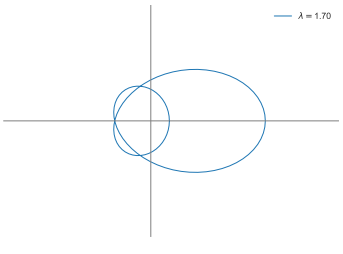
Code Python 7.5 – Calcul du déterminant intrinsèque de Kreiss–Lopatinskii.

7.2 Calcul de l'indice complexe

Maintenant que l'on peut tracer la courbe $\Delta(\mathbb{S})$ du déterminant intrinsèque de Kreiss–Lopatinskii, on veut calculer l'indice complexe de l'origine par rapport à $\Delta(\mathbb{S})$ pour pouvoir utiliser les Corollaire 5.15 et Corollaire 6.14.

7.2.1 Ligne polygonale

La courbe $\Delta(\mathbb{S})$ est approchée par une ligne polygonale. Pour ce faire, on discrétise le cercle unité $\mathbb{S}$ avec N points $\theta_0, \theta_1, \dots, \theta_{N-1} \in \mathbb{S}$. On trace ensuite les segments $[\Delta(e^{i\theta_j}), \Delta(e^{i\theta_{j+1}})]$ pour tout $j \in [\![0 : N-1]\!]$, où $\theta_N \overset{\text{def}}{=} \theta_0$.



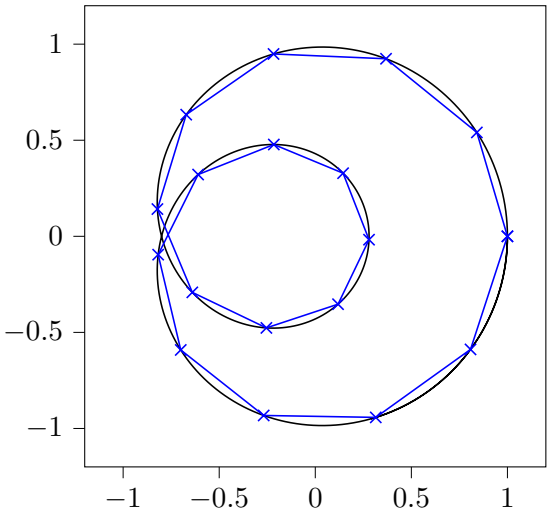


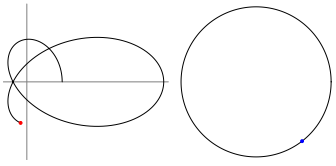
FIGURE 7.2 – Exemple de tracé de ligne polygonale.

Pour calculer l'indice complexe de cette ligne polygonale, il suffit de compter le nombre de tours que la ligne polygonale fait autour de l'origine. Pour cela, on s'inspire de l'algorithme présenté dans [O'R98] et complété dans [Sun21] et [Fra06]. On retranscrit ci-dessous l'algorithme utilisé en pratique.

```python
def is_left(P0, P1, P2):
    """take three points P0, P1, and P2
    return  =0 if P2 is on the line through P0 and P1
            >0 if P2 is on the left of the line
            <0 if P2 is on the right of the line"""
    return (P1[0] - P0[0]) * (P2[1] - P0[1]) - (P2[0] - P0[0]) * (P1[1] - P0[1])

def wn_PnPoly(P, V):
    """take a point P and a list V of vertex points of a polygon
    return wn the winding number of P with respect to the curve V"""
    wn = 0
    V = tuple(V[:]) + (V[0],)
    for i in range(len(V)-1):     # edge from V[i] to V[i+1]
        if V[i][1] <= P[1]:         # start y <= P[1]
            if V[i+1][1] > P[1]:      # an upward crossing
                if is_left(V[i], V[i+1], P) > 0: # P left of edge
                    wn += 1            # have a valid up intersect
        else:                       # start y > P[1] (no test needed)
            if V[i+1][1] <= P[1]:     # a downward crossing
                if is_left(V[i], V[i+1], P) < 0: # P right of edge
                    wn -= 1            # have a valid down intersect
    return wn
```

Code Python 7.6 – Calcul de l'indice complexe de l'origine par rapport à une ligne polygonale.



Le problème d'une discrétisation uniforme du cercle unité est que si la courbe se rapproche de l'origine, le calcul d'indice complexe est rendu difficile, comme on le voit à la Figure 7.3.

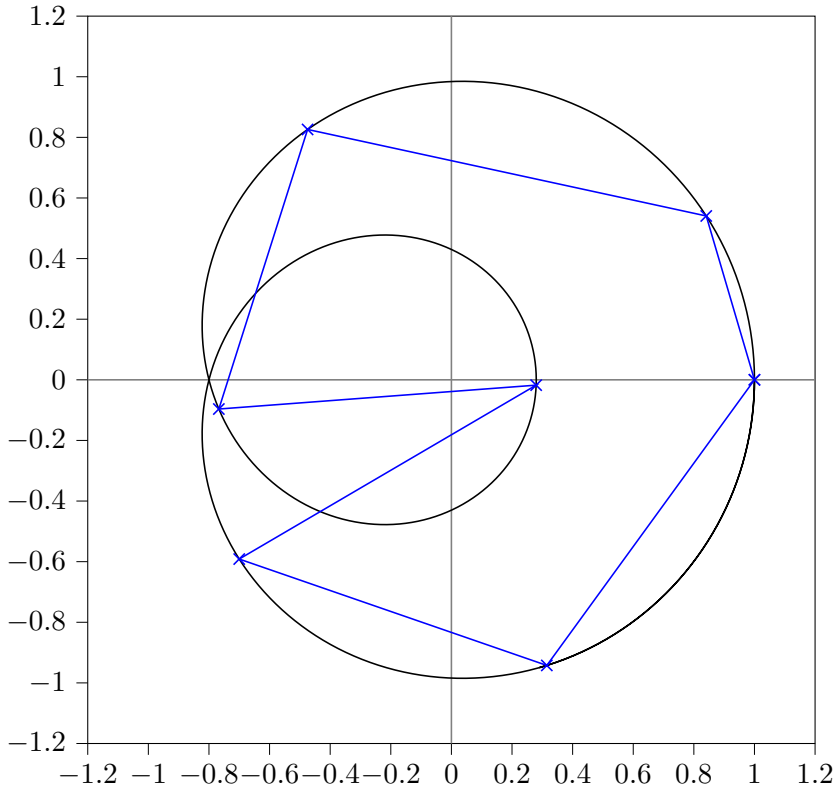


FIGURE 7.3 – Exemple de calcul incorrect de l'indice complexe.

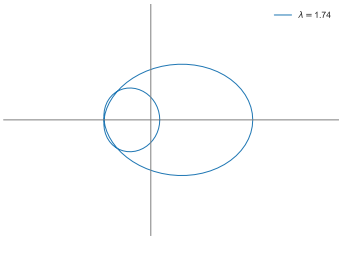
La discrétisation, utilisée dans la Figure 7.3, ne permet pas de calculer correctement l'indice complexe de la courbe. En effet, l'indice complexe de la courbe noire vaut 2 alors que l'indice de la ligne polygonale bleue vaut 1.

Dans la section suivante, on va raffiner localement la discrétisation du cercle unité pour avoir un calcul de l'indice plus robuste.

7.2.2 Raffinement de la discrétisation de la courbe

Quand la courbe $\Delta(\mathbb{S})$ est proche de l'origine, si la discrétisation n'est pas assez fine, l'indice complexe ne sera pas calculé correctement. On peut alors utiliser la procédure décrite dans les travaux de Garcia Zapata et Diaz Martin [GZDM12, GZDM14].

Le principe est de séparer le plan complexe $\mathbb{C}$ en les huit secteurs angulaires $S_0, \ldots, S_7$ présentés en Figure 7.4.



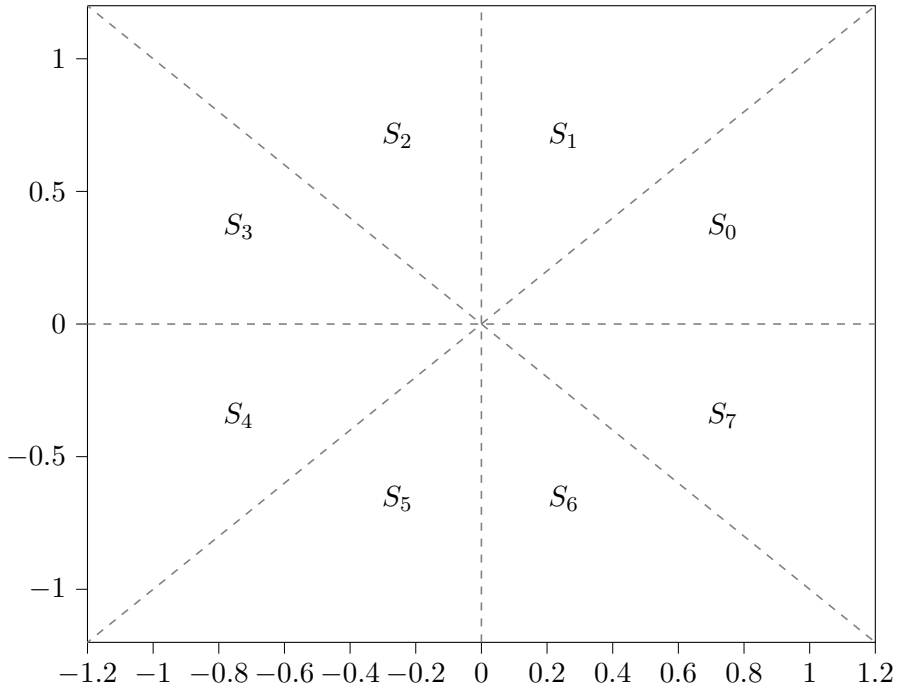


FIGURE 7.4 – Huit secteurs angulaires de la procédure de [GZDM12].

Dès lors qu'un segment de la discrétisation de la courbe $\Delta(\mathbb{S})$ coupe deux frontières de secteurs angulaires, on discrétise plus finement la courbe $\Delta(\mathbb{S})$ afin que deux points soient toujours dans des secteurs angulaires voisins (on parle de *secteurs voisins* quand ils partagent une demi-droite du plan complexe). Grâce à cela, la discrétisation est plus fine, localement au voisinage de l'origine, les secteurs étant centrés en l'origine.

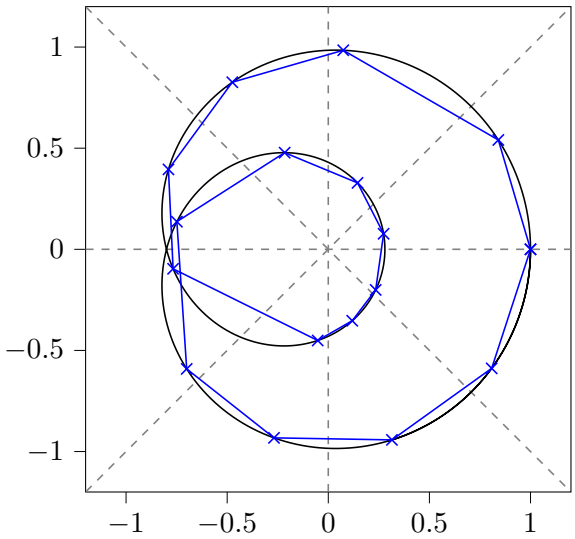
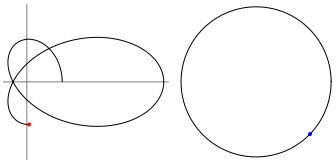


FIGURE 7.5 – Exemple de raffinement de la discrétisation.



Dans la Figure 7.5, tous les segments de la discrétisation possèdent leurs extrémités dans le même secteur ou dans deux secteurs voisins. De plus, l'indice de la courbe noire et l'indice de la ligne polygonale bleue sont tous les deux égaux à 2.

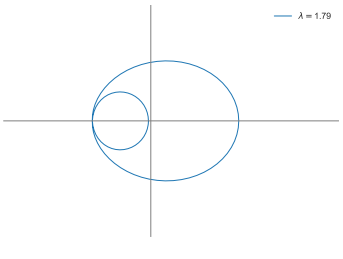
Dans le Chapitre 6, la Figure 6.6 (page 164) montre l'efficacité du raffinement en comparaison avec une discrétisation uniforme. Pour vérifier que les points de la discrétisation sont dans des secteurs angulaires voisins, on utilise les fonctions auxiliaires suivantes pour savoir dans quel secteur se trouve un point de la discrétisation.

```
def sector(z):
    """
    return the sector of z, it is an integer "a" between 0 and 7 which
    correspond to the angular sector [a*pi/4, (a+1)*pi/4[
    Warning : the argument is between -pi and pi and here "a" is between 0 and 7
    """
    ph = phase(z) * 4 / pi
    if ph < 0:
        return int(ph) + 7
    else:
        return int(ph)


def neighboor(sector1, z2):
    """
    sector1 is an integer between 0 and 7 and represent the angular sector [
    sector1*pi/4, (sector1+1)*pi/4[
    z2 is a complex number
    return True iff the complex z2 is in a neighboring sector of [sector1*pi/4,
    (sector1+1)*pi/4[
    """
    return abs((sector(z2) - sector1) % 8) <= 1
```

Code Python 7.7 – Calcul du secteur angulaire d'un complexe.

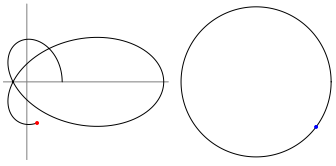
Ensuite, si un point de la discrétisation n'est pas dans un secteur angulaire voisin du point précédent, on raffine davantage la discrétisation en utilisant la fonction du Code Python 7.8.



```python
def parametrization(n_param, curve_formula):
    """
    n_param is an integer for the default discretization of [0,2pi]
    curve_formula is a fonction : z in the unit circle mapsto a complex and
    represent a curve
    return the discretization of the curve refine if it is needed (as [
    ZapataMartin2014] procedure).
    """
    dx = 2 * pi / n_param
    current_dx = dx
    current_param = 0
    Param = [current_param]
    curve = [curve_formula(1)]
    current_sector = sector(curve_formula(np.exp(1j * Param[0])))#get the sector
     of the first point
    c = 0
    s = 0
    while Param[-1] < 2 * pi:
        current_param = Param[-1] + current_dx
        current_curve_point = curve_formula(np.exp(1j * current_param))
        if neighboor(current_sector,current_curve_point):
            Param.append(current_param)
            curve.append(current_curve_point)
            current_sector = sector(current_curve_point)
            current_dx = dx
            c = 0
        elif c<40:
            current_dx = current_dx / 2
            c += 1
        else:
            current_param = Param[-1] + dx
            Param.append(current_param)
            curve.append(curve_formula(np.exp(1j * current_param)))
            current_sector = sector(curve[-1])
            current_dx = dx
            c = 0
        s += 1
    if Param[-1]>2*pi:
        Param[-1] = 0
        curve[-1] = curve_formula(1)
    return np.array(Param), np.array(curve)
```

Code Python 7.8 – Raffinement de la discrétisation de la courbe.

Malgré ce raffinement de discrétisation, l'indice complexe de la courbe peut ne pas être calculé correctement, comme on le voit dans la Figure 7.6 suivante.



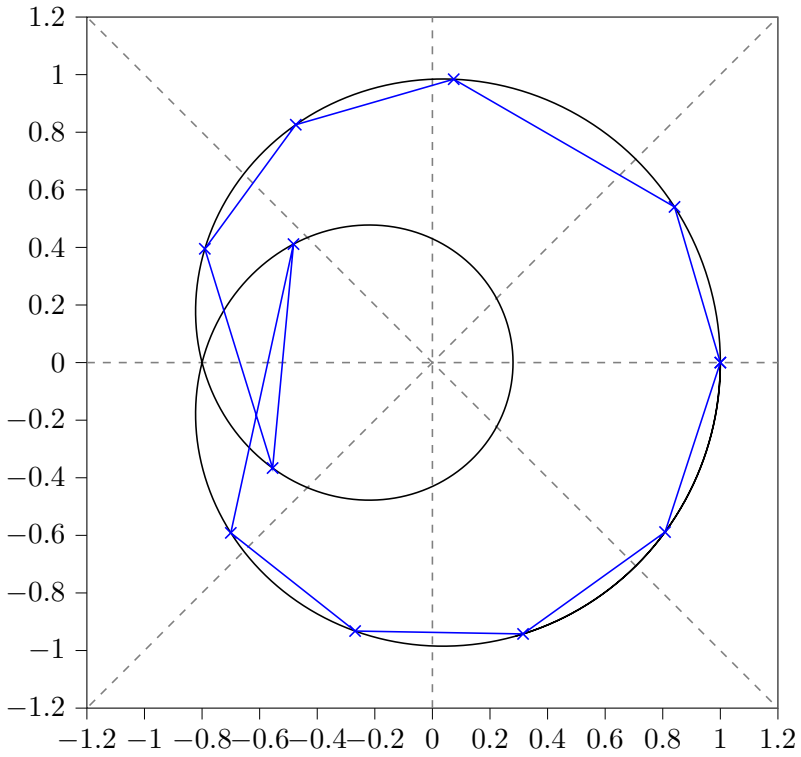


FIGURE 7.6 – Exemple avec raffinement de calcul incorrect de l'indice complexe.

Dans la Figure 7.6, tous les segments de la discrétisation possèdent leurs extrémités dans le même secteur ou dans deux secteurs voisins. Cependant, l'indice de la courbe noire est 2 et l'indice de la ligne polygonale bleue est 1.

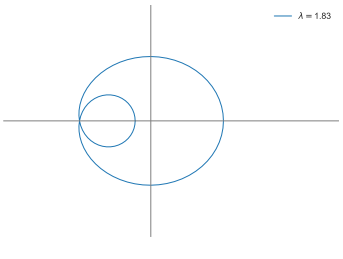
7.2.3 Régularité du déterminant intrinsèque de Kreiss–Lopatinskii pour un schéma totalement décentré

Les articles [GZDM12, Lem.3] et [GZDM14, Fig.6] donnent un critère supplémentaire pour être sûr d'obtenir un calcul exact de l'indice complexe. Pour cela, en supposant que la courbe est Lipschitz de constante L, il suffit de vérifier que pour chaque segment de la discrétisation, on a $|\theta_j - \theta_{j+1}| < \frac{|\Delta(e^{i\theta_j})|+|\Delta(e^{i\theta_{j+1}})|}{L}$, sinon on raffine davantage la discrétisation.

Dans le cas d'un schéma totalement décentré (cadre du Chapitre 5), grâce à la formule explicite (5.15) du déterminant intrinsèque de Kreiss–Lopatinskii, on peut démontrer que la courbe $\Delta(\mathbb{S})$ est lipschitzienne, autrement dit qu'il existe une constante L tel que

$$\forall \theta_1, \theta_2 \in [0, 2\pi], \quad |\Delta(e^{i\theta_1}) - \Delta(e^{i\theta_2})| \leqslant L|\theta_1 - \theta_2|.$$

Démonstration. On rappelle la formule explicite (5.15) du déterminant intrinsèque de Kreiss–



Lopatinskii :

$$\forall z \in \mathbb{S}, \quad \Delta(z) = (-1)^{r(m-r)} \det C(z) \left(\frac{a_{-r}}{a_0 - z} \right)^{m-r}.$$

Soit la fonction $\varphi : \theta \in [0, 2\pi] \mapsto \dfrac{P(e^{i\theta})}{(e^{i\theta} - a_0)^{\alpha}}$ où P est un polynôme et $\alpha \in \mathbb{N}$. La fonction $\theta \mapsto \Delta(e^{i\theta})$ peut s'écrire sous la même forme que φ. Montrons le caractère lipschitzien de φ. Pour tout $\theta \in [0, 2\pi]$, on a

$$\varphi'(\theta) = \frac{ie^{i\theta}(P'(e^{i\theta})(e^{i\theta} - a_0) - \alpha P(e^{i\theta}))}{(e^{i\theta} - a_0)^{\alpha+1}}.$$

Comme $|e^{i\theta} - a_0| \geqslant 1 - |a_0|$ puisque $|a_0| < 1$ sous de bonnes hypothèses (voir Lemme 5.27), on a

$$|\varphi'(\theta)| \leqslant \frac{|P'(e^{i\theta})(e^{i\theta} - a_0) - \alpha P(e^{i\theta})|}{(1 - |a_0|)^{\alpha+1}}.$$

En écrivant le polynôme P comme $\displaystyle\sum_{j=0}^{N} a_j X^j$, on a

$$|\varphi'(\theta)| \leqslant \underbrace{\frac{(\sum_{j=1}^{N} |a_j| j)(1 + |a_0|) + \alpha \sum_{j=0}^{N} |a_j|}{(1 - |a_0|)^{\alpha+1}}}_{L}.$$

Ainsi, par inégalité des accroissements finis, on a

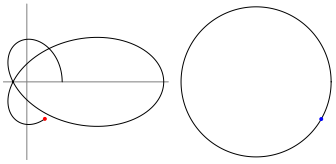
$$\forall \theta_1, \theta_2, \quad |\varphi(\theta_1) - \varphi(\theta_2)| \leqslant L|\theta_1 - \theta_2|.$$

Donc $\Delta(\mathbb{S})$ est une courbe lipschitzienne. □

En pratique, dans les figures du déterminant intrinsèque de Kreiss–Lopatinskii, cette condition n'a pas été implémentée, mais cela serait faisable en calculant explicitement la constante L.

7.3 Classe du bord

On a choisi de représenter indépendamment les bords et les schémas intérieurs. On a ainsi défini des classes Python liées aux conditions de bord. On représente dans la Figure 7.7 une classe par un rectangle contenant trois lignes : une première ligne pour son nom, une deuxième ligne pour ses attributs (variables utilisées dans la classe) et une troisième ligne pour indiquer ses méthodes (fonctions pour manipuler la classe). On représente par des flèches l'héritage des classes, autrement dit, on fait pointer les sous-classes vers leur classe mère. On ne fait apparaître que les nouveaux attributs et les nouvelles méthodes dans les classes héritées. En Figure 7.7, on représente les classes utilisées pour le bord.



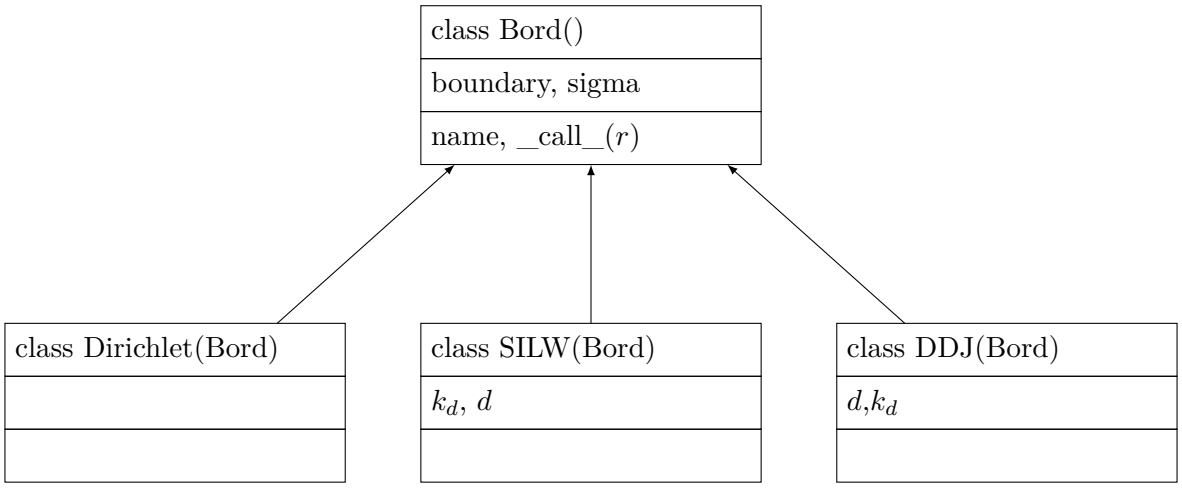


FIGURE 7.7 – Architecture des classes Python liées au traitement du bord numérique.

La méthode `__call__(r)` renvoie la matrice de bord B qui possède m lignes et r colonnes. L'entier r étant lié à l'équation intérieure et non au bord, r ne peut pas être un attribut de la classe Bord. En pratique, on utilise le code suivant pour obtenir la matrice B de bord.

```
kd = 2
d = 3
r = 2

boundary = SILW(kd,d)

B = boundary(r)
```

Code Python 7.9 – Exemple d'exécution pour définir la matrice de bord S2ILW3.

7.3.1 Bord SILW

La classe de bord SILW permet d'implémenter les conditions de bord Lax-Wendroff inverse simplifiée, étudiée en Section 1.5.2 (page 39). Pour Sk_dILWd, on doit construire la matrice de bord

$$B = \begin{pmatrix} \sum_{k=k_d}^{d-1} (-r+\sigma)^k p_{k,0}^{(d)} & \cdots & \sum_{k=k_d}^{d-1} (-r+\sigma)^k p_{k,d-1}^{(d)} \\ \vdots & & \vdots \\ \sum_{k=k_d}^{d-1} (-1+\sigma)^k p_{k,0}^{(d)} & \cdots & \sum_{k=k_d}^{d-1} (-1+\sigma)^k p_{k,d-1}^{(d)} \end{pmatrix} \tag{7.1}$$

avec les coefficients $p_{k,s}^{(d)}$ définis par le système (1.21) de la page 41.

On commence par construire les matrices du système (1.21) qui utilisent les coefficients introduits en Annexe E (page 217).

```python
def deriv(d):
    M = np.zeros((d,d), dtype = int)
    M[0,:] = np.ones(d)
    for i in range (1,d):
        for j in range (i,d+1):
            M[i,j-1] = M[i-1,j-1]*(j-i)
    return M

def coeff(d):
    M = np.zeros((d,d))
    M[0,0] = 1
    for n in range (1,d):
        for l in range (1,n+1):
            if l == 1 or l == n+1:
                M[n,l] = 1
            else:
                M[n,l] = M[n-1,l-1] + l * M[n-1,l]
    return M
```

Code Python 7.10 – Construction des matrices du système (1.21).

```python
def __call__(self, r,**kwargs):
    sigma = kwargs.get("sigma",0)
    B = np.zeros((r, self.d))
    for j in range(r):
        vec_x = np.zeros(self.d)
        vec_x[self.kd:] = (-(j+1)+sigma)**np.arange(self.kd,self.d)
        coeff_d = coeff(self.d)
        deriv_d = deriv(self.d)
        coeff_pk = Upper(deriv_d,Lower(coeff_d,vec_x))
        B[r-1-j,:] = coeff_pk
    return B
```

Code Python 7.11 – Calcul de la matrice B défini par (7.1).

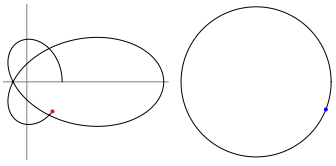
La ligne 2 du Code Python 7.11 permet de donner un argument optionnel σ et de mettre 0 si σ n'est pas défini.

Exemple 7.2. L'exécution du Code Python 7.9 donne la matrice suivante

$$B = \begin{pmatrix} 2 & -4 & 2 \\ 0.5 & -1 & 0.5 \end{pmatrix}.$$

Cette matrice correspond bien au bord décrit dans l'équation (5.25).

En remplaçant la ligne 5x du Code Python 7.9 par `boundary(r, sigma = 0.2)`, on obtient



la matrice suivante

$$B = \begin{pmatrix} 1.62 & -3.24 & 1.62 \\ 0.32 & -0.64 & 0.32 \end{pmatrix}.$$

7.3.2 Bord DDJ

La classe de bord DDJ permet d'implémenter la méthode de reconstruction décrite en Section 6.4.3 (page 159). Le nom DDJ vient de l'article Dakin, Després et Jaouen [DDJ18] qui utilise aussi cette condition de bord.

L'implémentation Python construit les matrices $\mathcal{Y}_-$ et $\mathcal{Y}_+$ de la Section 6.4.3 afin de calculer la matrice B.

7.4 Classe du schéma complet

Maintenant, on définit les classes Python liées aux schémas intérieurs. De la même façon que pour les classes liées au bord, on utilise des rectangles et des flèches pour représenter les classes et l'héritage entre les classes. En Figure 7.8, on représente les classes utilisées pour définir les schémas.

On a défini une classe `SchemeP0` pour traiter les schémas totalement décentrés comme celui de Beam-Warming et upwind. Dans cette sous-classe, on a redéfini le déterminant de Kreiss–Lopatinskii car il n'y a plus besoin de sélectionner les racines stables vu que toutes les racines de l'équation caractéristique (3.4) sont stables.

Un schéma intérieur est représenté, à la façon de Beam et Warming [BW93], par les deux attributs `self.int` et `self.center`. Le premier est une liste contenant les coefficients $a_{-r}, \dots, a_0, \dots, a_p$ et le deuxième permet de définir où est le coefficient a_0.

Exemple 7.3. Le schéma jouet, utilisé dans le Chapitre 2, défini par

$$U_j^{n+1} = -\frac{1}{3}U_{j-1}^n - \frac{1}{2}U_j^n + U_{j+1}^n - \frac{1}{6}U_{j+2}^n,$$

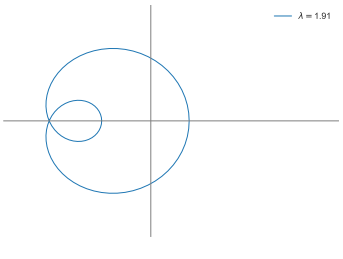
est représenté par

```
self.int = np.array([-1/3, -1/2, 1, -1/6])
self.center = 1
```

Les attributs `boundary` et `boundary_quasitoep` contiennent les matrices B et $\mathcal{B}$ qui représentent le bord.

La méthode `toep(J)` construit la matrice quasi-Toeplitz de taille J associée au schéma. La méthode `symbol(n)` donne une discrétisation de la courbe du symbole du schéma contenant n points, cela permet de tracer la Figure 1.4 (page 33). Les méthodes suivantes ont déjà été présentées en Section 7.1.

La classe `LaxWendroff` correspond au schéma (LW), défini à la page 24, la classe `ThirdOrder` correspond au schéma (O3), la classe `LW` correspond aux schémas Lax-Wendroff présentées



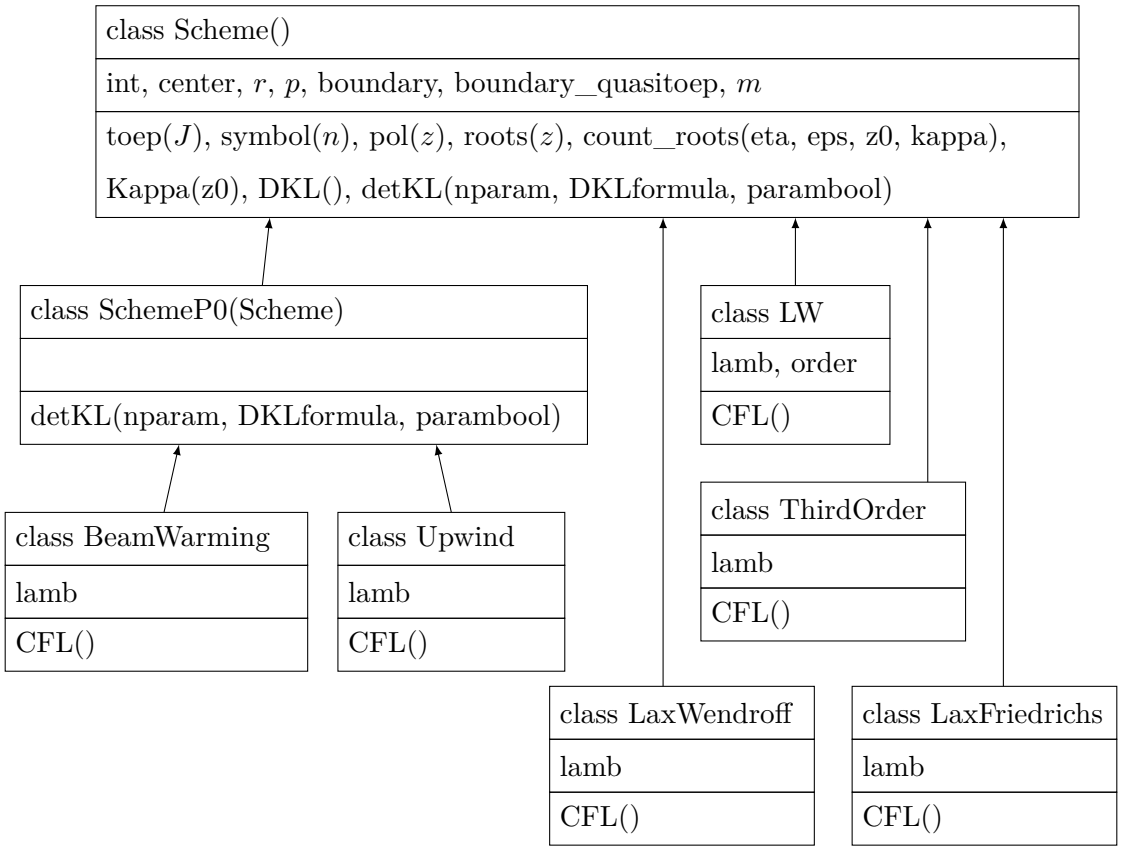


FIGURE 7.8 – Architecture des classes Python liées aux schémas.

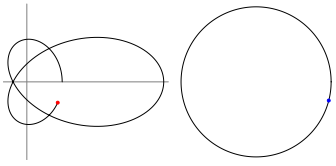
dans [LM06] et, par exemple, le schéma (LW5) est obtenu en prenant `order = 5`. La classe `LaxFriedrichs` correspond au schéma de Lax-Friedrichs défini par

$$U_j^{n+1} = \frac{1-\lambda}{2} U_{j+1}^n + \frac{1+\lambda}{2} U_{j-1}^n.$$

```python
class LaxFriedrichs(Scheme):
    def __init__(self,lamb, boundary = Dirichlet(),**kwargs):
        self.sigma = kwargs.get("sigma",0)
        self.lamb = lamb
        self.int = np.array([(1+lamb)/2,0,(1-lamb)/2])
        self.center = 1
        super().__init__(int=self.int,center=self.center, boundary = boundary,**kwargs)

    def CFL(self):
        return 1
```

Code Python 7.12 – Définition de la classe du schéma Lax-Friedrichs



On donne dans la suite des simulations numériques pour le schéma Beam-Warming avec des conditions de bord S2ILW3, pour faire le lien entre le spectre de la matrice quasi-Toeplitz et la courbe du déterminant intrinsèque de Kreiss–Lopatinskii.

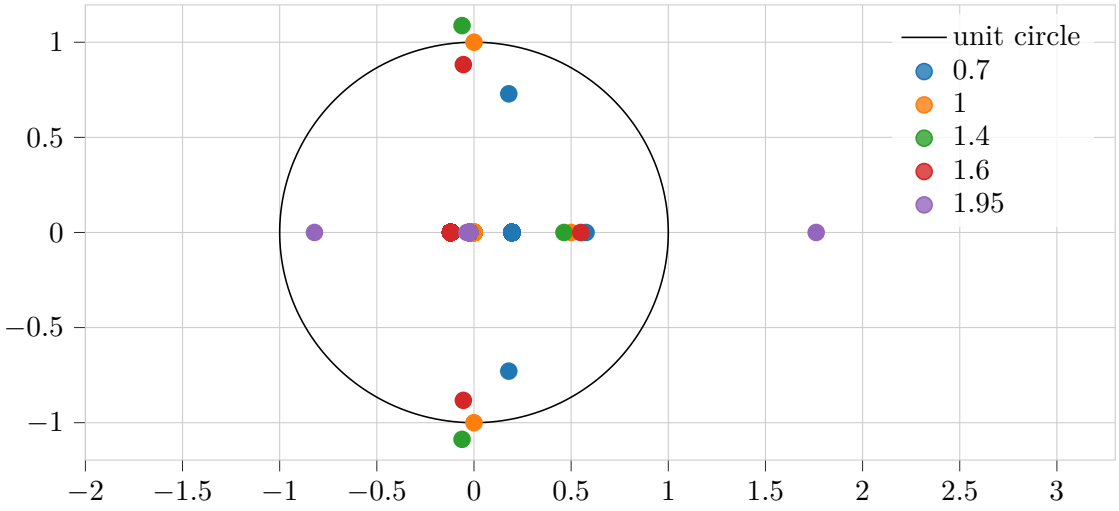


FIGURE 7.9 – Spectre de la matrice quasi-Toeplitz de taille $J = 100$ pour le schéma Beam-Warming muni des conditions de bord S2ILW3 pour différents λ.

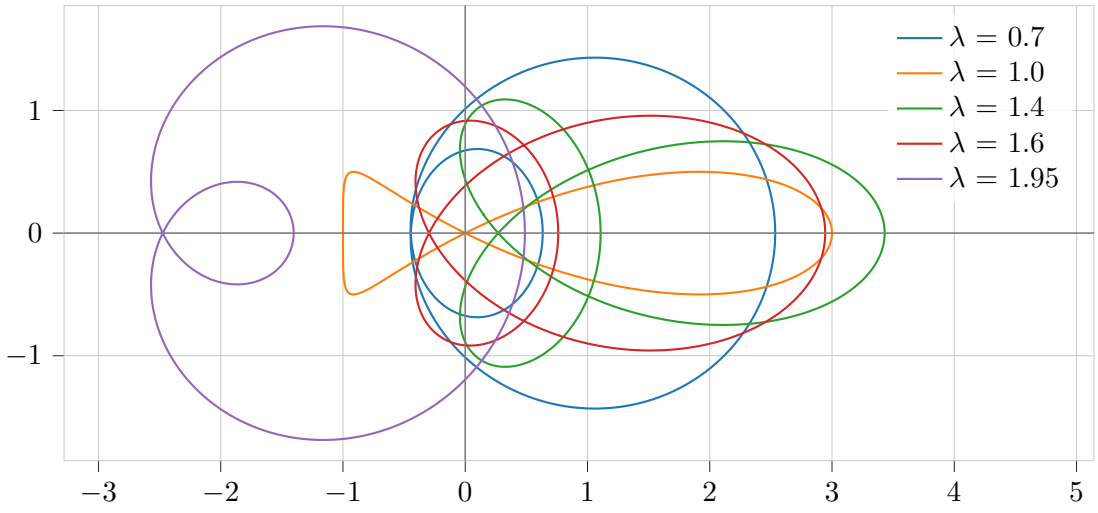
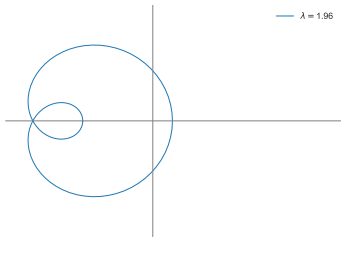


FIGURE 7.10 – Courbes du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma Beam-Warming muni des conditions de bord S2ILW3 pour différents λ.

Les Figures 7.9 et 7.10 sont à lire en parallèle. En effet, par exemple, en prenant $\lambda = 1.4$, on voit qu'il y a des valeurs propres de la matrice quasi-Toeplitz qui sont à l'extérieur du disque unité et on voit que la courbe du déterminant intrinsèque de Kreiss–Lopatinskii ne fait aucun tour autour de l'origine, donc le déterminant s'annule deux fois dans $\mathcal{U}$ (en utilisant le Corollaire 5.15). Alors que pour $\lambda = 1.6$, toutes les valeurs propres sont à l'intérieur du disque



unité et la courbe du déterminant fait deux tours autour de l'origine, donc le déterminant ne s'annule pas ni dans $\mathcal{U}$, si sur $\mathbb{S}$ (puisque la courbe ne passe pas par l'origine).

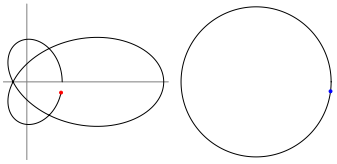
7.5 Explication des folioscopes

On présente ici les légendes et explications des *folioscopes* (ou *flipbook* en anglais) présents en bas de page du manuscrit :

- Le folioscope de droite entre les pages 21 et 94 représente la courbe du symbole du schéma Beam-Warming (BW) pour λ variant entre 0 et 2 pour respecter la CFL de la Proposition 1.15. En noir est représenté le cercle unité et en bleu est représenté la courbe Γ du symbole.
- Le folioscope de gauche entre les pages 21 et 94 représente le spectre de la matrice quasi-Toeplitz associée au schéma de l'Exemple 2.22, on fait varier la condition de bord entre celle de T_J et celle de $\widetilde{T_J}$ pour les matrices suivantes :

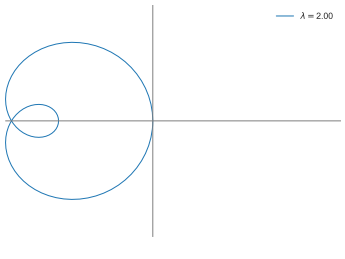
$$T_J = \begin{pmatrix} -\frac{1}{2} & 1 & -\frac{1}{6} & & & 0 \\ -\frac{1}{3} & -\frac{1}{2} & 1 & -\frac{1}{6} & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & -\frac{1}{6} \\ & & & \ddots & \ddots & 1 \\ 0 & & & & -\frac{1}{3} & -\frac{1}{2} \end{pmatrix} \text{ et } \widetilde{T_J} = \begin{pmatrix} -2.7 & 5.6 & -4.1 & 1.2 & & 0 \\ -\frac{1}{3} & -\frac{1}{2} & 1 & -\frac{1}{6} & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & -\frac{1}{6} \\ & & & -\frac{1}{3} & -\frac{1}{2} & 1 \\ 0 & & & 0.8 & -2.9 & 2.4 \end{pmatrix}.$$

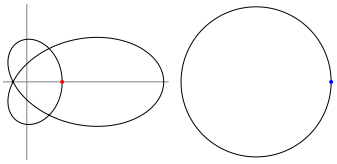
  On voit alors les valeurs propres isolées du spectre de $\widetilde{T_J}$ dont on parle à la Proposition 2.20 (page 54) sortir du spectre de la matrice T_J. La légende de la figure est la même que celle de la Figure 2.5 (page 56).
- Le folioscope de droite entre les pages 101 et 196 représente la courbe $\Delta(\mathbb{S})$ du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma de Beam-Warming muni de S2ILW3 (défini en (1.22), page 42) en faisant évoluer λ entre 0 et 2.
- Le folioscope de gauche entre les pages 101 et 196 représente le tracé de la courbe $\Delta(\mathbb{S})$ du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma de Beam-Warming muni de S2ILW3 pour $\lambda = 1.6$ (à gauche) tout en faisant parcourir à z le cercle unité $\mathbb{S}$ (à droite).
- Le folioscope de droite entre les pages 199 et 230 représente en haut la solution du schéma de Beam-Warming muni de S2ILW3 après 100 itérations temporelles avec les données $f(x) = 0$ pour tout $x \in [0, 1]$ et $g(t) = t^2 e^{-t}$ pour tout $t \geqslant 0$. On représente en bas le nombre de zéros du déterminant intrinsèque de Kreiss–Lopatinskii dans $\mathcal{U}$ en faisant évoluer λ entre 0 et 2. La figure du bas est similaire aux Figure 4.3 et Figure 5.5 (page 98 et page 130). Les deux figures permettent de mettre en relation la stabilité du schéma et



le nombre zéros du déterminant intrinsèque de Kreiss–Lopatinskii comme cela est fait au Corollaire 3.24 et aux Corollaires 5.15 et 6.14.

- Le folioscope de gauche entre les pages 199 et 230 représente le tracé de la courbe $\Delta(\mathbb{S})$ du déterminant intrinsèque de Kreiss–Lopatinskii pour le schéma leap-frog muni de la condition de bord (α) définie en Annexe D (à gauche) tout en faisant parcourir à z le cercle unité $\mathbb{S}$ (à droite). On observe que la courbe de $\Delta(\mathbb{S})$ fait bien deux tours autour de l'origine comme évoqué page 175.





Troisième partie

Annexes

Annexe A

Quelques éléments d'analyse complexe

On donne dans ce chapitre quelques résultats d'analyse complexe. On rappelle quelques résultats classiques que l'on ne démontrera pas (théorème des résidus ...), on peut par exemple en trouver des preuves dans les livres [Lan99], [Rud11], etc. On donne des preuves lorsque l'énoncé n'est pas usuel (théorème des résidus adapté ...) ou si le résultat est utilisé à plusieurs reprises dans le manuscrit (principe de l'argument, théorème de Rouché ...).

A.1 Notations

On rappelle que l'indice d'un point $w \in \mathbb{C}$ par rapport à une courbe fermée Γ, noté $\mathrm{Ind}_\Gamma(w)$, vaut

$$\mathrm{Ind}_\Gamma(w) = \frac{1}{2i\pi} \int_\Gamma \frac{dz}{z-w}.$$

De plus, on rappelle que le résidu d'un point $w \in \mathbb{C}$ par rapport à une fonction f holomorphe sur $\Omega \setminus \{w\}$ et méromorphe sur Ω où Ω est un ouvert simplement connexe peut être vu comme le coefficient a_{-1} de la série de Laurent de f en w, *i.e.*

$$f(z) = \sum_{n=-N}^{+\infty} a_n (z-w)^n, \quad \text{pour } z \text{ suffisamment proche de } w.$$

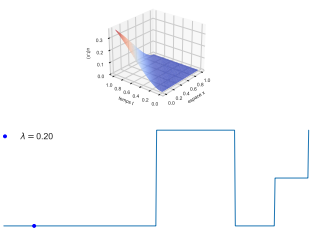
A.2 Théorème des résidus

Théorème A.1 (des résidus)**.** *Soit un ouvert simplement connexe $\Omega \subset \mathbb{C}$, soit $w_1, \dots, w_n \in \Omega$, soit f une fonction holomophe sur $\Omega \setminus \{w_1, \dots, w_n\}$ et méromorphe sur Ω et soit Γ une courbe fermée de $\Omega \setminus \{w_1, \dots, w_n\}$. On a alors*

$$\frac{1}{2i\pi} \int_\Gamma f(z)dz = \sum_{j=1}^{n} \mathrm{Ind}_\Gamma(w_j)\mathrm{Res}_f(w_j).$$

On admet ce résultat usuel.

Théorème A.2 (des résidus adapté)**.** *Soit un ouvert simplement connexe $\Omega \subset \mathbb{C}$, soit $n \in \mathbb{N}$, soit $w_1, \dots, w_n \in \Omega$. Soit $a \in \mathbb{C}$ et $r > 0$ tels que $B(a,r) \subset \overline{\Omega}$. Si f est une fonction holomophe*



sur $\Omega \setminus \{w_1, \dots, w_n\}$*, méromorphe sur* Ω *et continue sur* $\overline{\Omega} \setminus \{w_1, \dots, w_n\}$ *et si* $\Gamma : \theta \in [0, 2\pi] \mapsto a + re^{i\theta}$ *est un contour circulaire ne passant par aucun élément de* $\{w_1, \dots, w_n\}$*, alors*

$$\frac{1}{2i\pi} \int_\Gamma f(z)dz = \sum_{j=1}^{n} \mathrm{Ind}_\Gamma(w_j)\mathrm{Res}_f(w_j).$$

Démonstration. Si le contour Γ est dans Ω, on peut utiliser le Théorème A.1 des résidus classique car le contour est dans le domaine d'holomorphie. Sinon, on va utiliser un contour un peu plus petit pour être dans le domaine d'holomorphie et en passant à la limite, par le théorème de convergence dominée, on aura le résultat voulu.

Soit $\varepsilon \in \left]0, \min_{j=1}^{n} ||w_j - a| - r|\right[$, la borne de droite étant non nulle car, pour tout $j \in [\![1 : n]\!]$, on a $w_j \notin \Gamma$, on introduit la courbe $\Gamma_\varepsilon : \theta \mapsto a + (r - \varepsilon)e^{i\theta}$. On peut maintenant appliquer le Théorème A.1 des résidus classique car $\Gamma_\varepsilon \subset \Omega$. On a donc

$$\frac{1}{2i\pi} \int_{\Gamma_\varepsilon} f(z)dz = \sum_{j=1}^{n} \mathrm{Ind}_{\Gamma_\varepsilon}(w_j)\mathrm{Res}_f(w_j).$$

Le réel ε a été choisi de sorte qu'il n'y a pas d'élément de $\{w_1, \dots, w_n\}$ dans $\overline{B(a,r)} \setminus B(a, r-\varepsilon)$, ainsi, on a $\mathrm{Ind}_{\Gamma_\varepsilon}(w_j) = \mathrm{Ind}_\Gamma(w_j)$ pour tout $j \in [\![1 : n]\!]$. De plus, par convergence dominée, on a

$$\int_{\Gamma_\varepsilon} f(z)dz = \int_0^{2\pi} i(a + (r-\varepsilon)e^{i\theta}) f(a + (r-\varepsilon)e^{i\theta})d\theta \xrightarrow[\varepsilon \to 0]{} \int_0^{2\pi} i(a + re^{i\theta}) f(a + re^{i\theta})d\theta = \int_\Gamma f(z)dz$$

car $|i(a + (r-\varepsilon)e^{i\theta}) f(a + (r-\varepsilon)e^{i\theta})| \leqslant (|a| + r)\|f\|_\infty$ intégrable sur $[0, 2\pi]$. Ce qui donne le résultat voulu. □

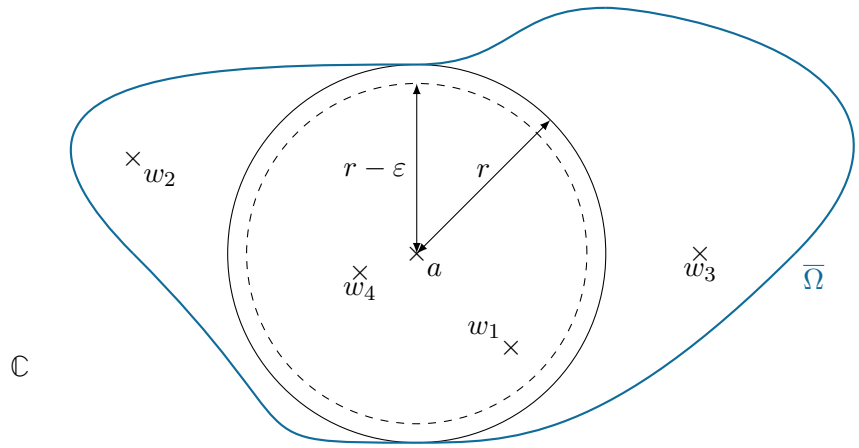


FIGURE A.1 – Illustration du Théorème A.2

Théorème A.3 (Principe de l'argument)**.** *Soit* Ω *un ouvert simplement connexe de* $\mathbb{C}$*, soit* f *une fonction méromorphe sur* Ω *et continue sur* $\overline{\Omega} \setminus F$ *où* F *est l'ensemble des zéros et des pôles de* f *dans* $\overline{\Omega}$ *et est supposé fini. Soit* $a \in \mathbb{C}$ *et* $r > 0$ *tels que* $B(a, r) \subset \overline{\Omega}$*. On suppose que le*

contour $\Gamma : \theta \in [0, 2\pi] \mapsto a + re^{i\theta}$ *ne passe pas par des éléments de* F*. On a alors*

$$\mathrm{Ind}_{f(\Gamma)}(0) = \#z\acute{e}ros_f(B(a,r)) - \#p\hat{o}les_f(B(a,r)) \tag{A.1}$$

comptés avec multiplicité, où $\#z\acute{e}ros_f(B(a,r))$ *est le nombre de zéros de* f *dans* $B(a,r)$ *et* $\#p\hat{o}les_f(B(a,r))$ *son nombre de pôles dans* $B(a,r)$*.*

Démonstration du Théorème A.3. Par définition de l'indice complexe, on a

$$\mathrm{Ind}_{f(\Gamma)}(0) = \frac{1}{2i\pi}\int_{f(\Gamma)} \frac{d\zeta}{\zeta - 0} = \frac{1}{2i\pi}\int_{\Gamma} \frac{f'(z)dz}{f(z)} \tag{A.2}$$

en intégrant suivant le contour $f : \Gamma \to f(\Gamma)$.

On remarque que l'ensemble fini F des zéros et des pôles de f dans $\overline{\Omega}$ correspond à l'ensemble des pôles de $\frac{f'}{f}$ dans $\overline{\Omega}$. Par le Théorème A.2 (des résidus) appliqué à la fonction $\frac{f'}{f}$ qui est holomorphe sur $\Omega \setminus F$ et continue sur $\overline{\Omega} \setminus F$, on a

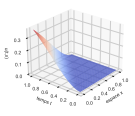
$$\frac{1}{2i\pi}\int_{\Gamma} \frac{f'(z)dz}{f(z)} = \sum_{\zeta \in F \cap B(a,r)} \mathrm{Res}_{\frac{f'}{f}}(\zeta). \tag{A.3}$$

En effet, comme Γ est un lacet simple positivement orienté, on a $\mathrm{Ind}_{\Gamma}(\zeta) = 1$ pour tout $\zeta \in F \cap B(a,r)$ et $\mathrm{Ind}_{\Gamma}(\zeta) = 0$ pour tout $\zeta \in F \setminus B(a,r)$.

- Si ζ est un zéro de f de multiplicité α_ζ. On a alors $f(z) = (z-\zeta)^{\alpha_\zeta} g_\zeta(z)$ où $g_\zeta(z)$ est holomorphe et ne s'annule pas en ζ. D'où $f'(z) = \alpha_\zeta (z-\zeta)^{\alpha_\zeta - 1} g_\zeta(z) + (z-\zeta)^{\alpha_\zeta} g'_\zeta(z)$. Ainsi,
$$\frac{f'(z)}{f(z)} = \frac{\alpha_\zeta}{z-\zeta} + \frac{g'_\zeta(z)}{g_\zeta(z)}.$$
Donc $\mathrm{Res}_{\frac{f'}{f}}(\zeta) = \alpha_\zeta$.
- Si ζ est un pôle de f de multiplicité β_ζ. On a alors $f(z) = \frac{g_\zeta(z)}{(z-\zeta)^{\beta_\zeta}}$ où $g_\zeta(z)$ est holomorphe et ne s'annule pas en ζ. D'où $f'(z) = \frac{g'_\zeta(z)}{(z-\zeta)^{\beta_\zeta}} - \frac{g_\zeta(z)\beta_\zeta(z-\zeta)^{\beta_\zeta - 1}}{(z-\zeta)^{2\beta_\zeta}}$. Ainsi,
$$\frac{f'(z)}{f(z)} = \frac{g'_\zeta(z)}{g_\zeta(z)} - \frac{\beta_\zeta}{z-\zeta}.$$
Donc $\mathrm{Res}_{\frac{f'}{f}}(\zeta) = -\beta_\zeta$.

Ainsi, grâce à (A.2) et (A.3), on obtient (A.1). □

Lemme A.4. *Soit* f *une fonction holomorphe sur* $\mathbb{C} \setminus \mathbb{D}$ *et continue sur* $\overline{\mathbb{C} \setminus \mathbb{D}}$*. On pose la fonction* $g : z \in \overline{\mathbb{D}} \mapsto f(z^{-1})$*. On a alors* $\mathrm{Ind}_{g(\mathbb{S})}(0) = -\mathrm{Ind}_{f(\mathbb{S})}(0)$.



Démonstration. Il suffit de faire le calcul suivant :

$$\begin{aligned}\mathrm{Ind}_{g(\mathbb{S})}(0) &= \frac{1}{2i\pi}\int_{g(\mathbb{S})}\frac{dz}{z} = \frac{1}{2i\pi}\int_0^{2\pi}\frac{ie^{i\theta}g'(e^{i\theta})}{g(e^{i\theta})}d\theta = \frac{1}{2i\pi}\int_0^{2\pi}\frac{-ie^{i\theta}f'(e^{-i\theta})}{e^{2i\theta}f(e^{-i\theta})}d\theta \\ &= \frac{1}{2i\pi}\int_0^{2\pi}\frac{-ie^{-i\theta}f'(e^{-i\theta})}{f(e^{-i\theta})}d\theta = \frac{1}{2i\pi}\int_0^{-2\pi}\frac{ie^{i\theta}f'(e^{i\theta})}{f(e^{i\theta})}d\theta \\ &= -\frac{1}{2i\pi}\int_{-2\pi}^{0}\frac{ie^{i\theta}f'(e^{i\theta})}{f(e^{i\theta})}d\theta = -\frac{1}{2i\pi}\int_0^{2\pi}\frac{ie^{i\theta}f'(e^{i\theta})}{f(e^{i\theta})}d\theta = -\mathrm{Ind}_{f(\mathbb{S})}(0).\end{aligned}$$

□

A.3 Théorème de Rouché

Théorème A.5 (Rouché)**.** *Soit Ω un ouvert simplement connexe de $\mathbb{C}$. Soient deux fonctions f et g holomorphes sur Ω et continues sur $\overline{\Omega}$. On note $F \subset \overline{\Omega}$ l'ensemble des zéros de f et de g dans $\overline{\Omega}$ et on suppose que F est fini. Soit $a \in \mathbb{C}$ et $r > 0$ tels que $B(a,r) \subset \overline{\Omega}$. On suppose que la courbe $\Gamma : \theta \in [0, 2\pi] \mapsto a + re^{i\theta}$ ne passe pas par des éléments de F.*

Si, pour tout $\kappa \in \Gamma$, on a

$$|f(\kappa) - g(\kappa)| < |g(\kappa)|,$$

alors le nombre de zéros de f dans $B(a,r)$ est égal au nombre de zéros de g dans $B(a,r)$ (comptés avec multiplicité).

Démonstration. On pose la fonction $h = \frac{f}{g}$ qui est donc méromorphe sur Ω et continue sur $\overline{\Omega}\setminus F$. On a alors, pour tout $\kappa \in \Gamma$,

$$|h(\kappa) - 1| < 1.$$

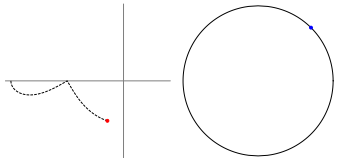
Autrement dit, $h(\Gamma) \subset B(1,1)$ et donc pour tout $\zeta \notin B(1,1)$, on a $\mathrm{Ind}_{h(\Gamma)}(\zeta) = 0$. En particulier, on a $\mathrm{Ind}_{h(\Gamma)}(0) = 0$ et par le principe de l'argument (Théorème A.3), on a

$$\mathrm{Ind}_{h(\Gamma)}(0) = 0 = \#\text{zéros}_h(B(a,r)) - \#\text{pôles}_h(B(a,r)),$$

comptés avec multiplicités. Or, par définition de h, les zéros de h dans $B(a,r)$ sont exactement les zéros de f dans $B(a,r)$ et les pôles de h dans $B(a,r)$ sont exactement les zéros de g dans $B(a,r)$. Ce qui permet d'obtenir

$$\#\text{zéros}_f(B(a,r)) = \#\text{zéros}_g(B(a,r)),$$

comptés avec multiplicités. □



Annexe B

Quelques éléments sur la transformée en $\mathcal{Z}$

On donne dans ce chapitre quelques résultats sur la transformée en $\mathcal{Z}$. Cette transformée est analogue à la transformée de Laplace :

$$\mathcal{L}(x)(p) = \int_0^{+\infty} e^{-pt} x(t) dt.$$

Comme on va le voir, la transformée en $\mathcal{Z}$ est le cas discret de la tranformée de Laplace. En effet, la fonction x devient une suite $(x_n)_n$ et on pose $z = e^{-p}$.

B.1 Définitions et formule d'inversion

Definition B.1 (Transformée en $\mathcal{Z}$)**.** Soit une suite $(x_n)_n \in \ell^2(\mathbb{N}, \mathbb{R})$. On définit la transformée en $\mathcal{Z}$ de $(x_n)_n$, notée $\widetilde{x}$, par

$$\forall z \in \mathcal{U}, \quad \widetilde{x}(z) = \sum_{n \geqslant 0} x_n z^{-n}.$$

Proposition B.2 (Transformée en $\mathcal{Z}$ inverse)**.** *Soit $(x_n)_n \in \ell^2(\mathbb{N})$. On définit la transformée en $\mathcal{Z}$ inverse de la manière suivante : pour $R > 1$, pour tout $n \in \mathbb{N}$,*

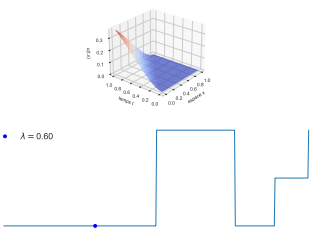
$$x_n = \frac{1}{2i\pi} \int_{\mathcal{C}(0,R)} z^{n-1} \widetilde{x}(z) dz = \frac{1}{2\pi} \int_0^{2\pi} \widetilde{x}(Re^{i\theta}) R^n e^{in\theta} d\theta.$$

Démonstration. Fixons $R > 1$ et $n \in \mathbb{N}$. On parcourt la courbe fermée $\mathcal{C}(0, R)$ par le chemin $\Gamma : \theta \mapsto Re^{i\theta}$ avec $\Gamma'(\theta) = Rie^{i\theta}$. On a alors

$$\begin{aligned} \frac{1}{2i\pi} \int_{\mathcal{C}(0,R)} z^{n-1} \widetilde{x}(z) dz &= \frac{1}{2\pi} \int_0^{2\pi} R^{n-1} e^{i\theta(n-1)} \widetilde{x}(Re^{i\theta}) Re^{i\theta} d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} R^n e^{in\theta} \sum_{k=0}^{+\infty} x_k R^{-k} e^{-ik\theta} d\theta. \end{aligned}$$

Par inversion de la somme et de l'intégrale grâce au théorème de Fubini, on obtient alors

$$\frac{1}{2\pi} \sum_{k=0}^{+\infty} x_k \int_0^{2\pi} R^{n-k} e^{i(n-k)\theta} d\theta \;=\; \frac{1}{2\pi} \sum_{k=0}^{+\infty} x_k R^{n-k} \int_0^{2\pi} e^{i(n-k)\theta} d\theta \;=\; \sum_{k=0}^{+\infty} x_k R^{n-k} \delta(n - k).$$



On trouve donc bien x_n car le seul terme non nul de la somme est pour $k = n$.

Justification de Fubini. Il faut montrer que $(k, \theta) \mapsto x_k R^{n-k} e^{i\theta(n-k)}$ est intégrable sur l'ensemble $\mathbb{N} \times [0, 2\pi]$ pour la mesure produit de la mesure de comptage et la mesure de Lebesgue. Or

$$\left(\sum_{k=0}^{\infty} |x_k| R^{n-k} \right)^2 \leqslant R^{2n} \sum_{k=0}^{+\infty} |x_k|^2 \sum_{k=0}^{\infty} R^{-2k} < +\infty$$

car $R > 1$ et $(x_n)_n \in \ell^2(\mathbb{N})$. Comme cette quantité est bornée, on peut l'intégrer sur $[0, 2\pi]$. Donc Fubini–Tonelli s'applique et on peut utiliser Fubini pour intervertir la somme et l'intégrale. □

B.2 Égalité de Parseval

Proposition B.3 (Égalité de Parseval)**.** *Soient $(x_n)_n \in \ell^2(\mathbb{N})$ et $R > 1$. On a*

$$\sum_{n \geqslant 0} R^{-2n} |x_n|^2 = \frac{1}{2\pi} \int_0^{2\pi} |\widetilde{x}(Re^{i\theta})|^2 d\theta.$$

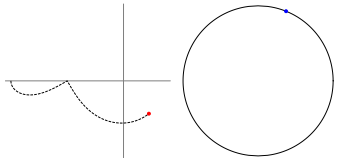
Démonstration. Soit $R > 1$. En utilisant un produit de Cauchy et le théorème de Fubini, on a

$$\begin{aligned}
\frac{1}{2\pi} \int_0^{2\pi} |\widetilde{x}(Re^{i\theta})|^2 d\theta &= \frac{1}{2\pi} \int_0^{2\pi} \left(\widetilde{x}(Re^{i\theta}) \right) \left(\overline{\widetilde{x}(Re^{i\theta})} \right) d\theta \\
&= \frac{1}{2\pi} \int_0^{2\pi} \left(\sum_{n=0}^{+\infty} x_n R^{-n} e^{-in\theta} \right) \left(\sum_{k=0}^{+\infty} \overline{x_k} R^{-k} e^{ik\theta} \right) d\theta \\
&= \frac{1}{2\pi} \int_0^{2\pi} \sum_{n=0}^{+\infty} \sum_{k=0}^{n} x_k \overline{x_{n-k}} R^{-k} R^{-(n-k)} e^{-ik\theta} e^{i(n-k)\theta} d\theta \\
&= \sum_{n=0}^{+\infty} \sum_{k=0}^{n} x_k \overline{x_{n-k}} R^{-n} \delta(n - 2k).
\end{aligned}$$

Ainsi, le symbole de Kronecker $\delta(n - 2k)$ impose que n soit pair, on peut donc écrire $n = 2\ell$.

$$\begin{aligned}
\frac{1}{2\pi} \int_0^{2\pi} |\widetilde{x}(Re^{i\theta})|^2 d\theta &= \sum_{\ell=0}^{+\infty} \sum_{k=0}^{2\ell} x_k \overline{x_{2\ell-k}} R^{-2\ell} \delta(2\ell - 2k) \\
&= \sum_{\ell=0}^{+\infty} x_\ell \overline{x_\ell} R^{-2\ell} = \sum_{\ell=0}^{+\infty} |x_\ell|^2 R^{-2\ell}.
\end{aligned}$$

Justification de Fubini. Même justification que dans la preuve précédente. Comme les deux séries $\sum_{n=0}^{+\infty} x_n R^{-n} e^{-in\theta}$ et $\sum_{k=0}^{+\infty} \overline{x_k} R^{-k} e^{ik\theta}$ sont absolument convergentes, leur produit de Cauchy l'est aussi. Ainsi, comme il est borné, il est intégrable sur $[0, 2\pi]$. Le théorème de Fubini s'applique. □



Quelques éléments de théorie spectrale

Les trois premières sections (Sections C.1, C.2 et C.3) sont un condensé des définitions et notations utiles pour comprendre les résultats de la Section C.4 et de la Section 2.3.3. Pour plus de détails, on peut s'appuyer sur le livre de Cheverry et Raymond [CR21]. On se place dans le cadre d'opérateurs bornés (car tous les opérateurs étudiés sont bornés, voir Proposition C.11), mais la plupart des définitions demeurent valables pour des opérateurs non bornés.

C.1 Quelques notations

Soient E et F deux espaces de Banach. Soit $T : E \to F$ un opérateur linéaire borné.

- l'*image* de T est notée $\mathrm{Ran}(T) = \{Tx \in F, x \in E\} \subset F$.
- le *noyau* de T est noté $\ker(T) = \{x \in E, Tx = 0\} \subset E$.
- la *dimension* d'un espace vectoriel E est notée $\dim E$.
- la *codimension* dans un espace vectoriel E d'un sous-espace vectoriel G est notée $\mathrm{codim}_E(G) \overset{\text{def}}{=} \dim(E/G)$ ou $\mathrm{codim}\,(G)$ si l'espace vectoriel ambiant est clair. Si E est de dimension finie, on a $\mathrm{codim}_E(G) = \dim E - \dim G$.
- on dit que T est *borné* s'il existe $M > 0$ tel que, pour tout $x \in E$, on a

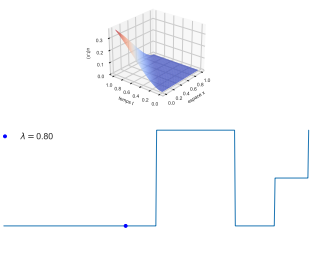
$$\|Tx\|_F \leqslant M\|x\|_E.$$

C.2 Opérateur compact et opérateur de Fredholm

Définition C.1 (Opérateur de Fredholm)**.** Soient E et F deux espaces de Banach. Un opérateur $T : E \to F$ linéaire borné est dit *de Fredholm* si $\dim\ker(T) < \infty$ et $\mathrm{codim}\,\mathrm{Ran}(T) < \infty$. De plus, son *indice de Fredholm* est défini par

$$\mathrm{ind}(T) = \dim\ker(T) - \mathrm{codim}\,\mathrm{Ran}(T).$$

Exemple C.2. Un opérateur $T : E \to F$ linéaire borné bijectif est de Fredholm et est d'indice nul. En effet, son noyau est réduit à $\{0\}$ et son image vaut $\mathrm{Ran}(T) = F$ donc de codimension



nulle.

Exemple C.3. Si E et F sont de dimension finie, alors tout opérateur linéaire borné $T : E \to F$ est de Fredholm et d'indice $\mathrm{ind}(T) = \dim E - \dim F$ (théorème du rang).

Définition C.4 (Opérateur compact)**.** Soient E et F deux espaces de Banach. Un opérateur $T : E \to F$ linéaire borné est dit *compact* si $T(B_E(0,1))$ est relativement compact dans F.

On va voir dans les propositions suivantes le lien entre opérateur compact et opérateur de Fredholm.

Proposition C.5. *Soit E un espace de Banach. Si $K : E \to E$ est un opérateur compact, alors $\mathrm{Id}_E + K$ est un opérateur de Fredholm.*

La preuve peut être trouvée dans le livre [CR21].

Proposition C.6. *Soient E et F deux espaces de Banach. Si $T : E \to F$ est un opérateur de Fredholm et si $K : E \to F$ est un opérateur compact, alors $T + K$ est un opérateur de Fredholm et* $\mathrm{ind}(T + K) = \mathrm{ind}(T)$.

La preuve peut être trouvée dans le livre [CR21].

C.3 Spectres

Soit E et F deux espaces de Banach. Soit $T : E \to F$ un opérateur linéaire borné.

Definition C.7 (Spectres)**.** Il existe plusieurs notions de spectre :

- le *spectre* $\Lambda(T) = \{z \in \mathbb{C},\ T - z \text{ n'est pas bijectif}\}$.
- le *spectre ponctuel* $\Lambda_p(T) = \{z \in \mathbb{C},\ T - z \text{ n'est pas injectif}\}$.
- le *spectre essentiel* $\Lambda_{ess}(T) = \{z \in \mathbb{C},\ T - z \text{ n'est pas Fredholm d'indice } 0\}$.
- le *spectre discret* $\Lambda_{dis}(T) = \{z \in \Lambda(T),\ z \text{ est isolée dans le spectre, de multiplicité algébrique finie et telle que } \mathrm{Ran}(T - z) \text{ est fermé}\}$.

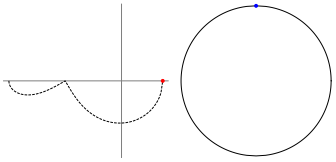
Proposition C.8. *On a les inclusions suivantes :*

- $\Lambda_p(T) \subset \Lambda(T)$.
- $\Lambda_{ess}(T) \subset \Lambda(T)$.
- $\Lambda_{dis}(T) \subset \Lambda_p(T)$.

Théorème C.9. *On a $\Lambda(T) = \Lambda_{ess}(T) \cup \Lambda_p(T)$.*

Démonstration. $\boxed{\supset}$ car $\Lambda_{ess}(T) \subset \Lambda(T)$ et $\Lambda_p(T) \subset \Lambda(T)$.

$\boxed{\subset}$ car, par contraposée, si $z \notin \Lambda_{ess}(T) \cup \Lambda_p(T)$, on a $z \notin \Lambda_{ess}(T)$ donc $T - z$ est Fredholm d'indice 0 et $z \notin \Lambda_p(T)$, donc $\ker(T-z) = \{0\}$. Ainsi $\dim\ker(T-z) = 0$ et comme $\mathrm{ind}(T-z) = 0$, on a $\mathrm{codim}\,\mathrm{Ran}(T - z) = 0$, d'où $T - z$ est bijectif, *i.e.* $z \notin \Lambda(T)$. □



C.4 Régularité des opérateurs Toeplitz et Quasi-Toeplitz

Cette section donne des détails sur les pistes de réflexion envisagées dans la Section 2.3 (page 57) à propos du spectre des opérateurs Toeplitz $T_{\mathbb{N}}$ et Quasi-Toeplitz $\widetilde{T_{\mathbb{N}}}$.

On rappelle ici la Définition 2.10 d'un opérateur Toeplitz $T_{\mathbb{N}}$.

Définition C.10 (Opérateur Toeplitz sur $\mathbb{N}$)**.** On définit l'opérateur Toeplitz sur $\mathbb{N}$ de la manière suivante :

$$T_{\mathbb{N}} : \left\{ \begin{array}{ccc} \ell^2(\mathbb{N}) & \to & \ell^2(\mathbb{N}) \\ u = (u_n)_{n\in\mathbb{N}} & \mapsto & ((T_{\mathbb{N}}u)_n)_{n\in\mathbb{N}}, \end{array} \right. \tag{C.1}$$

avec $\forall n \in \mathbb{N}, \quad (T_{\mathbb{N}}u)_n \overset{\text{def}}{=} \begin{cases} \sum_{j=-r}^{p} a_j u_{n+j} & \text{si } n \geqslant r, \\ \sum_{j=-n}^{p} a_j u_{n+j} & \text{si } n < r. \end{cases}$ où les coefficients $(a_k)_{k=-r}^{p}$ sont fixés.

Proposition C.11. *L'opérateur $T_{\mathbb{N}} : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$, défini en Définition C.10, est linéaire et borné.*

Démonstration. En reprenant les notations de la Définition C.10. L'opérateur $T_{\mathbb{N}}$ est clairement linéaire. De plus, on a

$$\begin{aligned} \sum_{n=0}^{+\infty} |(T_{\mathbb{N}}u)_n|^2 &\leqslant \max_{j=-r}^{p} |a_j| \left(\sum_{n=r}^{+\infty} \left| \sum_{j=-r}^{p} u_{n+j} \right|^2 + \sum_{n=0}^{r-1} \left| \sum_{j=-n}^{p} u_{n+j} \right|^2 \right) \\ &\leqslant \max_{j=-r}^{p} |a_j| (r+p+1) \left(\sum_{n=r}^{+\infty} \sum_{j=-r}^{p} |u_{n+j}|^2 + \sum_{n=0}^{r-1} \sum_{j=-n}^{p} |u_{n+j}|^2 \right) \\ &\leqslant \max_{j=-r}^{p} |a_j| (r+p+1)^2 \sum_{n=0}^{+\infty} |u_n|^2 \end{aligned}$$

en faisant des changements d'indice. Ainsi $T_{\mathbb{N}}$ est un opérateur linéaire borné. □

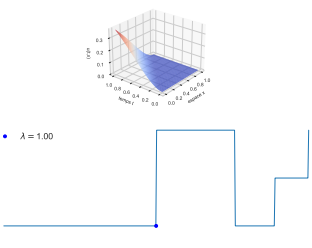
On rappelle qu'on note Γ la courbe du symbole γ :

$$\Gamma = \left\{ \xi \in [0, 2\pi] \mapsto \gamma(\xi) \overset{\text{def}}{=} \sum_{k=-r}^{p} a_k e^{ik\xi} \right\}.$$

Théorème C.12. *Soit $T_{\mathbb{N}}$ un opérateur de Toeplitz de symbole γ. On a la disjonction de cas :*

- *$z \in \Gamma$ si et seulement si $T_{\mathbb{N}} - z$ n'est pas de Fredholm.*
- *$z \in \{\mathrm{Ind}_\Gamma = 0\}$ si et seulement si $T_{\mathbb{N}} - z$ est bijectif.*
- *$z \in \{\mathrm{Ind}_\Gamma \neq 0\}$ si et seulement si $T_{\mathbb{N}} - z$ est de Fredholm et d'indice*

$$\mathrm{ind}(T_{\mathbb{N}} - z) = \mathrm{Ind}_\Gamma(z) \neq 0.$$



Remarque C.13. Ici, la notation $\mathrm{ind}(T)$ désigne l'indice de Fredholm, défini par

$$\mathrm{ind}(T) = \dim\ker(T) - \mathrm{codim}\,\mathrm{Ran}(T),$$

et la notation $\mathrm{Ind}_\Gamma(z)$ désigne l'indice complexe de z par rapport à la courbe Γ, défini par

$$\mathrm{Ind}_\Gamma(z) = \frac{1}{2i\pi}\int_\Gamma \frac{d\zeta}{\zeta - z}.$$

Démonstration.

- Si $z \in \Gamma$, alors $T_\mathbb{N} - z$ est de symbole nul, donc $T_\mathbb{N} - z$ n'est pas Fredholm (voir (1) de III 1.6 de [Nik17]). Réciproquement, si $T_\mathbb{N} - z$ n'est pas Fredholm, alors $T_\mathbb{N} - z$ est de symbole nul, donc $z \in \Gamma$.
- Si $z \in \{\mathrm{Ind}_\Gamma = 0\}$, alors $z \notin \Lambda(T_\mathbb{N})$ (voir [SS60]), ainsi $T - \mathbb{N}$ est bijectif. Réciproquement, si $T_\mathbb{N} - z$ est bijectif, alors $z \notin \Gamma$ par [SS60]. Ainsi, $T_\mathbb{N} - z$ est Fredholm (par le premier point) avec pour indice $\mathrm{ind}(T_\mathbb{N} - z) = 0$ (car bijectif). Or l'affirmation (3) de III 1.6 de [Nik17] nous dit qu'un opérateur de Toeplitz, qui est de Fredholm, est d'indice $\mathrm{ind}(T_\mathbb{N} - z) = \mathrm{Ind}_\Gamma(z)$, ce qui permet de dire que $z \in \{\mathrm{Ind}_\Gamma = 0\}$. (Attention, dans cet ouvrage leur indice complexe wind est l'opposé de notre indice Ind, car leur symbole est défini dans l'autre sens que le notre.)
- Si $z \in \{\mathrm{Ind}_\Gamma \neq 0\}$, $T_\mathbb{N} - z$ est Fredholm et il y a concordance entre l'indice de Fredholm et l'indice complexe (voir (3) de III 1.6 de [Nik17]). Réciproquement, si $T_\mathbb{N} - z$ est Fredholm d'indice non nul, on a $z \in \{\mathrm{Ind}_\Gamma \neq 0\}$.

□

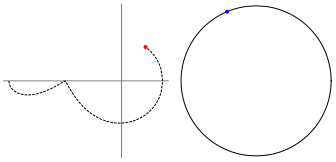
Proposition C.14. *Soient* $(k_{i,j})_{\substack{0\leqslant i\leqslant r-1\\ 0\leqslant j\leqslant m-1}} \in \mathbb{C}^{r\times m}$. *L'opérateur* $K : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ *défini, pour tout* $(u_n)_n \in \ell^2(\mathbb{N})$, *par* $(Ku)_n = \sum_{j=0}^{m-1} k_{n,j}u_j$ *si* $n < r$ *et* $(Ku)_n = 0$ *si* $n \geqslant r$, *est compact.*

Démonstration. L'image de l'opérateur K est de dimension finie car elle est dans $\mathrm{Vect}(\{u^{(0)}, \dots, u^{(r-1)}\})$ où, pour tout $i \in \mathbb{N}$, pour tout $n \in \mathbb{N}$,

$$u_n^{(i)} = \begin{cases} 1 & \text{si } n = i, \\ 0 & \text{sinon.} \end{cases}$$

De plus, K est clairement borné. Ainsi K est compact car $\overline{K(B(0,1))}$ est borné et fermé dans l'espace $\mathrm{Vect}(\{u^{(0)}, \dots, u^{(r-1)}\})$ de dimension finie. □

Remarque C.15. La Proposition C.14 est très importante car, combinée avec la Proposition C.6, cela montre que les opérateurs Quasi-Toeplitz $\widetilde{T_\mathbb{N}}$ sont des opérateurs de Fredholm de même indice que l'opérateur de Toeplitz $T_\mathbb{N}$ associé (sans bord). En effet, par définition de $\widetilde{T_\mathbb{N}}$, on a $\widetilde{T_\mathbb{N}} = K + T_\mathbb{N}$ avec K un opérateur rentrant dans le cadre des hypothèses de la Proposition C.14.



Quelques éléments sur le schéma leap-frog

On se fixe quatre entiers r, p, s et m. Pour définir les schémas multipas, on se donne la convention d'écriture suivante :

$$\begin{cases} U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} \sum_{\ell=-r}^{p} a_{\ell,\sigma} U_{j+\ell}^{n-\sigma}, & n \geqslant s,\ j \geqslant r, \quad \text{(D.1a)} \\ U_j^{n+1} = \displaystyle\sum_{\sigma=0}^{s} \sum_{\ell=0}^{m-1} b_{j,\ell}^{(\sigma)} U_\ell^{n-\sigma} + g_j^{n+1}, & n \geqslant s,\ j \in [\![0 : r-1]\!], \quad \text{(D.1b)} \\ U_j^n = f_j^n, & n \leqslant s,\ j \geqslant 0. \quad \text{(D.1c)} \end{cases}$$

On peut visualiser la discrétisation de ce schéma de la manière suivante :

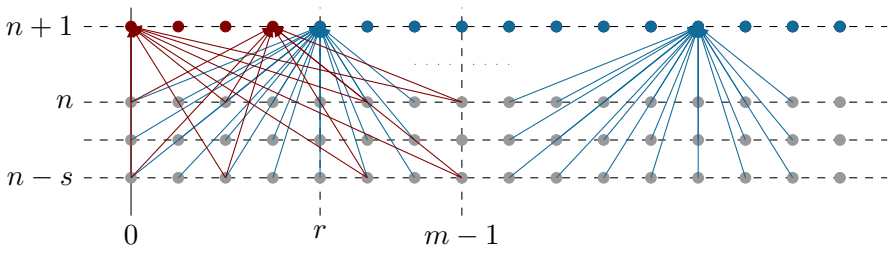
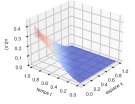
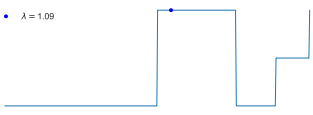


Figure D.1 – Discrétisation du schéma (D.1).

Comme dans la Section 6.7, pour faciliter l'écriture de la condition de bord (D.1b), on pose les matrices suivantes :

$$\forall \sigma \in [\![0 : s]\!], \quad B_\sigma \overset{\text{def}}{=} \begin{pmatrix} b_{0,0}^{(\sigma)} & \cdots & b_{0,m-1}^{(\sigma)} \\ \vdots & & \vdots \\ b_{r-1,0}^{(\sigma)} & \cdots & b_{r-1,m-1}^{(\sigma)} \end{pmatrix}. \quad \text{(D.2)}$$





D.1 Leap-frog

D.1.1 Définition

La schéma leap-frog est un schéma multipas avec deux pas de temps qui s'écrit

$$U_j^{n+1} = U_j^{n-1} + \lambda(U_{j+1}^n - U_{j-1}^n) \tag{D.3}$$

Pour identifier le schéma (D.3) avec les notations de (D.1a), on a $s = 1$, $r = 1$ et $p = 1$, avec pour coefficients

$$\begin{array}{ccc} a_{-1,0} = -\lambda & a_{0,0} = 0 & a_{1,0} = \lambda \\ a_{-1,1} = 0 & a_{0,1} = 1 & a_{1,1} = 0. \end{array}$$

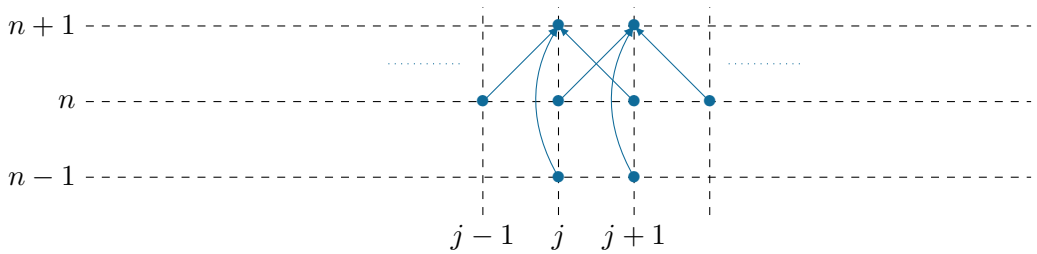


FIGURE D.2 – Discrétisation du schéma leap-frog (D.3).

D.1.2 Équation caractéristique

L'équation caractéristique d'inconnue κ s'écrit, pour tout $z \in \overline{\mathcal{U}}$,

$$\left(z - \frac{1}{z}\right) = \lambda\left(\kappa - \frac{1}{\kappa}\right) \quad \text{ou} \quad \lambda z\kappa^2 + (1 - z^2)\kappa - \lambda z = 0. \tag{D.4}$$

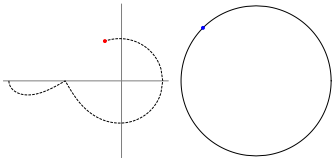
Le coefficient dominant et le coefficient constant sont bien non nuls pour tout $z \in \overline{\mathcal{U}}$, donc l'Hypothèse 6.27 (page 173) est bien satisfaite.

De plus, pour la condition CFL $\lambda \in]0, 1[$, le schéma de leap-frog est Cauchy-stable au sens de l'Hypothèse 6.28 (page 173). En effet, pour tout $\xi \in \mathbb{R}$, la matrice $\mathscr{A}(\xi)$, définie en Hypothèse 6.28, est de la forme

$$\mathscr{A}(\xi) = \begin{pmatrix} 2i\lambda \sin\xi & 1 \\ 1 & 0 \end{pmatrix}.$$

Les valeurs propres de $\mathscr{A}(\xi)$ sont $\pm\sqrt{1 - \lambda^2 \sin^2(\xi)} + i\lambda \sin\xi$ qui sont de module 1 mais simples tant que $\lambda \in]0, 1[$.

On sait que si $|z| > 1$ alors $|\kappa(z)| \neq 1$ (par le Lemme 3.4 (Hersh) généralisé aux schémas multipas [Cou13, Lem.3.7]) et que $\kappa_-(z)\kappa_+(z) = -1$, ainsi si $|z| > 1$, on peut ordonner les



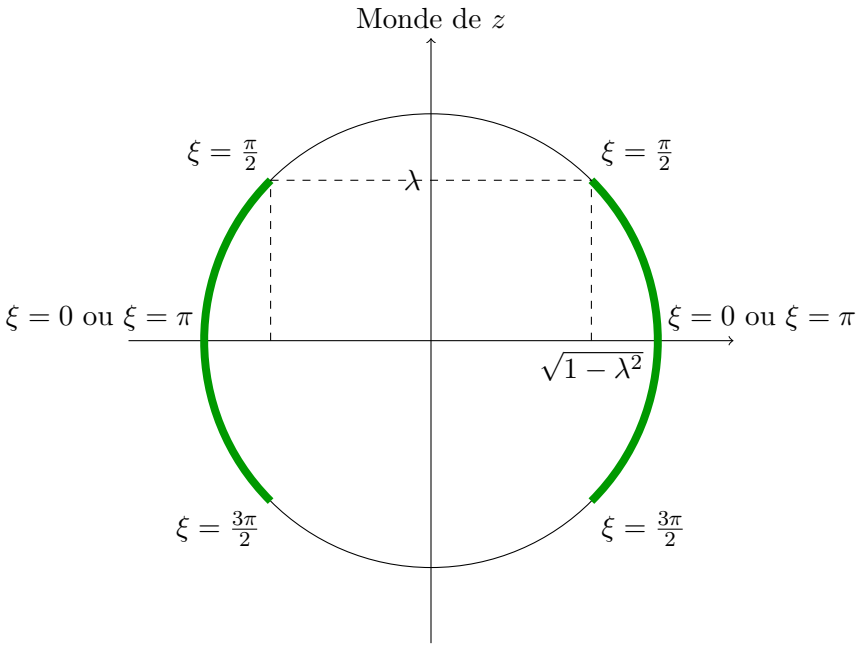


FIGURE D.3 – Les éléments z solution de (D.4) en vert en fonction de $\kappa = e^{i\xi}$.

racines κ de la manière suivante :

$$|\kappa_-(z)| < 1 < |\kappa_+(z)|.$$

On présente cette situation en Figure D.4. De plus, on trace la trajectoire de $\kappa_-(z)$ et $\kappa_+(z)$ quand z parcourt le cercle de centre 0 et de rayon 1.005 en Figure (a) de la Figure D.5.

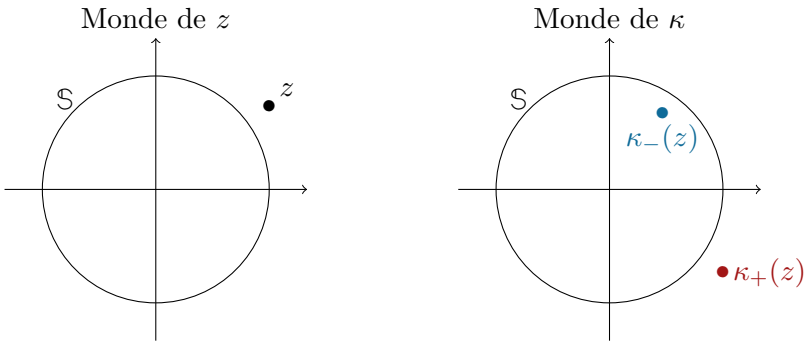
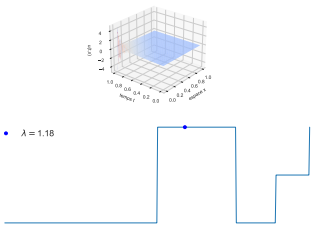


FIGURE D.4 – Le cas $|z| > 1$.



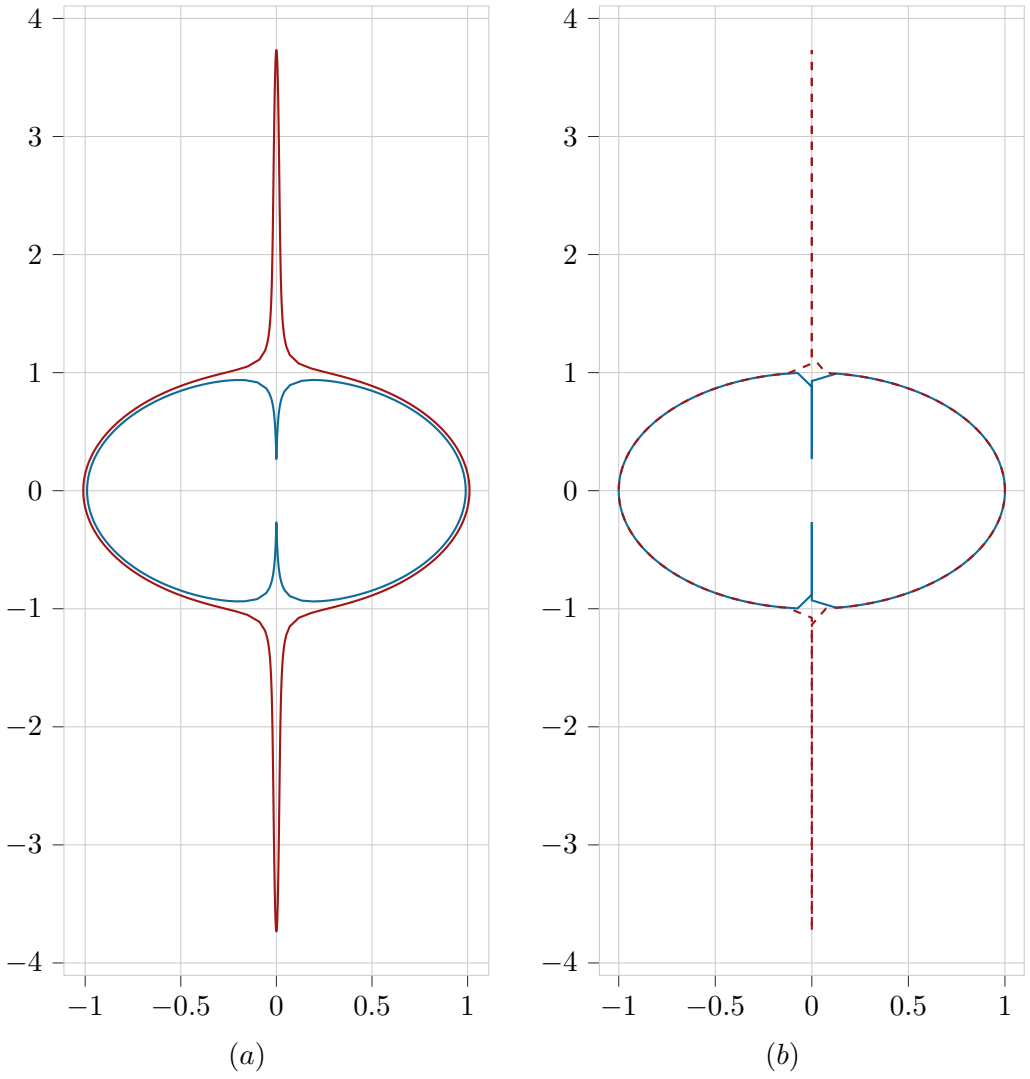
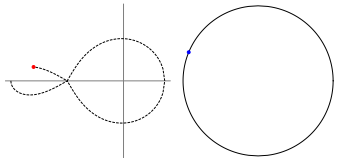


FIGURE D.5 – Trajectoire de $\kappa_-(z)$ et $\kappa_+(z)$ pour $|z| = 1.005$ (Figure (a)) et pour $|z| = 1$ (Figure (b)) pour une discrétization du cercle de 1000 points.



Si z est de module 1, les racines κ de l'équation (D.4) peuvent être de module 1. En faisant le calcul pour $z_0 = 1$, on trouve la Figure D.6. On trace la trajectoire de $\kappa_-(z)$ et $\kappa_+(z)$ quand z parcourt $\mathbb{S}$ en Figure (b) de la Figure D.5. En i et en $-i$, on devrait voir des angles droits pour les valeurs $z = \pm\sqrt{1-\lambda^2} \pm i\lambda$ qui sont les points de rebroussement que l'on observe en Figure D.3, il faudrait augmenter le nombre de points de la discrétisation pour voir plus précisément les angles droits.

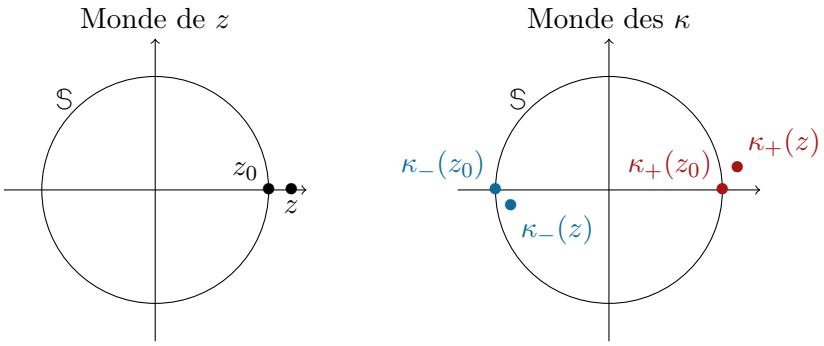


FIGURE D.6 – Le cas $|z| = 1$.

D.2 Conditions de bord

Trefethen [Tre84] présente quatre conditions de bords qui donnent différents types de stabilité et d'instabilités. On les réécrit ici dans le formalisme de (D.1b). Ce sont les quatre conditions de bord présentées dans les figures de pseudospectres à la Section 2.4.4 (page 66) et dans les figures de courbes de déterminant de Kreiss–Lopatinskii de la Section 6.7.2 (page 174).

Comme dans [Tre84], on se place dans le cas où $\lambda = 0.5$.

D.2.1 Mode stable

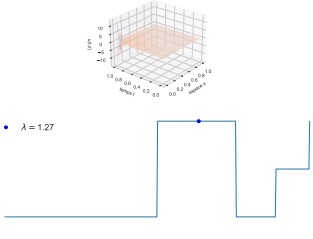
La condition de bord α de [Tre84] s'écrit

$$U_0^{n+1} = U_1^n. \tag{α}$$

Pour identifier la condition de bord (α) avec les notations de (D.2), on a $m = 2$, avec

$$B_0 = \begin{pmatrix} 0 & 1 \end{pmatrix} \quad \text{et} \quad B_1 = \begin{pmatrix} 0 & 0 \end{pmatrix}.$$

Ce cas correspond à un schéma stable car il n'a ni valeur propre ni valeur propre généralisée.



D.2.2 Mode instable croissant

La condition de bord β de [Tre84] s'écrit

$$U_0^{n+1} = U_1^{n-2}. \tag{β}$$

Grâce au schéma (D.3), on peut réécrire, $U_1^{n-2} = U_1^n - \lambda(U_2^{n-1} - U_0^{n-1})$, ainsi en identifiant la condition de bord (β) avec les notations de (D.2), on a $m = 3$, avec

$$B_0 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \quad \text{et} \quad B_1 = \begin{pmatrix} \lambda & 0 & -\lambda \end{pmatrix}.$$

Ce cas correspond à un schéma instable, car il possède des valeurs propres généralisées aux points $z = \pm e^{\pm i\pi/6}$ associés aux valeurs $\kappa = \pm i$ avec une vitesse de groupe nulle, car $z = \pm e^{\pm i\pi/6}$ correspond à des points de rebroussement que l'on peut observer sur la Figure D.3. Dans l'article [Tre84], il utilise le terme « rightgoing steady state solution » pour décrire ce phénomène.

D.2.3 Mode instable strictement croissant

La condition de bord γ de [Tre84] s'écrit

$$U_0^{n+1} = \frac{1}{2}(U_0^n + U_2^n). \tag{γ}$$

En identifiant la condition de bord (γ) avec les notations de (D.2), on a $m = 3$, avec

$$B_0 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \quad \text{et} \quad B_1 = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}.$$

Ce cas correspond à un schéma instable, car il possède une valeur propre généralisée en $z = 1$ associée aux valeurs $\kappa = \pm 1$ avec une vitesse de groupe non nulle. Dans l'article [Tre84], il utilise le terme « strictly rightgoing steady state solution » pour décrire ce phénomène.

D.2.4 Mode instable strictement croissant avec coefficient de réflexion infini

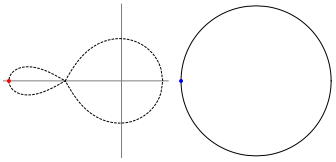
La condition de bord δ de [Tre84] s'écrit

$$U_0^{n+1} = U_1^{n+1}. \tag{δ}$$

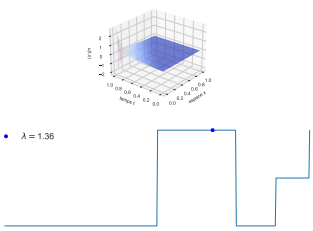
Grâce au schéma (D.3), on peut écrire $U_1^{n+1} = U_1^{n-1} + \lambda(U_2^n - U_0^n)$ et ainsi, en identifiant la condition de bord (δ) avec les notations de (D.2), on a $m = 3$, avec

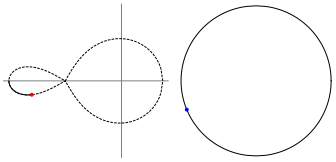
$$B_0 = \begin{pmatrix} -\lambda & 0 & \lambda \end{pmatrix} \quad \text{et} \quad B_1 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}.$$

Ce cas correspond à un schéma instable, car il possède une valeur propre généralisée en



$z = -1$ associée aux valeurs $\kappa = \pm 1$ avec une vitesse de groupe non nulle. Dans l'article [Tre84], il utilise le terme « strictly rightgoing steady state solution » pour décrire ce phénomène comme dans le cas de la condition de bord (γ). La différence avec ce dernier est le coefficient de réflexion, défini dans [Tre84], qui est infini dans le cas (δ).





Annexe E

Quelques éléments sur un coefficient binomial modifié

E.1 Définition

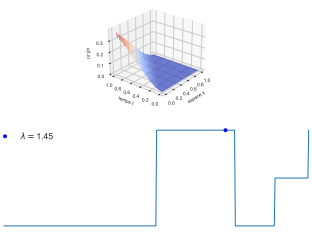
On introduit la notation suivante.

Définition E.1. Soit $n \in \mathbb{N}^*$ et soit $\ell \in [\![1 : n]\!]$. On définit le coefficient $\begin{bmatrix} n \\ \ell \end{bmatrix}$ par récurrence de la manière suivante :

$$\forall n \in \mathbb{N}, \quad \begin{bmatrix} n+1 \\ \ell \end{bmatrix} = \begin{cases} 1 & \text{si } \ell \in \{1, n+1\}, \\ \begin{bmatrix} n \\ \ell - 1 \end{bmatrix} + \ell \begin{bmatrix} n \\ \ell \end{bmatrix} & \text{si } \ell \in [\![2 : n]\!]. \end{cases}$$

Ce coefficient ressemble au coefficient binomial, on peut en donner un triangle de Pascal modifié.

		ℓ							
		1	2	3	4	5	6	7	$\cdots$
n	1	1							
	2	1	1						
	3	1	3	1					
	4	1	7	6	1				
	5	1	15	25	10	1			
	6	1	31	90	65	15	1		
	7	1	63	301	350	140	21	1	
	8	1	127	966	1701	1050	266	28	1
	$\vdots$								



On pose alors la matrice suivante :

$$P_d = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \ddots & & \vdots \\ \vdots & \begin{bmatrix} 2 \\ 1 \end{bmatrix} & \begin{bmatrix} 2 \\ 2 \end{bmatrix} & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & \begin{bmatrix} d-1 \\ 1 \end{bmatrix} & \begin{bmatrix} d-1 \\ 2 \end{bmatrix} & \cdots & \begin{bmatrix} d-1 \\ d-1 \end{bmatrix} \end{pmatrix} \in \mathcal{M}_d(\mathbb{C}).$$

E.2 Propriété

La matrice P_d est une matrice de changement de base entre la base canonique de $\mathbb{R}_{d-1}[X]$ et la base des polynômes $(\prod_{j=0}^{\ell-1}(X-j))_{\ell=0}^{d-1}$. On prouve cette assertion dans la proposition suivante.

Proposition E.2. *Pour tout $s \in \mathbb{R}$, on a la relation*

$$P_d \begin{pmatrix} 1 \\ s \\ s(s-1) \\ s(s-1)(s-2) \\ \vdots \\ s(s-1)\cdots(s-(d-2)) \end{pmatrix} = \begin{pmatrix} 1 \\ s \\ s^2 \\ s^3 \\ \vdots \\ s^{d-1} \end{pmatrix}.$$

Démonstration. Soit $s \in \mathbb{R}$. Montrons par récurrence sur n que, pour tout $n \in [\![1 : d-1]\!]$, on a

$$s^n = \sum_{\ell=1}^{n} \begin{bmatrix} n \\ \ell \end{bmatrix} s(s-1)(s-2)\cdots(s-\ell+1). \tag{E.1}$$

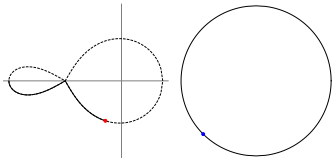
Initialisation : Pour $n = 1$. On a

$$s^n = s \text{ et } \begin{bmatrix} 1 \\ 1 \end{bmatrix} s = s.$$

Hérédité : Supposons que la formule (E.1) soit vraie au rang $n \in \mathbb{N}^*$. Montrons-la au rang $n+1$. On a

$$s^n X^s = \sum_{\ell=1}^{n} \begin{bmatrix} n \\ \ell \end{bmatrix} s(s-1)(s-2)\cdots(s-\ell+1) X^\ell X^{s-\ell}$$

par hypothèse de récurrence au rang n. En dérivant par rapport à X l'expression précédente,



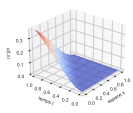
on obtient

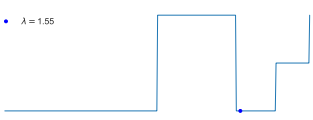
$$\begin{aligned}
s^{n+1}X^{s-1} &= \sum_{\ell=1}^{n} \begin{bmatrix} n \\ \ell \end{bmatrix} s(s-1)\cdots(s-\ell+1)\left(\ell X^{\ell-1}X^{s-\ell} + (s-\ell)X^{s-\ell-1}X^{\ell}\right) \\
&= \sum_{\ell=1}^{n} \begin{bmatrix} n \\ \ell \end{bmatrix} \ell s(s-1)\cdots(s-\ell+1)X^{s-1} + \sum_{\ell=2}^{n+1} \begin{bmatrix} n \\ \ell-1 \end{bmatrix} s(s-1)\cdots(s-\ell+1)X^{s-1} \\
&= \begin{bmatrix} n \\ 1 \end{bmatrix} sX^{s-1} + \begin{bmatrix} n \\ n \end{bmatrix} s\cdots(s-n)X^{s-1} + \sum_{\ell=2}^{n} \left(\begin{bmatrix} n \\ \ell-1 \end{bmatrix} + \ell \begin{bmatrix} n \\ \ell \end{bmatrix} \right) s\cdots(s-\ell+1)X^{s-1} \\
&= \begin{bmatrix} n+1 \\ 1 \end{bmatrix} sX^{s-1} + \begin{bmatrix} n+1 \\ n+1 \end{bmatrix} s\cdots(s-n)X^{s-1} + \sum_{\ell=2}^{n} \begin{bmatrix} n+1 \\ \ell \end{bmatrix} s\cdots(s-\ell+1)X^{s-1} \\
&= \sum_{\ell=1}^{n+1} \begin{bmatrix} n+1 \\ \ell \end{bmatrix} s(s-1)\cdots(s-\ell+1)X^{s-1}.
\end{aligned}$$

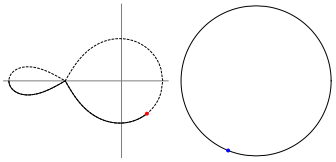
Ainsi, on a

$$s^{n+1} = \sum_{\ell=1}^{n+1} \begin{bmatrix} n+1 \\ \ell \end{bmatrix} s(s-1)\cdots(s-\ell+1).$$

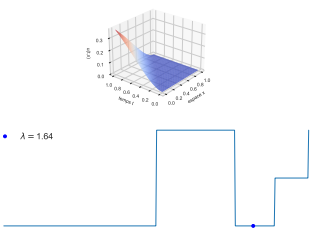
Conclusion : Ainsi, la formule (E.1) est vraie pour tout $n \in \mathbb{N}^*$. □

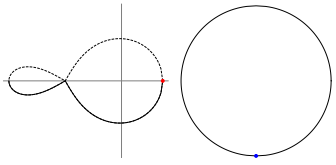


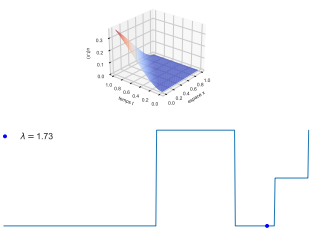


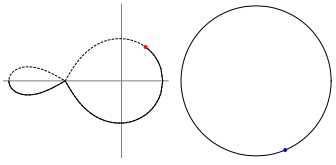


LISTE DES FIGURES









BIBLIOGRAPHY

[AES03] Anton Arnold, Matthias Ehrhardt, and Ivan Sofronov. Discrete transparent boundary conditions for the Schrödinger equation: fast calculation, approximation, and stability. *Communications in Mathematical Sciences*, 1(3):501–556, 2003.

[Ben22] Antoine Benoit. Stability of finite difference schemes approximation for hyperbolic boundary value problems in an interval. *Mathematics of Computation*, 91, 2022.

[BGS06] Sylvie Benzoni-Gavage and Denis Serre. *Multi-dimensional hyperbolic partial differential equations: First-order Systems and Applications*. OUP Oxford, 2006.

[BLBS23a] Benjamin Boutin, Pierre Le Barbenchon, and Nicolas Seguin. On the stability of totally upwind schemes for the hyperbolic initial boundary value problem. *accepted in IMA Journal of Numerical Analysis (IMAJNA)*, 2023.

[BLBS23b] Benjamin Boutin, Pierre Le Barbenchon, and Nicolas Seguin. Stability of finite difference schemes for the hyperbolic initial boundary value problem by winding number computations. Submitted, 2023.

[BNS+21] Benjamin Boutin, Thi Hoai Thuong Nguyen, Abraham Sylla, Sébastien Tran-Tien, and Jean-François Coulombel. High order numerical schemes for transport equations on bounded domains. *ESAIM: Proceedings and Surveys*, 70:84–106, 2021.

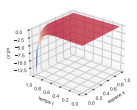
[BP12] Marc Briane and Gilles Pagès. *Théorie de l'intégration*. Vuibert, 2012.

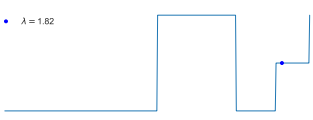
[BS00] Natalia Borovykh and Marc N. Spijker. Resolvent conditions and bounds on the powers of matrices, with relevance to numerical stability of initial value problems. *Journal of Computational and Applied Mathematics*, 125(1-2):41–56, 2000.

[BW93] Richard M. Beam and Robert F. Warming. The asymptotic spectra of banded Toeplitz and quasi-Toeplitz matrices. *SIAM Journal on Scientific Computing*, 14(4):971–1006, 1993.

[BWY82] Richard M. Beam, Robert F. Warming, and Helen Yee. Stability analysis of numerical boundary conditions and implicit difference approximations for hyperbolic equations. *NASA Publications*, 1982.

[CFL28] Richard Courant, Kurt Friedrichs, and Hans Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100(1):32–74, 1928.





[CL20] Jean-François Coulombel and Frédéric Lagoutière. The Neumann numerical boundary condition for transport equations. *Kinetic and Related Models*, 13(1):1–32, 2020.

[CN47] John Crank and Phyllis Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, 43(1):50–67, 1947.

[Cou13] Jean-François Coulombel. Stability of finite difference schemes for hyperbolic initial boundary value problems. In *HCDTE lecture notes. Part I. Nonlinear hyperbolic PDEs, dispersive and transport equations*, volume 6 of *AIMS Ser. Appl. Math.*, page 146. Am. Inst. Math. Sci. (AIMS), Springfield, MO, 2013.

[Cou19] Jean-François Coulombel. Transparent numerical boundary conditions for evolution equations: derivation and stability analysis. *Ann. Fac. Sci. Toulouse, Math. (6)*, 28(2):259–327, 2019.

[CR21] Christophe Cheverry and Nicolas Raymond. *A Guide to Spectral Theory*. Springer International Publishing, 2021.

[DDJ18] Gautier Dakin, Bruno Després, and Stéphane Jaouen. Inverse Lax–Wendroff boundary treatment for compressible Lagrange-remap hydrodynamics on cartesian grids. *Journal of Computational Physics*, 353:228–257, 2018.

[Ehr10] Matthias Ehrhardt. Absorbing boundary conditions for hyperbolic systems. *Numer. Math., Theory Methods Appl.*, 3(3):295–337, 2010.

[EM77] Bjorn Engquist and Andrew Majda. Absorbing boundary conditions for the numerical simulation of waves. *Mathematics of Computation*, 31:629–651, 1977.

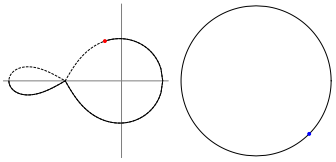
[Fra06] W Randolph Franklin. Pnpoly-point inclusion in polygon test. *Web site: http://www. ecse. rpi. edu/Homepages/wrf/Research/Short_Notes/pnpoly. html*, 2006.

[GKO13] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. *Time-dependent problems and difference methods*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[GKS72] Bertil Gustafsson, Heinz-Otto Kreiss, and Arne Sundström. Stability theory of difference approximations for mixed initial boundary value problems. II. *Mathematics of Computation*, 26(119):649–649, 1972.

[Gol77] Moshe Goldberg. On a boundary extrapolation theorem by Kreiss. *Mathematics of Computation*, 31(138):469–477, 1977.

[GR63] Sergei K. Godunov and Victor S. Ryabenkii. Spectral stability criteria for boundary-value problems for non-self-adjoint difference equations. *Russian Mathematical Surveys*, 18(3):1–12, 1963.



[GT78] Moshe Goldberg and Eitan Tadmor. Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. i. *Mathematics of Computation*, 32(144):1097–1107, 1978.

[GT81] Moshe Goldberg and Eitan Tadmor. Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. ii. *mathematics of computation*, 36(154):603–626, 1981.

[Gus08] Bertil Gustafsson. *High Order Difference Methods for Time Dependent PDE*. Number 38. Springer Series in Computational Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg., Berlin, 2008.

[GW99] Claude Gasquet and Patrick Witomski. *Fourier Analysis and Applications: Filtering, Numerical Computation, Wavelets*, volume 30. Texts in applied mathematics. Springer-Verlag, New York., 1999.

[GZDM12] Juan Luis García Zapata and Juan Carlos Díaz Martín. A geometric algorithm for winding number computation with complexity analysis. *Journal of Complexity*, 28(3):320–345, 2012.

[GZDM14] Juan Luis García Zapata and Juan Carlos Díaz Martín. Finding the number of roots of a polynomial in a plane region using the winding number. *Computers & Mathematics with Applications*, 67(3):555–568, 2014.

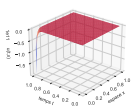
[Her77] Charles Hermite. Sur la formule d'interpolation de Lagrange. (Extrait d'une lettre de M. Charles Hermite à M. Borchardt). *Journal für die reine und angewandte Mathematik*, 84:70–79, 1877.

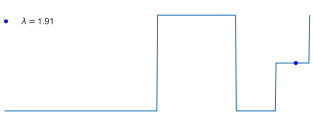
[Her63] Reuben Hersh. Mixed problems in several variables. *Journal of Mathematics and Mechanics*, 12(3):317–334, 1963.

[HMvdW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[IS83] Arieh Iserles and Gilbert Strang. The optimal accuracy of difference schemes. *Transactions of the American Mathematical Society*, 277(2):779–803, 1983.

[Kre62] Heinz-Otto Kreiss. Über Die Stabilitätsdefinition Für Differenzengleichungen Die Partielle Differentialgleichungen Approximieren. *BIT Numerical Mathematics*, 2:153–181, 1962.





[Kre66] Heinz-Otto Kreiss. Difference approximations for hyperbolic differential equations. *Numerical Solution of Partial Differential Equations (Proc.Sympos. Univ. Maryland, 1965)*, pages 51–58, 1966.

[Kre68] Heinz-Otto Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. I. *Mathematics of Computation*, 22(104):703–703, 1968.

[Lan95] Serge Lang. *Differential and Riemannian Manifolds*. Springer New York, New York, 1995. OCLC: 958525799.

[Lan99] Serge Lang. *Complex analysis*, volume 103. Graduate texts in mathematics. Springer-Verlag, New York., 4th edition, 1999.

[Lap75] Genadii Ivanovich Laptev. Conditions for the uniform well-posedness of the cauchy problem for systems of equations. In *Doklady Akademii Nauk*, volume 220, pages 281–284. Russian Academy of Sciences, 1975.

[LB23] Pierre Le Barbenchon. Boundaryscheme: Python package for numerical scheme with boundary. https://doi.org/10.5281/zenodo.7773742, 2023.

[LLS22] Tingting Li, Jianfang Lu, and Chi-Wang Shu. Stability analysis of inverse Lax–Wendroff boundary treatment of high order compact difference schemes for parabolic equations. *Journal of Computational and Applied Mathematics*, 400:113711, 2022.

[LM06] Frieder Lörcher and Claus-Dieter Munz. Lax–Wendroff-type schemes of arbitrary order in several space dimensions. *IMA Journal of Numerical Analysis*, 27(3):593–615, 2006.

[LR56] Peter D. Lax and Robert D. Richtmyer. Survey of the stability of linear finite difference equations. *Communications on Pure and Applied Mathematics*, 9(2):267–293, 1956.

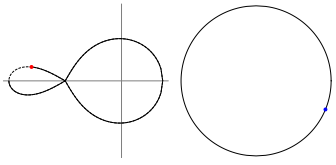
[LS91] Hermanus Wilhelmus Johannes Lenferink and Marc Nico Spijker. On a generalization of the resolvent condition in the kreiss matrix theorem. *Mathematics of computation*, 57(195):211–220, 1991.

[LSZ16] Tingting Li, Chi-Wang Shu, and Mengping Zhang. Stability analysis of the inverse Lax–Wendroff boundary treatment for high order upwind-biased finite difference schemes. *Journal of Computational and Applied Mathematics*, 299:140–158, 2016.

[LSZ17] Tingting Li, Chi-Wang Shu, and Mengping Zhang. Stability analysis of the inverse Lax–Wendroff boundary treatment for high order central difference schemes for diffusion equations. *Journal of Scientific Computing*, 70(2):576–607, 2017.

[LT84] Randall J. Leveque and Lloyd N. Trefethen. On the resolvent condition in the Kreiss Matrix Theorem. *BIT Numerical Mathematics*, 24(4):584–591, 1984.

[LW60] Peter D. Lax and Burton Wendroff. Systems of conservation laws. *Communications on Pure and Applied Mathematics*, 13:217–237, 1960.



[Mét00] Guy Métivier. The block structure condition for symmetric hyperbolic systems. *Bulletin of the London Mathematical Society*, 32(6):689 – 702, 2000.

[Mil67] John Miller. On power-bounded operators and operators satisfying a resolvent condition. *Numerische Mathematik*, 10:389–396, 1967.

[Mor64] Keith William Morton. On a matrix theorem due to Heinz-Otto Kreiss. *Communications on Pure and Applied Mathematics*, 17:375–379, 1964.

[MS66] John Miller and Gilbert Strang. Matrix theorems for partial differential and difference equations. *Mathematica Scandinavica*, 18(2):113–133, 1966.

[MSP+17] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in Python. *PeerJ Computer Science*, 3:e103, January 2017.

[MZ04] Guy Métivier and Kevin Zumbrun. Symmetrizers and continuity of stable subspaces for parabolic-hyperbolic boundary value problems. *Discrete & Continuous Dynamical Systems - A*, 11(1):205–220, 2004.

[Nik17] Nikolaï Nikolski. *Matrices et opérateurs de Toeplitz*. Calvage et Mounet, 2017.

[O'R98] Joseph O'Rourke. *Computational geometry in C*. Cambridge university press, 1998.

[RT92] Lothar Reichel and Lloyd N. Trefethen. Eigenvalues and pseudo-eigenvalues of toeplitz matrices. *Linear Algebra and its Applications*, 162-164:153–185, 1992.

[Rud11] Walter Rudin. *Analyse réelle et complexe: cours et exercices*. Dunod, Paris, 3e éd edition, 2011. OCLC: 999564674.

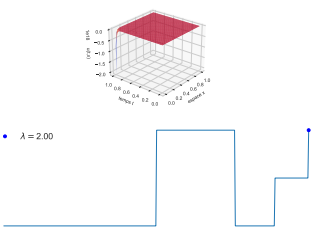
[Spi98] Marc Spijker. Numerical Stability - Stability Estimates and Resolvent Conditions in the Numerical Solution of Initial Value Problem, 1998.

[SS60] Palle Schmidt and Frank Spitzer. The Toeplitz matrices of an arbitrary laurent polynomial. *Mathematica Scandinavica*, 8(0):15–38, 1960.

[Str04] John C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Society for Industrial and Applied Mathematics, Philadelphia, 2nd ed edition, 2004.

[STW02] Marc N. Spijker, Stefania Tracogna, and Bruno D. Welfert. About the sharpness of the stability estimates in the Kreiss matrix theorem. *Mathematics of Computation*, 72(242):697–714, 2002.

[Sun21] Daniel Sunday. *Practical Geometry Algorithms: With C++ Code*. Amazon Digital Services LLC - KDP Print US, 2021.



[Tad81] Eitan Tadmor. The equivalence of L2-stability, the resolvent condition, and strict h-stability. *Linear Algebra and Its Applications*, 41:151–159, 1981.

[TE05] Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, N.J, 2005.

[Thu86] Michael Thuné. Automatic GKS stability analysis. *SIAM J. Sci. Statist. Comput.*, 7(3):959–977, 1986.

[Toe11] Otto Toeplitz. Zur Theorie der quadratischen und bilinearen Formen von unendlichvielen Veränderlichen. *Mathematische Annalen*, 70(3):351–376, 1911.

[Tre83] Lloyd N. Trefethen. Group velocity interpretation of the stability theory of Gustafsson, Kreiss, and Sundström. *J. Comput. Phys.*, 49(2):199–217, 1983.

[Tre84] Lloyd N. Trefethen. Instability of difference models for hyperbolic initial boundary value problems. *Communications on Pure and Applied Mathematics*, 37(3):329–367, 1984.

[TS10] Sirui Tan and Chi-Wang Shu. Inverse Lax-Wendroff procedure for numerical boundary conditions of conservation laws. *Journal of Computational Physics*, 229(21):8144 – 8166, 2010.

[TS13] Sirui Tan and Chi-Wang Shu. *Inverse Lax–Wendroff Procedure for Numerical Boundary Conditions of Hyperbolic Equations: Survey and New Developments*, pages 41–63. Springer US, Boston, MA, 2013.

[TT99] Kim-Chuan Toh and Lloyd N. Trefethen. The Kreiss matrix theorem on a general complex domain. *SIAM Journal on Matrix Analysis and Applications*, 21(1):145–165, 1999.

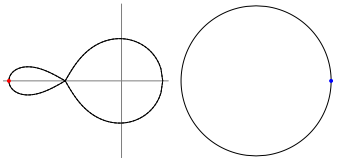
[TWSN12] Sirui Tan, Cheng Wang, Chi-Wang Shu, and Jianguo Ning. Efficient implementation of high order inverse Lax–Wendroff boundary treatment for conservation laws. *Journal of Computational Physics*, 231(6):2510–2527, 2012.

[VS15] François Vilar and Chi-Wang Shu. Development and stability analysis of the inverse Lax Wendroff boundary treatment for central compact schemes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(1):39–67, 2015.

[WB76] Robert F. Warming and Richard M. Beam. Upwind second-order difference schemes and applications in aerodynamic flows. *AIAA Journal*, 14(9):1241–1249, sep 1976.

[Wid89] Harold Widom. On the singular values of Toeplitz matrices. *Zeitschrift für Analysis und ihre Anwendungen*, 8(3):221–229, 1989.

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



COLLEGE DOCTORAL BRETAGNE

MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Étude théorique et numérique de la stabilité GKS pour des schémas d'ordre élevé en présence de bords

Mot clés : schémas numériques, condition de bord, stabilité forte, théorie GKS, déterminant de Kreiss–Lopatinskii

Résumé : Dans ce manuscrit, nous étudions la stabilité forte des schémas numériques explicites à un pas à coefficients constants, posés sur le demi-espace et possédant un bord à gauche. On suppose que ces schémas sont consistants avec l'équation de transport scalaire uni-dimensionnelle comportant une donnée de bord à gauche. Grâce au théorème de Kreiss et à la théorie développée par Gustafsson, Kreiss et Sundström, la stabilité forte est équivalente à l'absence de zéros du déterminant de Kreiss–Lopatinskii à l'extérieur du disque unité ouvert. On va alors décrire une stratégie numérique permettant de compter les zéros du déterminant de Kreiss–Lopatinskii afin de pouvoir conclure sur la stabilité forte du schéma.

La première partie de ce manuscrit décrit plusieurs approches de la stabilité et introduit les objets nécessaires à la compréhension des contributions, notamment la théorie de Gustafsson, Kreiss et Sundström et le déterminant de Kreiss–Lopatinskii. La deuxième partie est dédiée aux résultats théoriques et aux stratégies numériques pour le cas particulier des schémas totalement décentrés et pour le cas général.

L'enjeu est de trouver des stratégies efficaces et robustes pour étudier la stabilité de ces schémas, notamment au travers d'outils numériques et de la condition de Kreiss–Lopatinskii uniforme représentée par le déterminant de Kreiss–Lopatinskii.

Title: Theoretical and numerical analysis of GKS-stability for high order finite difference schemes with boundaries

Keywords: numerical scheme, boundary condition, strong stability, GKS theory, Kreiss–Lopatinskii determinant

Abstract: We study the strong stability of one-step explicit finite difference schemes set on the half-space with a left boundary condition. We work on schemes which are consistent with the scalar advection equation. Thanks to Kreiss theorem and GKS theory, the strong stability is equivalent to the absence of zero of the Kreiss–Lopatinskii determinant outside the open unit disk. Then we describe a numerical strategy to count the number of zeros of the Kreiss–Lopatinskii determinant in this domain.

The first part deals with different approaches to work on stability and introduce the tools needed to understand the contributions. The second part presents the details of the theoritical results and the numerical strategies for the particular case of totally upwind scheme and for the general case.

The goal is to introduce and to study a robust and efficient numerical strategy to handle strong stability, thanks to numerical tools and the uniform Kreiss–Lopatinskii condition.